Detection of functional modules in genomic and metagenomic datasets

Kumulative Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Sebastian Gil Anthony Konietzny

aus Düsseldorf

Sankt Augustin, November 2015

aus dem Institut für Informatik
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathemathisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent:	Prof. Dr. Alice C. McHardy
Korreferent:	Prof. Dr. Martin J. Lercher
Tag der mündlichen Prüfung:	09. Mai 2016

Selbstständigkeitserklärung

Ich versichere an Eides statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf erstellt worden ist. Arbeiten Dritter wurden entsprechend zitiert. Diese Dissertation wurde bisher in dieser oder ähnlicher Form noch bei keiner anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Sankt Augustin, den ....

(Sebastian Konietzny)

Statement of authorship

I hereby certify that this dissertation is the result of my own work, and that no other person's work has been used without acknowledgement. This thesis was created in accordance with the principles of good scientific practice of the Heinrich-Heine-University. It has not been submitted in the same or similar form to other institutions. I have not previously failed a doctoral examination procedure.

Summary

Background The diversity of microbial species is fascinating, and some microbes have extremely useful capabilities for technical applications. In many cases, the molecular mechanisms that are underlying cellular processes are valuable design templates for industrial processes in medical and biotechnological applications. One interesting example is the production of renewable biofuels. Plant biomass, such as grass and wood, can be chemically converted into biofuels. Appropriate industrial techniques already exist but need improvements to become cost-efficient. Strikingly, lignocellulose-degrading microbial species are capable to degrade plant biomass efficiently. But the underlying molecular mechanisms of the degradation processes are only partially understood. This example demonstrates the need for appropriate (computational) methods to investigate cellular processes of microbes. In particular, the development of sophisticated bioinformatics methods and specialized data mining techniques represents an important key factor to unravelling the complex mechanisms of cellular processes.

In order to understand the mechanisms of cellular processes in microbes, one needs to analyze the activities and molecular functions of proteins. Proteins are encoded in the genes of organisms, and result as products of gene expression. With modern DNA sequencing techniques, it became a highly automated and relatively cheap process to access the gene repertoires ('genomes') of organisms. As a consequence, thousands of sequenced genomes became available in public databases, and the numbers are rapidly increasing. Moreover, modern techniques enabled metagenome studies of microbial communities, i.e. the sequencing of environmental DNA probes without the need of cultivating organisms in the laboratories. A so-called metagenome thus represents the mixed genetic material of a microbial community of species. The fact that metagenomics is cultivation-independent enables scientists to study microbial organisms that were not accessible before. Estimates say that up to 99% of microbial species cannot be cultivated. Thus, metagenomics offers great opportunities to broaden our understanding of the microbial world. Overall, the millions of gene sequences already deposited in

public databases represent a rich basis for large-scale computational studies of their putative biochemical functions. Therefore, large parts of genetic studies are nowadays conducted *in silico* by using various bioinformatics tools.

Cellular processes typically correspond to one or more functional modules, which represent groups of functionally interacting proteins. Common examples of functional modules are metabolic pathways, protein complexes, and signal transduction chains. Studying the composition of functional modules is an important challenge because it paves the way to exploiting microbial proteins for improvements of biotechnological techniques. The problem here is to identify interacting proteins given only their gene sequences, and to understand the cross-effects between individual protein functions. A common approach for detecting interacting proteins is referred to as phylogenetic profiling. Its basic assumption is that functionally coupled genes tend to co-evolve, which suggests that protein-protein interactions (PPIs) are detectable from gene co-occurrence patterns across sets of genomes. This principle enables a computational identification of pairwise interactions of proteins and of groups of interacting proteins.

Key Challenges The key challenge of this PhD project was to develop new machine-learning-based methods for the computational detection of functional modules based on the principles of phylogenetic profiling. Notably, only a few previous studies had analyzed the applicability of standard phylogenetic profiling methods on large collections of genomes before, and the analysis of metagenomic datasets was largely untouched.

Results The author's main scientific contributions are the development and evaluation of two new methods for functional module inference for genomic and metagenomic input datasets (Konietzny *et al.*, 2011, 2014). These methods are based on probabilistic topic models, which originally stem from the field of text mining, and the idea of applying such models to gene sets of (meta-)genomes is new. Topic models are Bayesian graphical models which are known to be robust against noise in the input data. This property is important for the analysis of gene presence/absence patterns because currently available methods for DNA sequencing and gene prediction can produce erroneous outputs.

Moreover, the newly developed methods discussed in this thesis enable the identification of genomic elements, that is, proteins and entire functional modules that are linked to specific capabilities of cells ('phenotypic traits' of organisms, or 'phenotype' for short). Therefore, they represent valuable instruments for the identification of biocatalysts from microbes which might enable innovations in biotechnology and medical health care. As a test case, we investigated the microbial phenotype of plant biomass degradation, and demonstrated the successful application of the methods. Microbial degradation of lignocellulose is an important process for future improvements in the industrial production of biofuels, and the obtained results indicated an involvement of protein families which were not known to be important before. Thus, the methods allow the identification of new interesting candidate protein families for further research. Moreover, the methods can be used for screening the genomes of microbes with uncharacterized phenotypic traits to distinguish phenotype-positive from phenotype-negative candidates. Thus, with these new computational methods the screening of organisms can efficiently be done *in silico* without the need of running expensive experimental tests in laboratories. Linking the phenotypic traits of organisms to functional modules as well as predicting the phenotypes of organisms from their genomes are important contributions to microbial genome research. As a future perspective, the new methods could be used to study a variety of interesting cellular activities such as, for example, the synthesis of antibiotics, the degradation of environmental pollutants, or the pathogenicity of certain bacterial species.

Zusammenfassung

Hintergrund Die Diversität mikrobieller Spezies ist faszinierend; viele von ihnen besitzen äußerst nützliche Eigenschaften für technische Anwendungen. In vielen Fällen dienen die molekularen Mechanismen, welche zellulären Prozessen zugrunde liegen, als Entwurfsmuster für industrielle Prozesse, z. B. für medizinische und biotechnologische Anwendungen. Ein interessantes Beispiel dafür ist die Herstellung von erneuerbaren Treibstoffen. Pflanzliche Biomasse, wie Gräser und Holz, kann auf chemischem Wege in Biotreibstoff umgewandelt werden. Die entsprechenden industriellen Techniken existieren bereits, aber sie müssen verbessert werden, um kosteneffizienter zu werden. Interessanterweise sind Lignocellulose-abbauende Bakterien dazu fähig, Pflanzenmasse auf sehr effiziente Weise abzubauen. Die dem zugrundeliegenden molekularen Mechanismen sind allerdings erst teilweise entschlüsselt worden. Dieses Beispiel demonstriert die Notwendigkeit von geeigneten Methoden, um die Funktionsweise von zellulären Prozessen in Bakterien zu untersuchen. Insbesondere bioinformatische Methoden und spezialisierte Datamining-Techniken stellen Schlüsselfaktoren für die Aufklärung der komplexen Wirkungszusammenhänge in zellulären Prozessen dar.

Um die Wirkungsweise von zellulären Prozessen in Bakterien nachzuvollziehen, müssen zunächst die Aktivitäten und molekularen Funktionen einzelner Proteine verstanden werden. Proteine werden in den Genen eines Organismus kodiert und sind das Produkt der Genexpression. Mittels moderner DNA-Sequenzierungstechniken können die Genrepertoires (Genome) von Organismen auf kostengünstige, automatisierte Weise entschlüsselt werden. Als Folge wurden inzwischen Tausende von sequenzierten Genomen in öffentlichen Datenbanken veröffentlicht und die Menge wächst weiterhin rasant an. Darüber hinaus haben es moderne Techniken ermöglicht, Metagenomstudien von mikrobiellen Gemeinschaften durchzuführen, das heißt, eine direkte Sequenzierung von DNA-Proben ohne die bisherige Notwendigkeit, Kulturen von Bakterien im Labor züchten zu müssen. Ein Metagenom stellt deshalb eine Mixtur von DNA-Fragmenten unterschiedlicher Organismen dar, die eine mikrobielle Gemeinschaft bilden. Die Tat-

sache, dass Metagenomstudien keinen Kultivierungsschritt voraussetzen, ermöglicht es Wissenschaftlern, Bakterienspezies zu untersuchen, für die dies zuvor nicht möglich war. Laut einer Schätzung ist es für bis zu 99 % der existierenden mikrobiellen Spezies gar nicht möglich, sie in Kultur zu züchten. Aus diesem Grund bieten uns die technischen Möglichkeiten der Metagenomik ein breiteres Verständnis des mikrobiellen Kosmos, als dies zuvor möglich war. Zusammengenommen repräsentieren die Millionen von Proteinsequenzen, die bereits in Datenbanken gesammelt wurden, eine reichhaltige Datengrundlage, um mit Computerprogrammen die einzelnen Proteinfunktionen zu studieren. Dies ist der Hintergrund dafür, dass heutzutage ein Großteil genetischer Studien mittels bioinformatischer Methoden am Computer durchgeführt wird.

Zelluläre Prozesse basieren in der Regel auf einem oder mehreren funktionellen Modulen, die Gruppen funktionell gekoppelter Proteine entsprechen. Typische Beispiele für funktionelle Module sind metabolische Stoffwechselpfade, Proteinkomplexe oder Signaltransduktionspfade. Es ist eine wichtige Aufgabe, die Zusammensetzung funktioneller Module auf Proteinebene zu untersuchen, damit das daraus gewonnene Wissen für die Verbesserung biotechnologischer Verfahren genutzt werden kann. Das wesentliche Problem besteht darin, die funktionellen Wechselwirkungen zwischen Proteinen zu verstehen, und zwar einzig auf der Grundlage der Informationen über die DNA-Sequenzen der Proteine. Ein sehr gebräuchlicher Ansatz zur Detektion von Proteininteraktionen wird als “phylogenetisches Profiling” bezeichnet. Dem liegt die Annahme zugrunde, dass funktionell gekoppelte Gene zu paralleler Evolution (Ko-Evolution) neigen, so dass sich Protein-Protein-Interaktionen (PPIs) anhand von konservierten Kookkurrenzmustern von Genen in Genomen vorhersagen lassen. Dieses Prinzip bildet die Grundlage dafür, paarweise Interaktionen zwischen Proteinen, oder auch Gruppen interagierender Proteine mittels Computerverfahren vorherzusagen.

Herausforderungen Die zentrale Herausforderung im Rahmen der Forschungsprojekte dieser Doktorarbeit bestand darin, neuartige Methoden zur computergestützten Vorhersage von funktionellen Modulen zu entwickeln. Dazu sollten Verfahren des

maschinellen Lernens auf die Prinzipien des phylogenetischen Profilings angewendet werden. Nur wenige Vorläuferstudien hatten die Anwendbarkeit klassischer Profiling-Methoden auf sehr großen Mengen von bakteriellen Genomen getestet und die Analyse von Metagenomen mittels phylogenetischem Profiling fand zuvor praktisch nicht statt.

Ergebnisse Die wichtigsten wissenschaftlichen Beiträge des Autors dieser Arbeit liegen in der Entwicklung und Evaluierung zweier neuer Methoden, um funktionelle Module in genomischen und metagenomischen Datensätzen zu identifizieren. Die Methoden basieren auf so genannten probabilistischen "Themen-Modellen" (topic models), welche normalerweise im Text-Mining Verwendung finden. Die Idee, diese Modelle auf den Genrepertoires von (Meta-)Genomen anzuwenden ist neu. Topic-Modelle sind Bayes'sche grafische Modelle und als solche bekannt dafür, gegenüber verrauschten und mit Ungewissheit behafteten Eingabedaten robust zu sein. Diese Eigenschaft der Modelle ist von besonderer Wichtigkeit für die Analyse von Genprofilen (Vorhandensein oder Abwesenheit von Genen in verschiedenen Genomen), weil die zur Verfügung stehenden DNA-Sequenzierungsmethoden und Genvorhersagealgorithmen fehlerhafte Ergebnisse liefern können.

Mit den entwickelten Methoden lassen sich auch gezielt genomische Komponenten identifizieren, die spezifischen Eigenschaften (dem Phänotyp) von Bakterienzellen zugrunde liegen. Dabei kann es sich um einzelne Proteine oder vollständige funktionelle Module handeln, die mit dem Phänotyp assoziiert sind. Die neuen Methoden stellen deshalb wertvolle Instrumente zur Identifikation von Biokatalysatoren (Enzymen) dar, die zu Innovationen in den Bereichen Biotechnologie und Medizin beitragen könnten. Als Beispielszenario wurde der Phänotyp mikrobiellen Lignocelluloseabbaus von uns untersucht und daran die Nützlichkeit unserer Methoden demonstriert. Die Wirkungsweise des Abbaus von Lignocellulose durch bestimmte Bakterienspezies stellt eine wichtige Inspirationsquelle für zukünftige Verbesserungen bei der industriellen Gewinnung von Biotreibstoffen dar. Unsere Ergebnisse liefern Hinweise auf die Rollen von bisher nicht beachteten Proteinfamilien in diesen Prozessen. Dies zeigt das Potenzial unserer Metho-

den, neuartige Proteinfamilien zu entdecken, die wichtig für biotechnologisch relevante zelluläre Prozesse sind. Darüber hinaus können die Methoden benutzt werden, um Bakteriengenome zu klassifizieren, das heißt, automatisch zu erkennen, ob ein Genom die für einen bestimmten Phänotyp typischen genomischen Komponenten besitzt oder nicht. Auf diese Weise kann eine automatische Vorsortierung von Organismen am Computer durchgeführt werden, um Spezies eines bestimmten Phänotyps von anderen Spezies zu trennen. Dies erspart kostenintensive Tests im Labor. Neue Methoden zur Aufklärung der Beziehungen zwischen dem sichtbaren Phänotyp einer Zelle und den zugrunde liegenden funktionellen Modulen, sowie zur Vorhersage eines bestimmten Phänotyps anhand von genomischen Charakteristika stellen somit wichtige Beiträge zur mikrobiellen Genomforschung dar. Basierend auf den hier präsentierten Ergebnissen können diese neuen Methoden zukünftig genutzt werden, um weitere interessante zelluläre Aktivitäten zu untersuchen, beispielsweise die Synthese von Antibiotika, den bakteriellen Abbau von Umweltgiften, oder die Pathogenität bestimmter mikrobieller Organismen.

Danksagungen

Ich möchte mich bei meinen Betreuern, Koautoren und Kollegen für die gute Zusammenarbeit und Unterstützung bei der Umsetzung der Forschungsarbeiten bedanken. Besonderer Dank gebührt Frau Professor McHardy, deren Ideen und Erfahrungen eine wichtige Grundlage dieser Arbeit bilden.

Der Rückhalt meiner Familie und Freunde kann mit Worten nur unzureichend wertgeschätzt werden. Für mich war dieser Rückhalt fundamental wichtig und ich bin ihnen sehr zu Dank verpflichtet. Dies schließt vor allem meine engste Familie, meine Eltern und meinen Bruder mit ein. Aber auch eine ganze Reihe von sehr guten Freunden.

Besonders danken möchte ich an dieser Stelle dreien meiner Kollegen, die mich immer mit Rat unterstützt haben und auf diese Weise enge Freunde geworden sind. Christina Kratsch, Aaron Weimann und Lars Steinbrück waren ein enorm wichtiger Rückhalt für mich, ganz besonders in schwierigen Zeiten. Auf ihre fachliche Meinung und ihre Hilfsbereitschaft bei der Lösung von Problemen war immer Verlass. Das wird mir – zusammen mit vielen anderen schönen Erlebnissen während der Promotion – immer in Erinnerung bleiben. Dafür möchte ich mich bei allen Beteiligten bedanken.

Contents

Part I

Introduction and Background

CHAPTER 1

Introduction

The central promise of molecular life sciences is to provide solutions for problems in the fields of medical care and cost-efficient biotechnological production. This can be achieved by studying successful concepts from nature, such as the various capabilities of microorganisms to produce substances of medical or industrial importance through series of biochemical reaction steps arranged in metabolic pathways (Ogawa and Shimizu, 1999). The goal is to use biological processes as templates for industrial applications, or to identify new drug candidates from the protein molecules of microbial organisms (Huisman and Gray, 2002).

We therefore need to understand the molecular mechanisms of biological processes, which has traditionally been approached by cultivating microorganisms in the laboratory, followed by an extensive experimental and genetic characterization. The traditional work flow involves time- and labor-intensive steps in the laboratory ('wet lab') (Arnold, 2001), but within the last twenty years the approach has been revolutionized by new high-

throughput methods of functional genomics (Hawkins *et al.*, 2010) and next-generation DNA sequencing (NGS) techniques (Mardis, 2008). These innovations have greatly accelerated the genetic analysis of organisms because they deliver large quantities of experimental data in a short time and at low costs. Moreover, NGS techniques also enabled so-called metagenome studies, i.e. large-scale studies of microbial communities.

In the meantime, the increasing amounts of data have shifted the scientific focus away from data generation to automated data interpretation and motivated the development of new bioinformatics software (Mardis, 2011). Thus, it is the study of microorganisms with combined methods of functional genomics and bioinformatics which promises future innovations for technical applications. An important class of bioinformatics tools are *statistical inference* methods; where the process of inference refers to fitting statistical models to real world datasets in order to understand key properties and patterns of the datasets.

However, despite the remarkable progress in all fields of genome research, it remains a challenge to unravel the molecular mechanisms of biological processes in cases where it is not straightforward to transfer knowledge from well-studied model organisms.

The goal of the PhD project was to develop computational methods for the detection of functional modules, i.e. sets of functionally coupled proteins that are realizing together cellular processes in microorganisms. New methods are required to *i)* broaden our understanding of the diversity of biological processes that are realized in the microbial world, as well as to *ii)* investigate specific processes in targeted analyses of biotechnologically relevant microbes. With the methods developed during the PhD project, challenges *i)* and *ii)* both get addressed. The methods provide means of explorative analyses of the dominant functional modules in large sets of (meta-)genomes, as well as targeted investigations of the protein families underlying biological processes with a particular relevance for biotechnological or medical questions.

There already existed so-called genomic context (GC) methods for the detection of protein-protein interactions (Section 3.5), and some of these have also been used to predict functional modules (Section 3.6.3). A particular subclass of GC-methods are

called phylogenetic profiling methods; they typically analyze patterns of co-occurring genes across the gene sets of different organisms to predict functional relationships between genes (Section 3.6). Most of the methods try to predict pairs of interacting proteins only, so the inference of functional modules is not in direct focus and needs to be addressed separately, typically by examining cluster structures in protein-protein interaction networks. Unfortunately, GC-methods tend to produce high rates of false positive predictions, and especially phylogenetic profiling still needs to be improved (Subsection 3.6.3). Moreover, there is a need to develop robust inference methods capable of analyzing potentially incomplete metagenomic datasets (Section 2.3).

An important novelty of this thesis is the successful adaptation of so-called probabilistic topic models (PTMs) for functional module mining (Chapter 5). PTMs are specific types of Bayesian inference methods used in text mining to perform *latent semantic analysis* (LSA). The goal of LSA is to extract the central semantic concepts underlying a set of text documents (Landauer and Dumais, 1997; Landauer *et al.*, 1998), i.e. the dominant 'topics' that are discussed. Topics can be modelled as probabilistic clusters of words, describing groups of words with a common semantic context, such as words related to either 'music' or 'politics', for example. Interestingly, we found conceptual similarities between latent semantic analysis in texts and phylogenetic profiling on genomes (Section 5.1). Based on this key observation, PTMs were successfully adapted for functional module mining as part of this thesis (Konietzny *et al.*, 2011, 2014).

Despite the aforementioned similarities in concepts, it should be noted that there are fundamental differences between the challenges of method development for text mining and the discovery of biological processes. In text mining applications, a qualitative evaluation of the inferred clusters (topics) is straightforward. In most practical cases, researchers possess the necessary background knowledge about a domain to intuitively assess the semantic coherence of a group of words that was predicted. By contrast, it is far more challenging to assess the biological significance of a predicted functional module, that is, to distinguish a functionally coherent module from a randomly formed cluster of protein families. The qualitative assessment of a predicted module in fact requires

detailed investigations of the individual functions of each protein family involved. This usually takes a lot of time for literature research to find evidence. Moreover, it becomes even more difficult by the fact that functional characterizations are available only for a minority of the protein families with sequenced members. By consequence, it needs an efficient strategy to assess the degree of functional coherence of a predicted group of gene or protein families. In particular, the intended large-scale approach of functional module mining, with hundreds of predicted modules, makes it time-intensive to investigate individual modules in sufficient detail. Therefore, together with the PTM-based method for functional module mining, a new evaluation framework was proposed in Konietzny *et al.* (2011). The framework enables a qualitative assessment of large sets of functional modules as well as taking into account both protein families with and without previous functional characterizations.

The remainder of this section highlights the potential benefits from studying the cellular processes of microbes and motivates the challenges for this PhD project.

1.1 Prospects of microbial genome research for technical innovations

Microorganisms are omnipresent in nature (Lawrence, 1999). In particular, bacterial and archaeal species[1] inhabit all different kinds of environments, including those with extreme conditions. Over billions of years, microbes have evolved and developed efficient biochemical processes in adaptation to the requirements of their natural environments. The result today is a rich diversity of microbial species, many of which possess specialized abilities to avoid competition with other species by occupying environmental niches (Lawrence, 1999; Rocap *et al.*, 2003). This evolutionary process has produced a wealth of specialists' solutions to handle a variety of biochemical settings, which makes microorganisms a promising resource for innovations in biotechnology (Lorenz and Eck, 2005). On today's markets, more than 500 industrial products and about 150 industrial processes are already based on microbial enzymes (Adrio and Demain, 2014). It is expected for the future that about 50,000 conventional chemical products could be substituted by substances produced from renewable natural resources with biotechnological methods (Jarrell, 2009).

Functional genomics is vital to the discovery of biochemical agents from natural resources that have the potential to improve processes in the chemical industry, such as the production of drugs or biofuels, because it aims to explain the cause-and-effect-chains linking the genes of organisms ('genomes') to the observable properties of their cells ('phenotypes'). The key to understanding the molecular mechanisms behind specific properties of cells is to analyze the concerted actions of proteins that perform basic biochemical reactions (Bork *et al.*, 1998). Proteins are organized in functional modules, which are the organizational units of cellular processes (Hartwell *et al.*, 1999).

In functional genomics, the first step is to establish a complete catalogue of all proteins that are encoded in the genome of an organism (Hieter and Boguski, 1997). By isolation and sequencing of the DNA, that is, determining the sequence of nucleotides

[1] Sometimes also summarized as prokaryotes, however, this term is becoming outdated.

along the DNA molecule, one can locate the positions of genes in the genome sequence and determine their individual nucleotide sequences, which is an important prerequisite for all further investigations. Two important successive steps in functional genomics are to determine the biochemical functions of the proteins that are encoded in the genome and to unravel their functional dependencies and interplay. Functional modules are of particular interest for medical and biotechnological applications because their analysis may explain aspects of the cell's activities that cannot be ascribed to a single protein.

1.2 Exploring microbial diversity: Dynamics and progress

In the last fifteen years, a rapid progress in wet-lab high-throughput techniques, DNA-sequencing protocols and bioinformatics methods has shifted the scientific focus from single-organism-studies to large-scale comparative studies of all currently known bacterial species (Achtman, 2012; Bentley and Parkhill, 2004; Kyrpides, 2009; Wren, 2000). For such studies, the combined genetic information of hundreds or thousands of microbial species stored in public databases is processed and data mining methods are applied to identify patterns in the data that enable the discovery of unknown biological concepts, or to test scientific hypotheses on a large-scale level (Medini *et al.*, 2008). The power of this approach promises further glimpses into the complex universe of microbial phenomena that will hopefully lead to an in-depth understanding of the biochemical processes when sufficient data can be analyzed (Bansal, 2005). However, the microbial diversity of species and biochemical functions is remarkable (Horner-Devine *et al.*, 2004), and the currently sequenced datasets are of a limited taxonomic range (Rinke *et al.*, 2013).

Different from taxonomic classification systems for higher animals, the definition and distinction of microbial species based on genetic similarity[2] is difficult, due to the evolutionary dynamics caused by the short-term generation cycles of bacteria, and the exchange of genetic material between related and non-related organisms (Achtman and

[2]The classical distinction based on the success of reproduction with sexual partners of the same or different species is not useful for microbes, as they reproduce by cell division without the need of a sexual partner.

Wagner, 2008). As a result, the estimates of the total number of existing microbial species are uncertain, ranging from tens of thousands to more than a billion (Dykhuizen, 2005; Schloss and Handelsman, 2004). During the time period of this thesis, a dramatic increase in sequenced microbial genomes has occurred and it is still ongoing. It should be noted, however, that a large fraction of sequencing projects have not been finished yet and that the genome sequences can be highly fragmented and annotations being marked as 'draft annotations'. The number of finished sequencing projects has grown within five years from 863 in 2009 to 12,060 in 2014, while currently more than 50,000 sequencing projects have been officially registered[3]. This remarkable increase of genomic sequences was due to the development and improvements of next-generation sequencing techniques, which sped up the process of sequencing by several orders of magnitude and dropped the cost per sequenced base pair (bp) considerably (Mardis, 2011).

With the advent of NGS methods, microbiologists have stepped into a new era concerning the discovery of novel microbial species, as previous methodological limitations could be overcome. Up to 99% of all existing microbial species cannot be cultivated with the standard laboratory techniques because their unknown physiological needs cannot be served adequately (Hugenholtz, 2002). This has led to a significant bias in the phylogenetic diversity of species that are represented in databases (Hugenholtz, 2002). Therefore, cultivation-independent methods were needed. A key technology to overcome these hurdles is metagenomics. In a metagenome study, we take a random sample of DNA from a natural habitat – without targeting individual organisms. The DNA sample thus represents a pooled mixture of DNA fragments from the microbial community that resides at the examined location. This approach differs from the traditional way of isolating cells of single organisms from the environment to grow them in cultures for obtaining their DNA in large amounts. Due to improvements of DNA amplification techniques, even small amounts of DNA are sufficient for analyses nowadays (Mardis, 2008), which made this approach feasible.

[3]See data at http://www.genomesonline.org/statistics

1.3 Motivation and research aims

As pointed out, deciphering the groups of proteins that are underlying cellular processes, especially those related to phenotypes of medical or biotechnological interest, is a key to future technical innovations and efficient industrial processes. However, despite the progress in the development of bioinformatics software, an accurate functional characterization of the gene functions and functional modules encoded in the vast amounts of sequencing data that have accumulated inside public databases has become a major bottleneck. Our knowledge about biological processes mainly stems from a few model organisms and thus should be considered to be biased and limited, seen with respect to the enormous diversity of microbial species. In fact, we might even expect to discover unknown pathways within well-studied model organisms (Date and Marcotte, 2003; Saito *et al.*, 2010). Therefore, the development of computational methods for the inference of functional modules from genomes and metagenomes is very important (De Filippo *et al.*, 2012). In particular, there is a need for *de novo* methods that do not rely on *a priori* information about known pathways, and thus allow for an exploratory discovery of completely novel functional modules.

At the time of writing this thesis, an online literature search using the buzzword 'functional modules' resulted in more than 1,000 hits in the titles or abstracts of scientific articles listed in Pubmed[4], an online repository for publications in the life sciences, illustrating the importance of this prominent research field. Scientific interest in functional modules and, more generally, in protein-protein interactions is also grounded in the promises of the so-called 'guilt-by-association principle' (Aravind, 2000), which states that the function of a protein can be determined from its interaction partners. As a consequence, genomic context methods, which are able to identify functional interaction partners of a protein, can serve to support gene function prediction methods based on guilt-by-association (Huynen *et al.*, 2000).

GC-methods detect functional interactions between proteins by considering the genomic context of genes, and thus do not require *a priori* knowledge about the

[4] www.ncbi.nlm.nih.gov/pubmed, accessed 12/08/15.

biochemical functions of the proteins. This also makes them candidates for functional module mining in a *de novo* fashion. However, as pointed out in a recent review by Muley and Ranjan (2013), only few studies have addressed the benefits and limitations of genomic context methods for the reconstruction of biological pathways from the gene sets of genomes.

The central aim of this thesis was to develop advanced computational methods for the detection and reconstruction of functional modules in large collections of genomes and metagenomes. The intended approach should target functional modules in a *de novo* exploratory fashion, and was motivated by the need to improve and extend our general knowledge about biological processes and their functional modules, especially for metagenomes. The methods were outlined from a machine learning perspective to detect functional modules by analyzing the distribution patterns of genes across datasets. This idea is related to phylogenetic profiling (Pellegrini *et al.*, 1999), a special type of the genomic context concept.

In contrast to traditional genomic context methods, the outlined methods should be designed to cope with the special issues of metagenomic datasets. Metagenomic datasets comprise gene inventories of microbial community members, but – with only a few exceptions – they lack further data from high-throughput experiments such as e.g. gene expression profiles, due to technical limitations which still need to be overcome, and because the necessary analysis steps are expensive (Parro *et al.*, 2007; Simon and Daniel, 2011; Warnecke and Hugenholtz, 2007). Therefore, limiting the required input data to gene distribution patterns makes the methods suitable for the analysis of metagenomes. Moreover, the gene inventories of the community members derived from the metagenomic sequence data are usually incomplete and lack positional information about the exact locations of genes on the original genomes. This partial lack of information makes it difficult to apply genomic context methods such as gene neighborhood detection and phylogenetic profiling. In order to cope with this problem, we considered robust Bayesian methods to exploit gene co-occurrence patterns in metagenomes.

Apart from a general exploration of biological processes, there is a need to specifically

identify the genomic elements causing known biotechnologically relevant features of microbial cells ('phenotype[5]-related genomic elements'). There are thus two distinct paradigms for the analysis of cellular processes. On the one hand, an unbiased (unspecific) large-scale exploration of the existing functional modules in microbes, and on the other hand, a targeted search for phenotype-related protein families and functional modules. From a machine-learning perspective, the two paradigms could be reflected in the choice of *unsupervised* or *supervised* inference methods, respectively. In an unsupervised setting, one would not distinguish between organisms with different phenotypes, whereas supervised methods could be trained to use known phenotypes of organisms as a guideline to the discovery of phenotype-related genomic elements. Both types of methods should be considered to solve the problem of functional module detection with respect to the exploratory and targeted approaches discussed. In any case, potential functional modules implied by the new methods would represent functional contexts that can guide biologists to characterize the functional roles of the proteins involved – following the guilt-by-association principle.

[5]In general, phenotypic traits of cells are all observable or measurable properties of cells; in contrast to the *genotype*, which refers to the genetic make up of cells. However, we often use the term phenotype in an exclusive way to refer to a specific aspect of cells caused by cellular processes that are targets for biotechnological research.

1.4 Outline

This document represents a cumulative dissertation which is based on three peer-reviewed articles; published in international journals with a focus on bioinformatics and biotechnology. The thesis is composed of four parts, of which the first places the articles into a larger scientific context (Chapters 1 to 3), and introduces probabilistic topic models (Chapter 4) which represent a fundamental concept for the presented work.

The two following parts present the publicized articles of the author that summarize the research activities on unsupervised inference of functional modules (Chapters 5 to 7) and the detection of phenotype-defining genomic elements (Chapters 8 to 10), respectively. The thesis closes with a summary of the results (Chapter 11), and a discussion of further research questions that are related to this work, and could be addressed in the future (Chapter 12).

The articles are presented in chronological order in the published versions of the respective journals, and the author's contributions to each research project are described (Section 5.2, Section 8.3 and Section 8.4).

CHAPTER 2

Functional genomics of microbial species

This chapter introduces the biological concepts underlying the presented articles. Moreover, it highlights the prospects and existing challenges of microbial genome research which motivated our research projects.

2.1 Overview

Genomic research provides the basis for the computational analysis of biological processes that exist in microbial species. Its techniques deliver the gene repertoires that are encoded in the genomes of organisms (Section 2.2). In a similar way, metagenomics provides us with the gene sequences of microbial communities, but the community structure of metagenomic samples requires some additional analysis steps (Section 2.3).

The research on biological processes has a long tradition. A few basic cellular activities are assumed to be encoded in the genomes of all microbial cells (Section 2.2).

These so-called housekeeping functions represent a core set of biological processes which are well-conserved in evolution. Housekeeping functions are a central aspect of microbial life that have been studied quite intensively. In contrast to this, many biotechnological or medical studies concentrate on specialized capabilities of microbes, which are typically less well understood.

The goal of protein- and process-level functional annotation is to understand the intricate interplay of gene products in the realization of the biological processes of interest (Section 2.4). Bioinformatics has helped to analyze and interpret the gene contents of thousands of genomes and also metagenomes, but, despite this success, many cellular activities and functions of individual proteins could not be characterized yet (Section 2.7).

Further concepts of particular importance in the context of the presented articles are assignments of gene products to controlled vocabularies (Section 2.5), protein-protein interaction networks in cells, and reference databases for functional modules (Section 2.6 for both).

2.2 The layout of bacterial genomes and their housekeeping genes

Most bacterial and archaeal genomes are organized in the form of a single circular DNA molecule (the 'chromosome') that encodes the gene repertoire of the organism, but in some cases genes are also encoded on small additional circular DNA molecules called plasmids (Bentley and Parkhill, 2004). Across different species, bacterial genome sizes can vary between less than 200,000 and more than 13,000,000 nucleotide base pairs (bps) (Rocha, 2008), and there can be differences of up to one million bps even on the level of different strains of the same species (Bentley and Parkhill, 2004). Bacterial genomes have a high coding density of genes, with approximately one gene per one thousand bps, and small intergenic DNA stretches in between (Bentley and Parkhill, 2004). Thus, the sizes of the gene repertoires can vary between a few hundred and a few thousand

genes depending on the organism, and even on the strain level there can be significant differences of the gene content (Achtman and Wagner, 2008; Tettelin *et al.*, 2008).

Comparative genome analyses have shown that genomes from related species often show a high level of synteny, that is, that the linear order of genes is maintained in large parts over short evolutionary time scales (Bentley and Parkhill, 2004). In general, organizational features of genomes seem to be highly conserved across species. One of the reasons might be that the structure of genomes is influenced by close interactions of cellular processes with the bacterial chromosome (Rocha, 2008). In contrast to this, the composition of gene repertoires is flexible to an astonishing extent across (even closely related) organisms (Medini *et al.*, 2005; Rocha, 2008).

Notably, there can be high degrees of redundancy in gene sets, as it is very common for genes to exist in several copies in a genome, but the copy numbers of essential genes in bacterial genomes tend to be small (Kepes *et al.*, 2012). Processes of gene amplification are frequent in bacteria; they are assumed to be essential for the genesis of new biochemical functions, and represent a mechanism of adaptive evolution for cells to overcome short-term selective pressures (Andersson and Hughes, 2009). Finally, the gene repertoires of organisms are influenced by gene transfer events between organisms ('horizontal gene transfer', HGT) (Fondi *et al.*, 2009).

Microbial cells are complex systems. They depend on a set of basic activities, such as the uptake of nutrients from the environment to gain energy, or the adaptation of their inner state in response to changing outer conditions. The gene repertoires of bacterial cells thus include a core set of housekeeping genes which are essential for maintaining the basic functions of life, that is, metabolic homeostasis, reproduction and evolution (Luisi *et al.*, 2002). Several efforts have been made to characterize a hypothetical minimal set of essential genes that are necessary to maintain a functional bacterial cell (Henry *et al.*, 2010); for example, Gil *et al.* (2004) have compiled a set that consists of only 206 protein-coding genes. Such characterizations are difficult because the requirements of cells depend on varying characteristics of their environments, but there is growing evidence that the order of magnitude of the gene number reported by

Gil *et al.* is likely correct (Henry *et al.*, 2010). Table 2.1 summarizes the main cellular processes that are encoded by the gene set presented by Gil *et al.*, and thus represents an overview of the housekeeping functions that are assumed to exist in most bacterial cells (Figure 2.1). Notably, more than half of the 206 genes of the proposed minimal gene set are involved in RNA metabolism, which demonstrates the large proportion of housekeeping proteins that act on the DNA and RNA molecules of the cell. Processes of intracellular information storage and processing are well-conserved ancient functions, whereas metabolic processes are flexible in the sense that many alternative metabolic pathways exist that can satisfy the basic metabolic requirements for life (Henry *et al.*, 2010). Apart from these results, another interesting observation is that the functions of many of the genes that were identified to be essential for cellular life still remain unclear or unknown (Gil *et al.*, 2004; Henry *et al.*, 2010), which shows the boundaries of the common knowledge about key biological functions and processes.

2.3 Sequencing of genomes and metagenomes

Genome sequencing usually follows the so-called 'shotgun approach', which basically means that a DNA molecule is not processed as a whole but in many tiny pieces of only 40-250 bp lengths. With a few exceptions[1], modern DNA NGS techniques can only sequence these very short DNA stretches. Finally, the original DNA sequence needs to be reconstructed from the short DNA reads by means of computational methods. A number of computational post-processing steps are necessary to obtain the original sequence information from the extremely short and sometimes erroneous DNA sequence fragments (Wooley *et al.*, 2010).

For metagenomes, the task of sequencing is more difficult due to the pooled DNA sample from various (partly unknown) species of a microbial community. The task can be split into two major challenges. In the assembly stage, overlapping reads are identified

[1] Nanopore (Clarke *et al.*, 2009) and PacBio (McCarthy, 2010) sequencing are two novel approaches to sequencing long DNA strands as a whole. However, these techniques are currently still in an early stage of market release and have not been widely used so far.

Category	**Subcategory**	**Subtypes**
Information storage and processing	DNA metabolism	Basic replication machinery DNA repair; restriction and modification
	RNA metabolism	Basic transcription machinery Translation RNA degradation
Protein processing	Folding	Posttranslational modification Activity of chaperones
	Secretion	Recognition of signal peptides
Cell structure	Cell shape	Cytoskeleton
	Cell division	
	Substrate transport	ABC transporters; PTS transporter for glucose; transporters for mono- and divalent cations
Energetic and intermediary metabolism	Central metabolism	Glycolysis; gluconeogenesis; pyruvate metabolism; TCA cycle
	Electron transport chain and proton motive force generation	
	Pentose phosphate pathway	
	Biosynthesis of amino acids	
	Biosynthesis of lipids	
	Biosynthesis of nucleotides	
	Biosynthesis of cofactors	

Table 2.1: The core features of a minimal bacterial cell. The description is based on the functional characterization of the minimal bacterial gene set proposed by Gil *et al.*. The authors assumed that the minimal cell would be able to obtain basic components from the environment, which also includes amino acids. Therefore, some functions, such as the biosynthesis of amino acids, were merely included in the table for the sake of completeness.

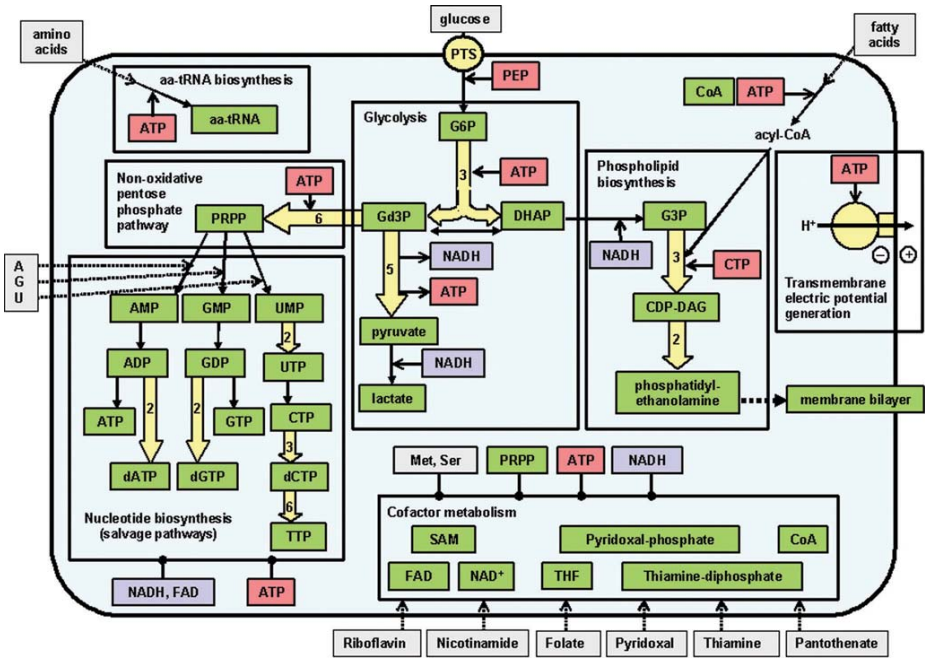


FIG. 1. A minimal metabolism. The minimal cell can obtain its more basic components from the environment: glucose, fatty acids, amino acids, adenine, guanine, uracil, and coenzyme precursors (nicotinamide, riboflavin, folate, pantothenate, and pyridoxal). Each box includes the metabolic transformations classified in major groups of pathways: glycolysis, phospholipid biosynthesis, nonoxidative pentose-phosphate pathway, nucleotide biosynthesis, synthesis of enzymatic cofactors, and synthesis of protein precursors, i.e., aminoacyl-tRNAs (aa-tRNA). Arrows with discontinuous lines represent incorporation from the environment. Single continuous arrows represent single enzymatic steps, whereas wide arrows represent several enzymatic steps (the number within the arrow indicates the number of steps). Lines with a final black point indicate the necessity of metabolites for some of the transformations inside the corresponding box. Metabolic intermediates and final pathway products are in green boxes. Metabolites acting as a source of chemical energy are in red boxes. Reducing-power cofactors are in light blue boxes. Abbreviations (besides the accepted symbols and those defined in the text): PEP, phosphoenolpyruvate; G6P, glucose-6-phosphate; Gd3P, glyceraldehyde-3-phosphate; DHAP, dihydroxyacetonephosphate; G3P, *sn*-glycerol-3-phosphate; CDP-DAG, CDP-diacylglycerol; SAM, *S*-adenosylmethionine; THF, tetrahydrofolate. Metabolic precursors of external origin are in gray boxes.

Figure 2.1: A model metabolism of a minimal cell based on the minimal gene set proposed by Gil *et al.* (2004). (Image source: Gil *et al.* (2004))

and fused into longer sequences (contiguous sequences, contigs for short) to enable later steps of the analysis pipeline. The resulting contigs then need to be partitioned into distinct sets that reflect their organismal origin, a process called binning (McHardy and Rigoutsos, 2007). Nevertheless, despite the progress in both problem fields, the analysis of metagenome data, which are typically incomplete and can be noisy depending on the sequencer platform, still remains considerably more difficult compared to genomic sequences of isolate species. In a metagenomic sequencing experiment, the sequencing depth for less abundant members of the community is often insufficient, which, in addition to the uncertainties about the organismal origins of the DNA fragments, affects the success of genome recovery (Dröge and McHardy, 2012). As a consequence, the

recovered gene repertoires of metagenomic datasets are often incomplete.

2.4 Protein- and process-level annotations

Functional genomics aims at understanding cells at all levels of their activities. In particular, this has been condensed in the goal of systems biology, that is, to model complex biological systems in mathematical frameworks to enable *in silico* simulation studies for hypothesis testing (Di Ventura *et al.*, 2006). This endeavor requires a precise knowledge of the biological processes that exist in various microbes. A concept related to this is *process-level* annotation, that is, the characterization of the processes that an organism is capable of (Stein, 2001) (Figure 2.2).

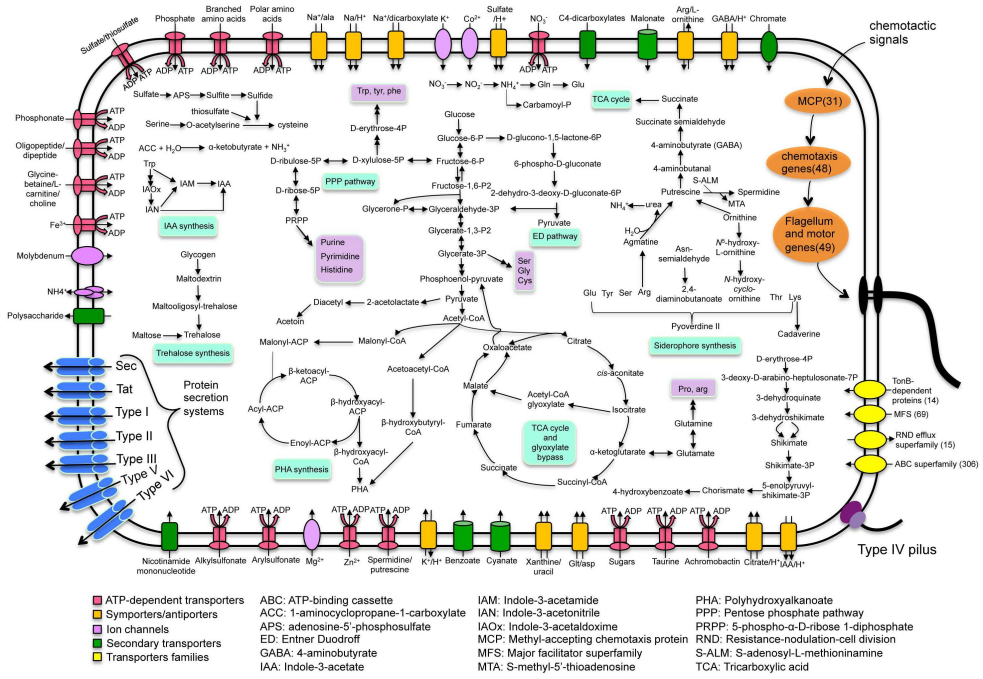


Figure 2.2: Schematic overview of metabolic pathways and transport systems in *Pseudomonas* sp. UW4. *Pseudomonas* sp. UW4 is a plant growth-promoting bacterium. This example serves as a high-level overview of the typical biological processes in bacterial cells. (Image source: Duan *et al.* (2013))

To fully understand the mechanisms of a cellular process, one needs to analyze

the individual molecular functions of its building blocks (Reed *et al.*, 2006). Proteins represent the most relevant functional units; however, additional molecules, such as catalytic RNAs (ribozymes), can be involved, but less frequently (Walsh, 2001). All these elements are encoded in the genome of an organism, and emerge as products of gene expression. We therefore need to determine the biochemical activities of gene products. This is referred to as *protein-level* functional genome annotation (Stein, 2001).

Multifunctional proteins typically consist of several functional subdomains, that is, they possess a modular protein domain architecture. For example, two separate structural parts of proteins are usually responsible for DNA binding and DNA modification. This also explains why there might be close similarities between parts of the amino acid sequences of proteins with seemingly different functions. Similar to gene families, there exist various conserved protein domain families, many of which have been catalogued in protein family/domain databases like e.g. Pfam (Bateman *et al.*, 2004).

Combinations of DNA sequencing techniques, gene detection software, genetic experiments, and computational tools for gene function prediction are routinely used for genome annotation (Stein, 2001). In order to obtain a high-quality annotation for an organism, that is, a sufficient description of the gene repertoire and the functions of individual genetic elements, all lines of evidence need to be integrated and reviewed by biologist experts (Valencia, 2005). This is a labor-intensive process; however, due to progresses in the development of computational tools, and the general performance increase of the compute infrastructures, draft genome annotations can be obtained in a mostly automated fashion nowadays (Aziz *et al.*, 2008; Meyer *et al.*, 2003). Most of the sequencing centers provide bioinformatics tools for automated genome annotation, and several comprehensive repositories providing data of annotated genomes and metagenomes are accessible online, such as the Integrated Microbial Genomes (IMG) database (Markowitz *et al.*, 2012), IMG with microbial samples (IMG-M) (Markowitz *et al.*, 2008), CAMERA (Seshadri *et al.*, 2007), and the metagenomics RAST (MG-RAST) server (Meyer *et al.*, 2008).

As described, protein-level annotation is concerned with describing the function of a

gene product. Several physico-chemical parameters, such as the reaction mechanism, or stoichiometric coefficients, determine the molecular function of gene products; thus a function can be described on different levels of detail. For example, one could explicitly list all relevant physico-chemical parameters in the description of the gene, or describe its biochemical activity on a more course-grained level, such as, for example, 'DNA-binding protein'. The level of detail that is required for the descriptions of functions depends on the applications. Sometimes it is sufficient to characterize the most relevant functions of genes on a superficial level, whereas medical studies often require precise models of cellular activities. In practice, trade-offs between the technical possibilities and their associated costs determine the targeted level of detail. A broad functional characterization of the protein set encoded in a (meta-)genome can easily be obtained, while generating sufficiently detailed knowledge for simulations of metabolic networks (e.g. flux balance analysis (Raman and Chandra, 2009)), or turnover rates of metabolites (Kell, 2004) needs tremendous efforts and would be too expensive in most cases. Therefore, sophisticated methods of systems biology are typically limited to a few well-studied model organisms and cannot be applied to the majority of sequenced bacterial species, and microbial community members.

2.5 Controlled vocabularies

The function of a gene can be described in a textual form. This allows to formulate a comprehensive list of details in a human-readable format. However, frequently occurring inconsistencies in the usage of words and formats can lead to ambiguities in the descriptions, and, therefore, textual descriptions in natural language are not amenable to an automated computational processing in large-scale comparative genomes studies. For this reason, controlled vocabularies, such as the Enzyme Commission Classification system (EC numbers), were defined, which allow for a consistent annotation of gene functions across genomes (Friedberg, 2006). EC numbers are numerical identifiers that code for specific metabolic reactions. They can be assigned to genes to characterize the

molecular function of the gene products. However, EC numbers are only defined for metabolic functions of enzymes. Another example of a controlled vocabulary is the Gene Ontology (GO) (Ashburner *et al.*, 2000). The concept of the GO vocabulary allows for the description of multiple levels of function, i.e. molecular function and process-level activities, which makes GO, in principle, a powerful tool for gene annotation. However, GO was originally developed for the annotation of eukaryotic genomes and was thus for a long time said to be less accurate for the annotation of bacterial and archaeal genomes (Mao *et al.*, 2005). As yet another example, the specialized 'Carbohydrate-Active EnZymes database' (CAZy) provides a controlled vocabulary for proteins that process carbohydrate molecules such as glucose or cellulose.

For an illustration of the concepts, Figure 2.3 shows the experimental GO-annotation of a multifunctional protein that is involved in several biological processes including cutting of RNA molecules (Radivojac *et al.*, 2013). The protein consists of seven protein domains which have been annotated in terms of Pfam families. This example illustrates the complexity of the task to predict the functional activities of a single protein.

Controlled functional vocabularies are only useful for annotating genes with known functions, and thus may not cover large parts of microbial gene repertoires. As an alternative, identifiers of gene/protein families can serve as a controlled vocabulary to increase the coverage of genes. Gene annotation then corresponds to the identification of a member of a gene (or protein) family. By annotating the gene with an identifier that refers to the family, all characteristics previously ascribed to the family can be implicitly adopted for the gene. This concept is merely based on sequence homology between the members of a family, and does not require knowledge about the molecular functions of the family members. The motivation behind this annotation approach is to consistently name gene products across genomes. Many of the existing gene/protein families could already be characterized, and were collected in public databases (e.g. Pfam (Bateman *et al.*, 2004), COG (Tatusov *et al.*, 1997), KEGG orthology (Kanehisa *et al.*, 2004), and eggNOG (Jensen *et al.*, 2008)). As mentioned before, the Pfam database is also an example of a protein domain database.

2.6 Protein-protein interactions and functional modules

The function of a protein can have multiple aspects and it usually depends on the cellular context (Jeffery, 1999). We need to take the interaction partners of proteins into account if we want to correctly annotate the functions of a protein. For example, proteins may work together in metabolic pathways, where a pathway is a process that converts chemical substances through a series of biochemical reactions. If a protein is involved in two different pathways, its interactions with other proteins define two different functional contexts that determine the respective function of the protein.

Protein-protein interactions (PPIs) can be physical, as, for instance, for members of a common protein complex, and they can be functional (Minguez and Dopazo, 2010). An example of a functional relationship is a pair of proteins that catalyze two successive steps in a metabolic pathway, where the input of the second reaction is the product of the first. A third category of relationships are genetic linkages. These can be determined in gene knockout experiments, and represent dependency relationships between genes; i.e. loss of the activity of one gene affects the activities of its interaction partners. For example, if the product of one gene is a suppressor for the expression of another gene, than the loss of the first gene might lead to an activation of the second.

We refer to a group of proteins as a functional module if they participate jointly in a biological process. A metabolic pathway is a typical example of a functional module because it corresponds to a set of proteins (enzymes) that catalyze the individual reaction steps. Apart from metabolic processes, functional modules may also represent protein complexes, or signal transduction cascades.

Several experimentally validated functional modules that were discovered through process-level annotation of model organisms have been collected in public databases. A comprehensive overview of web-accessible databases for metabolic pathways and protein interaction networks can be found in the meta-database Pathguide (Bader *et al.*, 2006). Prominent examples are the pathway database of the Kyoto Encyclopedia for Genes and Genomes (KEGG) (Kanehisa *et al.*, 2004), and the BioCyc collection (Caspi *et al.*,

2008) of several organism-specific pathway databases, which also includes one of the best curated databases for the model organism *Escherichia coli* (EcoCyc, Keseler *et al.*, 2011). Moreover, pathways from EcoCyc and several other BioCyc databases have been integrated into the MetaCyc database in an attempt to provide a single resource for experimentally verified pathways from a wide variety of organisms (Krieger *et al.*, 2004). Altogether, these databases represent valuable sources of knowledge about metabolic and signal transduction pathways, as well as validated protein-protein interactions. The general principles that emerge from the comparative study of these known functional modules can then be transferred to new organisms (Caspi *et al.*, 2008; Kanehisa *et al.*, 2008).

As most of the processes in cells are intertwined, proteins are embedded in cell-wide protein-protein interaction networks (PPINs) (Figure 2.4). Examples of networks, apart from networks of metabolic enzymes, are regulatory and signal transduction networks, which encode internal routes of information processing that are needed for the cell's self-maintenance and interactions with the environment (Aittokallio and Schwikowski, 2006; Klemm and Bornholdt, 2005). Within the networks, the boundaries of different processes are hard to define (Green and Karp, 2006). In many cases, networks can be decomposed into densely-connected sub-networks, some of which correspond to functional modules (Brohee and van Helden, 2006; Papin *et al.*, 2004). However, the definitions of individual functional modules are not unambiguous. For example, the ambiguous choices of metabolic pathway boundaries, even in well studied metabolic networks, resulted in differences of pathway definitions in databases such as KEGG and BioCyc. KEGG pathways (also called maps) are on average 4.2 times larger than BioCyc pathways, and summarize multiple different but chemically related biological processes (not necessarily from the same organism) in a substrate-centric reaction graph (Green and Karp, 2006). BioCyc pathways, in contrast, are compact organism-specific pathways that are often genetically regulated as a unit (Green and Karp, 2006). More recently, Altman *et al.* (2013) conducted a comparison of the KEGG and MetaCyc databases, reporting 237 KEGG pathway maps (average number of reactions: 28.84)

and 1,846 atomic MetaCyc pathways (average number of reactions: 4.37), respectively[2].

In many cases, biotechnologically relevant capabilities of cells are caused not only by the action of a single protein but the concerted action of proteins that belong to either the same functional module, or to a couple of interacting modules. An example for this is the degradation of plant cell walls, which has applications for the production of biofuels (Section 8.1). The plant cell wall has an intricate composite structure, and, therefore, several sub-processes are involved in degrading the various components.

2.7 Discovery of novel protein families

Genome sequencing projects have already identified a wealth of protein-coding nucleotide sequences from a broad range of organisms. During the last ten years, millions of protein sequences have been deposited in public databases such as UniProtKB/TrEMBL (Figure 2.5). In addition, various large-scale metagenomic studies, such as the Global Ocean Sampling (GOS) expedition, contributed a bulk of new data (Godzik, 2011). Looking at the vast amount of sequence data, not every sequence codes for a unique molecular function. Instead of this, sequences from different organisms often share sequence similarity with each other which indicates a functional similarity of the encoded gene products. The space of sequences clusters into distinct protein families (Levitt, 2009). Members of a family are likely to be related by a common ancestor sequence that diversified through mutational events in the course of evolutionary history. As protein families are clusters of similar sequences, they also serve as a compact representation of the parts of the protein universe that we already know (Levitt, 2009). With every new sequencing project, we collect a sample from the overall sequence space.

[2]Thus, different from the earlier study by Green and Karp (2006), KEGG pathways were found to be even more than 6 times larger than atomic MetaCyc pathways.

Buried in the massive amounts of data, we expect to find sequences that encode novel biocatalysts with the potential to improve existing industrial processes for a broad range of applications (Singh, 2010; Wilson and Piel, 2013). A common approach to finding the 'gold nuggets' is to combine functional genomics with bioinformatics. If some sort of *prior* knowledge about a chemical process is available, bioinformatics approaches can be guided to specifically detect potential biocatalysts that are likely to be useful for an industrial application. However, biocatalysts can have a plenitude of biochemical properties, and each industrial process comes with its own demands. Thus, it is necessary to perform a functional screening of the protein functions in the laboratory (Figure 2.6), and only very few candidates will be useful in the end. Screenings are associated with extra costs and are thus ideally applied to small sets of pre-filtered, highly promising candidates. It is hence clear that bioinformatics methods are expected to predict the potential biochemical functions of proteins as accurate as possible.

Today, there exist more than fifty computational tools for predicting the potential biochemical functions of proteins (Radivojac *et al.*, 2013); most of these try to annotate the function of a protein by detecting sequence similarity (sequence homology) to already characterized proteins from databases, such that already known functions can be transferred (Friedberg, 2006; Lee *et al.*, 2007). Homology-based methods make use of prior knowledge about the biochemical functions of similar sequences, and are less useful for the *de novo* prediction of functions. They are not suitable to annotate sequences of protein families that were not already characterized before. Thus, there is a need for homology-independent methods to explain the functions of novel sequences. In particular, this is the case for metagenomics (Teeling and Glockner, 2012), because metagenome studies typically target weakly characterized species in microbial communities that are likely to have novel and unique cellular activities.

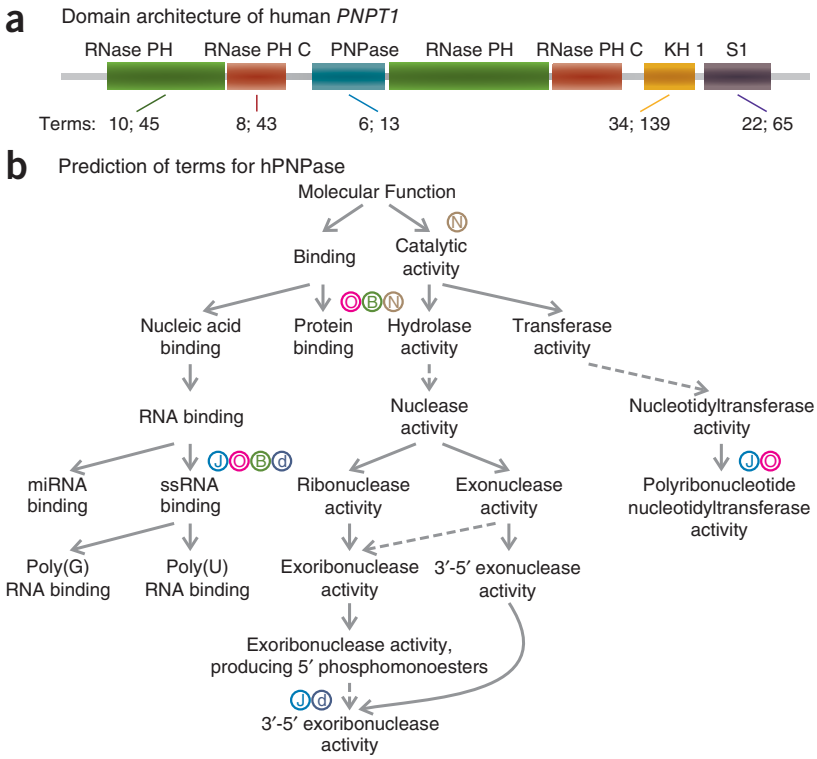


Figure 2.3: GO-annotation of the human mitochondrial polynucleotide phosphorylase 1 (PNPT1) gene. (a) Domain architecture of the human PNPT1 gene according to the Pfam classification. **(b)** Human PNPase belongs to a family of exoribonucleases, which hydrolyze single-stranded RNA. In complex with other components of the mitochondrial degradasome, hPNPase mediates the translocation of small RNAs into the mitochondrial matrix. It is also proposed to be involved in several biological processes including cell-cycle arrest, cellular senescence and response to oxidative stress. PNPT1 serves here as an example for the complexity of protein functions, which can be described on different levels of detail. (Image source: Radivojac *et al.* (2013))

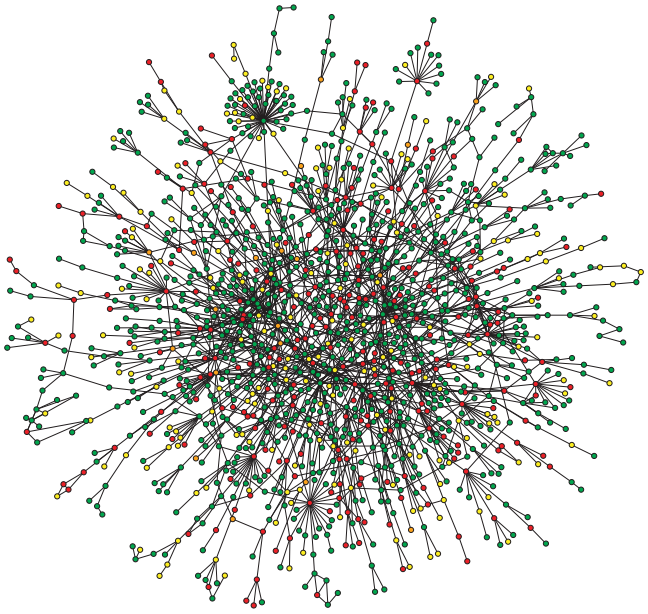


Figure 2.4: Yeast protein interaction network. A map of protein-protein interactions in *Saccharomyces cerevisiae* ('*baker's yeast*'). (Image source: Barabasi and Oltvai (2004))

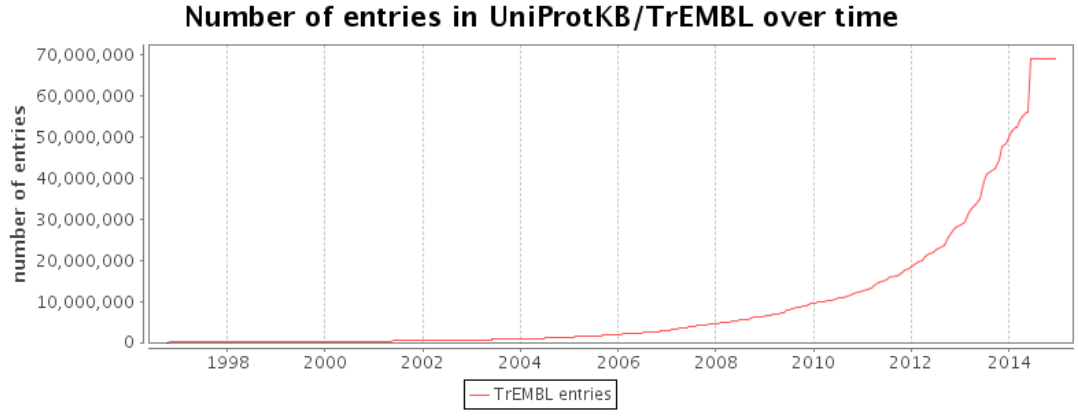


Figure 2.5: The number of deposited protein sequences has drastically increased in the course of the last ten years.

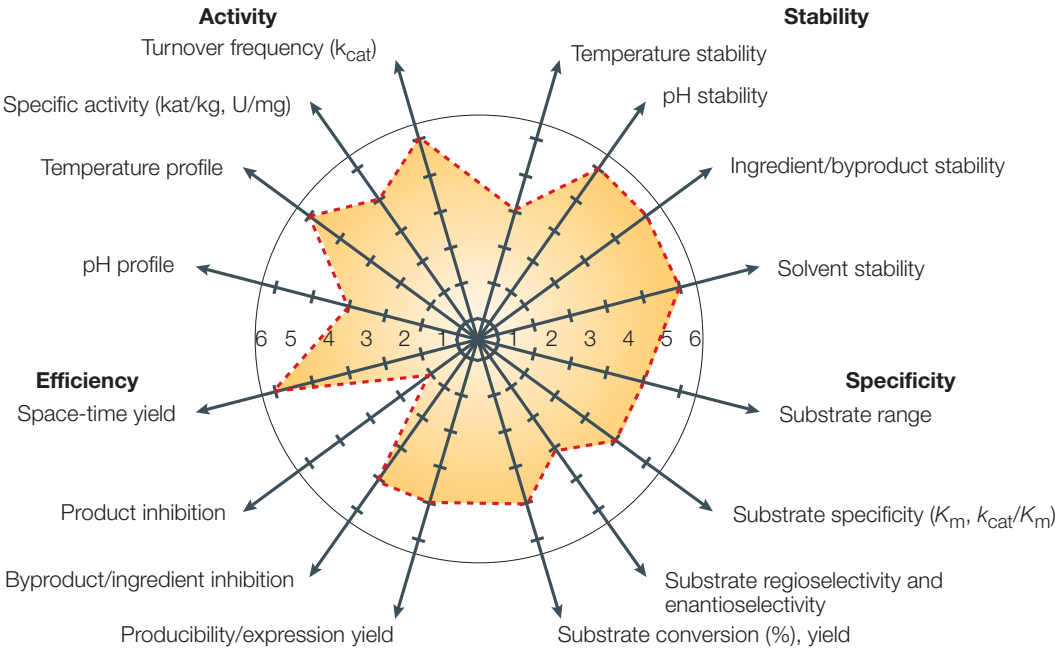


Figure 1 | **Multi-parameter footprint analysis.** This figure illustrates the ideal biocatalyst concept. Each enzyme candidate from the metagenome is ranked, from low (rating of 1) to high (rating of 6) using a specific set of criteria, to produce a multi-parameter fingerprint (shown in yellow). Criteria include *in vitro* enzyme activity, efficiency, specificity and stability. This decision matrix reveals the strengths and weaknesses of every candidate enzyme, so that the most promising candidate enzymes from diverse enzyme libraries can be selected for further process development by re-screening, protein engineering or directed evolution methods. kat, catalytic reaction rate; k_{cat}, catalytic constant; K_m, Michaelis constant; U, unit.

Figure 2.6: Criteria for the selection of biocatalysts. Functional genomics provides the DNA sequences of potential biocatalysts. In the laboratory, candidate biocatalysts need to be tested with respect to the requirements and conditions of a specified industrial process. The procedure is referred to as functional screening of candidate libraries. (Image source: Lorenz and Eck (2005))

Automated protein function prediction is a competitive field of research that has achieved a remarkable progress in recent years (Radivojac *et al.*, 2013). In combination with high-throughput functional genomics and curation efforts of experts from microbiology, it was possible to annotate a large fraction of the known protein space, at least with basic biochemical functions or memberships to conserved families. Unfortunately, automated prediction methods seldomly achieve levels of annotation quality that would be needed for a targeted selection of novel biocatalysts. Therefore, automatically

generated annotations usually represent basic biochemical functions only, and the functional annotations need to be refined by either laboratory experiments or by the use of specialized bioinformatics tools that are designed to target biocatalysts of interest.

Jaroszewski *et al.* (2009) reported in 2009 that around 30%-40% of the available sequences were classified as so-called 'hypothetical proteins' for which no function is known. Even well-studied model organisms like *Escherichia coli*, which has been under intensive investigation since the first publication of its genome sequence in 1997, possess 40% of genes that have no experimentally defined functions (Saito *et al.*, 2010). This sheds lights on the boundaries of our knowledge about the functions of proteins, even though their gene sequences are known.

CHAPTER 3

Background on the computational study of functional modules

This chapter describes existing methods for the computational discovery of protein-protein interactions and functional modules, as well as the challenges that need to be faced.

3.1 Overview

Two categories, *informed* and *de novo* methods, can be distinguished from another. Informed methods analyze the characteristics of known pathways in order to identify (groups of) proteins with similar features in newly analyzed organisms (Section 3.4). This is a form of knowledge transfer based on the recognition of known patterns. Therefore, informed methods do not perform well in the detection of novel pathways

in species with a unique metabolism. By contrast, *de novo* methods do not require *a priori* knowledge about specific pathways; they are designed to detect general patterns shared by the elements of a biological process. Such patterns are thought to be caused by the existence of functional dependencies between the proteins of a process, an implicit assumption in the field that lead to two related hypotheses about the structural cohesiveness/modularity of cellular processes (Section 3.3). Most of the existing *de novo* methods can be summarized under the term genomic context methods (Section 3.5). A specific example for genomic patterns that indicate functional dependencies between proteins are the joint occurrences of gene families across genomes. This type of pattern is the target of phylogenetic profiling methods, and also represents the central information source for the methods developed in this thesis. Therefore, the general challenges for phylogenetic profiling, and the previously obtained results with these methods are discussed in more detail (Section 3.6).

No matter how functional modules are predicted, the functional contexts that they represent can be used to complement sequence homology-based methods in the annotation of protein functions (Section 3.7). This assumption is grounded in the 'guilt-by-association principle' (Aravind, 2000), which therefore represents an additional motivation for the study of functional modules.

3.2 Granularity

The computational analysis of biological processes is a strongly diversified field of research, owing to the complexity and diversity of cellular networks as well as the different possible levels of detail in the descriptions of molecular functions (Raman and Chandra, 2009). For some biological processes, it is possible to precisely describe the involved reaction steps (e.g. considering the relevant thermodynamic constraints). However, this requires precise measurements, and prior knowledge about these aspects.

There exist sophisticated models of metabolic networks and regulatory circuits for a few model organisms (Durot *et al.*, 2009), but for the majority of genomes, and,

in particular, for metagenomes, a profound lack of basic knowledge prevents such applications (Raes and Bork, 2008; Vieites *et al.*, 2009). Flux balance analysis (FBA) is an example of models that require a sufficient level of genome annotation quality, as it needs the stoichiometric coefficients of the metabolic reactions as input (Raman and Chandra, 2009). Whole-cell computational models have also been designed for a couple of model organisms (Karr *et al.*, 2012; Tomita, 2001).

The majority of available annotations for (meta-)genomes comprise a broad functional characterization of the inventories of gene products, but only very few provide enough information for more sophisticated modeling approaches such as flux balance analysis. The reason is that DNA sequencing has more quickly become a cheap high-throughput method than metabolic analysis techniques. However, we can reduce the level of detail, and create more abstract models of processes. This allows for a gradual reduction of the complexity. Of course, we end up with a loss of information, but as long as we manage to retain the most relevant aspects, it will then become possible to fit our models with the available data, even if the measurements in our experiments were not very precise.

The definition of functional modules as sets of proteins that are underlying cellular processes does not cover the details of biochemical reactions, but it represents a useful compromise in practice. We can thus hope to detect functional modules by analyzing the protein repertoires of organisms, and studying the protein-protein interactions within a potential functional module can pave the way for targeted analyses in the laboratory.

The following discussion concentrates on methods which are assumed to operate on sets of proteins without considering precise mechanisms of protein-protein interactions, or the sequential layout of metabolic pathways.

3.3 Evolutionary and functional cohesiveness of functional modules

Most of the methods for studying cellular processes implicitly build on the concept of modularity, which states that elements of a certain group (that is, a module) are

tightly linked with each other, but less related with elements from other groups (Wagner *et al.*, 2007). The aspect of close relatedness within a group is also termed cohesiveness. In 1999, Hartwell *et al.* published an article discussing the modular organization of functions in cells. Since then, there has been a general conviction that cellular molecules organize into functional modules, where they act in concert to realize higher-level biological processes that cannot be ascribed to any of the involved molecules individually (Hartwell *et al.*, 1999). This synergetic effect is assumed to impose dependencies between the members of a functional module. As a consequence, one can expect to see patterns of modularity in cellular networks (Rives and Galitski, 2003), and in the evolutionary patterns of gene families across genomes (Cordero *et al.*, 2008; Martin *et al.*, 2003).

Modularity in cellular networks

Various studies have analyzed the modularity assumption for the topology of cellular networks (Barabasi and Oltvai, 2004; Kreimer *et al.*, 2008). It has been shown that protein-protein interaction networks can, in large parts, be decomposed into local regions of densely-connected proteins that are likely to reflect groups of proteins from the same biological processes (Altaf-Ul-Amin *et al.*, 2006; Pereira-Leal *et al.*, 2004). Furthermore, Ravasz *et al.* (2002) have pointed out that modules seem to be organized in a hierarchical manner, especially in metabolic networks. Hierarchical relationships increase the complexity of the networks, and hamper the unambiguous delineation of module boundaries within the networks (Barabasi and Oltvai, 2004).

Evolutionary cohesiveness

The second aspect of modularity refers to the assumption of evolutionary cohesiveness of the proteins that constitute a functional module. Evolutionary cohesiveness means that the evolutionary histories of the proteins of a functional module are coupled (Cordero *et al.*, 2008; Snel and Huynen, 2004). As a result, functional modules should be identifiable by conserved patterns across genomes, that is, one would expect to find either a large fraction (presence of the module), or a small fraction of the module's

proteins (absence) in the genomes of species (Snel and Huynen, 2004). The expectation of such patterns can be intuitively motivated. In order to be able to perform a certain process, all proteins of this process need to be encoded in the genome of an organism. If, however, some of these proteins get lost during evolution, e.g. due to mutation events, the process would get disturbed and finally inactivated. This consequence would likely cause the successive loss of the remaining proteins of the functional module, in the evolutionary descendants of the organism, because maintenance of the now dispensable genes would not be beneficial for the organisms anymore.

Several studies have put the assumption of evolutionary cohesiveness to the test (Peregrin-Alvarez *et al.*, 2003; Snel and Huynen, 2004; Yamada *et al.*, 2006). They examined the conservation of groups of genes whose products constitute functional modules across microbial genomes. Strikingly, the results have indicated that functional modules corresponding to known pathways are considerably less conserved than previously thought (Peregrin-Alvarez *et al.*, 2003; Snel and Huynen, 2004). In their landmark study, Snel and Huynen analyzed the conservation patterns for nine different collections of functional modules. The collections came from different data sources (manually curated pathways, results of high-throughput experiments, computational predictions), and represented different types of modules, that is, metabolic pathways, protein complexes, and transcriptional modules. The modules of this heterogeneous dataset showed a great variance in the degree of their evolutionary cohesiveness. Only half of the modules were significantly more cohesive than would be expected from a random group of genes. It seemed as if typical functional modules might only have a moderate degree of evolutionary cohesiveness. However, Snel and Huynen had pooled data from various resources, including computational predictions of modules. Their estimate of the mean cohesiveness of modules needs to be considered with caution. It should not be generalized due to the different information sources used for the definitions of the modules, the heterogeneity of their types, and the differences in the associated confidence levels. For instance, the majority (74%) of metabolic pathways from the EcoCyc database were shown to be significantly cohesive, whereas this was only the

case for a minority (35%) of the analyzed KEGG pathways, and for a small fraction of computationally predicted transcriptional clusters (13%). The differences with respect to the two metabolic pathway databases are likely due to the differences in their definitions of pathway boundaries; KEGG pathways are typically larger than those of EcoCyc, and represent compositions of related processes which would be defined as single entities in the EcoCyc database (Section 2.6). Thus, the larger sizes of KEGG pathways as well as their composite structures possibly explain the weaker conservation patterns.

In any case, the study by Snel and Huynen revealed that a large proportion of the known metabolic pathways and protein complexes in *E.coli* and yeast correspond to groups of genes that are significantly conserved across genomes. In line with this finding, Glazko and Mushegian (2004) reported the successful computational recovery of metabolic pathways from an analysis of gene distribution patterns across multiple microbial genomes. However, the authors also noted that most of the pathways could be recovered only partially. Thus, rather than being evolutionary cohesive as a whole, many of the known cellular processes seem to decompose into cohesive sub-modules and less cohesive parts, which was similarly observed in a study by Yamada *et al.* (2006).

The evolutionary age of cohesive modules

Further insights into the cohesive evolution of functional modules have been provided in a study by Campillos *et al.* (2006). The authors used a phylogenetic model based on maximum parsimony to track the evolutionary histories of individual modules across a reconstructed phylogenetic species tree. In total, more than 400 cohesive functional modules could be identified based on the estimated frequencies of joint gene gain/loss events in the tree. The authors classified the set of modules into three different age classes, that is, ancient, intermediate and young modules. It was known that cellular processes of central housekeeping functions are well conserved across microbial species, and possibly represent features that were already present in the universal ancestor (Peregrin-Alvarez *et al.*, 2003). Therefore, it is an interesting result that cohesive modules were not exclusively ancient, but spanned all three age classes with similar

proportions. The authors noted that many of the young modules, which occur in only a few species, and of the intermediate modules might represent a variety of complex functions still to be discovered.

Diversity of the metabolism

It is an important observation that metabolic enzymes tend to be conserved in evolution, but that the composition of metabolic pathways can vary significantly across different organisms (Peregrin-Alvarez *et al.*, 2003). This reflects the flexible nature of the cellular metabolism that provides many alternative paths leading to the same effects (Henry *et al.*, 2010). Consequently, many traditional metabolic pathways do not seem to be evolutionary cohesive as a whole (Spirin *et al.*, 2006; Yamada *et al.*, 2006). In particular, pathways of the central metabolism, such as glycolysis or the TCA cycle, were found to possess flexible evolutionary patterns (Peregrin-Alvarez *et al.*, 2003) that make it difficult to identify them by searching for conserved modular patterns (Spirin *et al.*, 2006).

Additional evidence for evolutionary modularity

In summary, the combined results of studies suggest that the hypothesis of evolutionary cohesiveness can in fact be a guideline for the discovery of functional modules, although they also shed light on the limitations of this assumption.

Apart from this, the cohesive evolution of modules is also indicated by the fact that neighboring genes in genomes often comprise genes with related functions (Rocha, 2008). Genes thus tend to form clusters in genomes, also called operons, which facilitates their joint expression. According to the selfish-operon hypothesis (Lawrence and Roth, 1996), the compact organization of genes allows the integration of the members of a functional module into a contiguous fragment of DNA, thus forming a mobile genomic element which can be transferred between organisms by horizontal gene transfer (HGT) (Fondi *et al.*, 2009). Mechanisms like this may explain the distribution patterns of evolutionary young modules, as pointed out by Campillos *et al.*.

3.4 Informed methods

Informed methods take the proteins of a known functional module from a model organism as input, and search for their homologs in other species. The module gets predicted to exist in the genome of an organism if the genome contains genes for most of the proteins that make up the module. In this case, it is likely that the necessary functions needed to carry out the cellular process are encoded in the genome. An example for the use of this principle is the PathoLogic algorithm (Paley and Karp, 2002), which can be used to map the content of the MetaCyc pathway database to the genomes of new species (Caspi *et al.*, 2008).

As public databases nowadays store several hundreds of well characterized metabolic pathways, these resources have also become useful for data mining and machine learning applications. The contents of these databases can serve as the input for statistical learning methods which allow for the extraction of general pathway characteristics such as the average number of pathways per organism, or the average number of proteins that belong to the metabolic network of an organism. Dale *et al.* (2010) compiled a comprehensive list of 123 characteristic features from the pathway set of the MetaCyc database, and used them for a supervised training of pathway prediction models that can be applied to genomes. Interestingly, the authors reported that a small number of features carried most of the information of whether a pathway is present in an organism. Once more, it should be noted that an approach of this type represents a form of knowledge transfer. In the study of Dale *et al.*, the strong dependence on *a priori* knowledge about the target pathways became obvious by the observation that the number of known enzymes that could be re-identified in the target organism was one of the main determinants for the accuracy of pathway prediction. Moreover, the example also illustrates the importance of an accurate prediction of the enzyme set of the target organism as a prerequisite for the success of the methods.

Taken together, informed methods can be used to detect known pathways in new species. They are less useful for the analysis of species with a new and unique metabolism.

In particular, their application to metagenomes is limited in cases where one can expect unknown species in the microbial communities, and because of the partial and incomplete nature of annotations that are typically available for metagenomes.

3.5 The importance of genomic context for *de novo* predictions

Various *de novo* methods can be used to predict functional interactions between proteins (Lee *et al.*, 2007). They are commonly referred to as genomic context methods (GC-methods). The methods differ in the types of signals that they use as input, but they all implicitly rely on the modularity assumptions for functional modules (Section 3.3). The different genetic patterns analyzed by these methods are all thought to be caused by the close interactions of the proteins that are jointly involved in higher-level functions of the organisms. This becomes most obvious in the case of methods detecting gene fusion events. It is known that genes whose products have very closely related functions show a tendency to be fused into a single gene instance during the course of their evolution. Thus, the detection of gene fusions is a strong indicator for the functional coupling of genes (Enright *et al.*, 1999). The physical fusion of genes is an extreme form of shared genomic context, but it turns out that wider definitions of a shared context are useful as well. The proximity of neighboring genes in genomes is also an indicator for functional relationships, and can be analyzed by GC-methods (Overbeek *et al.*, 1999). Finally, phylogenetic profiling methods go one step further, and take a coarse perspective on genomic context. In phylogenetic profiling, conserved co-occurrence patterns of gene families, which are observed across multiple species, are considered to be a sufficient indicator for functional relationships (Pellegrini *et al.*, 1999).

While all the aforementioned types of methods use genomic patterns as input, it may of course appear most intuitive to predict functional interactions from joint patterns of gene activities. And indeed, so-called co-expression methods exist which link genes by finding similar patterns in their expression profiles across various physiological conditions

(van Noort *et al.*, 2003).

It is known that the accuracy of GC-methods suffers from high rates of false positive predictions (Karimpour-Fard *et al.*, 2008; Muley and Ranjan, 2013). In particular, the prediction of metabolic interactions continues to be a challenging task for all types of methods (Muley and Ranjan, 2013). By combining the predictions of different methods, the quality of the results can be improved (von Mering *et al.*, 2003b). With respect to this, it should be noted that gene expression data are usually not available for the analysis of metagenomes, because high-throughput gene expression measurements have so far only been done for a few metagenomic datasets (Simon and Daniel, 2011). Similarly, the detection of gene clusters in metagenomic datasets is less effective than in genomic datasets due to a high level of DNA sequence fragmentation even after the assembly stage. Thus, the analysis of metagenomes with genomic context methods is more difficult.

GC-methods typically focus on the identification of pairs of interacting proteins. Groups of functionally related proteins can then be derived in a successive step from the individual interactions by means of clustering. It is thus possible to use genomic context methods for the discovery of complete biological pathways, but only few studies have addressed their benefits and limitations for this purpose (Muley and Ranjan, 2013). Previous applications of phylogenetic profiling to the recovery of functional modules will be discussed in Section 3.6.3.

The STRING database

The STRING database is closely related to genomic context methods, and represents a repository for known and predicted protein-protein interactions (von Mering *et al.*, 2005). Each interaction is defined as a pair of proteins that is associated with individual evidence scores for different information sources, including genomic context methods as well as text mining results from literature analysis, experimental data, and known pathways from public databases. The scores reflect the confidence that a particular type of evidence truly indicates a functional relationship (von Mering *et al.*, 2003a). The scoring scheme is based on benchmarks on pathways from the KEGG database, and was designed in such a way that it can be used in a consistent way for all different types of information sources, which also allows to summarize the individual evidence values in a combined score.

The information maintained in STRING can be displayed in a network graph, where nodes represent proteins, and edges represent pairwise interactions that are associated with predicted confidence scores. The resulting graph is called the STRING functional interaction network (Figure 3.1). Notably, we used high-confidence edges of the STRING network for the validation of one of our methods (Section 5.2).

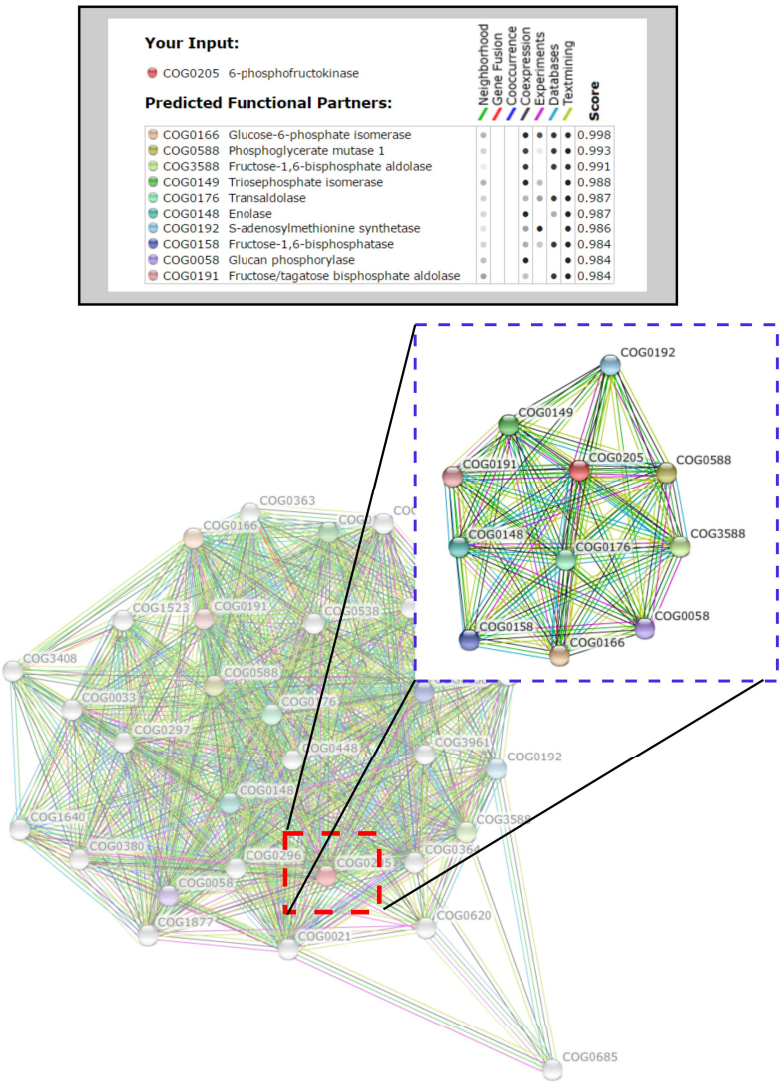


Figure 3.1: A partial view of the STRING functional interaction graph. The web interface of STRING was queried with the gene family identifier 'COG0205' (representing a central enzyme of the *glycolysis* metabolic pathway, *6-phosphofructokinase*). The query result displays the embedding of 'COG0205' into the STRING functional interaction graph, and the number of displayed surrounding interactions can be controlled by a parameter. The image was arranged to show the interaction network around 'COG0205' at two different zoom levels. Each colored line in the graph represents evidence for a pairwise functional interaction between COGs – based on different evidence sources including genomic-context methods, integrated information from pathway databases, and classical text mining results (conducted on abstracts of scientific articles). (Image source: Composition of screenshots taken from `http://string-db.org`)

3.6 Phylogenetic profiling

Phylogenetic profiling is a genomic context technique that attempts to analyze the co-evolution patterns of gene families by comparing the similarity of their presence / absence profiles across a set of different organisms.

As described in the motivation of this thesis (Section 1.3), the analysis of co-occurrence patterns of gene or protein families is at the heart of the thesis. This section describes the underlying assumptions which are closely related to the modularity of cellular processes (Section 3.3). Subsection 3.6.2 discusses challenges and problems that may affect the results of phylogenetic profiling. Finally, Subsection 3.6.3 provides a summary of the state of the art in phylogenetic profiling.

3.6.1 The principle of phylogenetic profiling

The underlying hypothesis for phylogenetic profiling is that functionally linked proteins are likely to evolve in a correlated fashion (Pellegrini *et al.*, 1999). Gaasterland and Ragan, and Pellegrini *et al.* are commonly recognized as being the first to demonstrate the validity of this assumption - with their studies being published in the years 1998 and 1999, respectively. Since their initial findings, a multitude of related methods have been released which all build on the same basic assumption of correlated evolutionary histories (see Kensche *et al.* (2008) for a comprehensive review).

The phylogenetic profile of a protein is often represented as a numerical vector that encodes the presence, or absence of the protein in a list of genomes. Pellegrini *et al.*, and many of their successors used binary profiles, assuming that the existence of a protein in all genomes can be detected without uncertainty. However, this is usually not the case. For a given protein, homologous sequences in other organisms can be detected but without strong guarantees that the identified sequences possess the same molecular functions (Lee *et al.*, 2007). In most cases, however, the degree of sequence homology is a sufficient measure of confidence for the congruence of functions (Tian and Skolnick, 2003), and, therefore, similarity scores have been used to define real-valued entries of

phylogenetic profiles in an attempt to model the uncertainty of homology-based function transfer more accurately (Marcotte, 2000). A real-valued phylogenetic profile for a given protein of a target organism is constructed by finding the most similar protein sequence within each of a list of reference genomes, and using the obtained similarity scores for defining the entries of the profile. However, even with techniques like this, phylogenetic profiles can only represent coarse approximations of the true evolutionary histories, taking into account the many factors that affect the evolution of gene families in microbial genomes. For this reason, more sophisticated methods of phylogenetic profiling have been developed that measure the correlation of the evolutionary rates of different gene families based on multiple sequence alignments and reconstructed phylogenetic trees ('mirror-tree method') (Goh *et al.*, 2000; Pazos *et al.*, 2005; Pazos and Valencia, 2001).

Co-occurrence profiling methods search for pairwise dependencies between proteins, and various measures for assessing the pairwise similarity (or distance) of phylogenetic profiles have been tested (Kensche *et al.*, 2008). In the strictest form, a perfect match of the profiles is required (Liberles *et al.*, 2002; Yanai and DeLisi, 2002), but the majority of profiling methods tolerate a certain amount of discrepancies between the two profiles - using different L_p-norm measures such as the Hamming or Euclidian distance, or statistical correlation measures such as Pearson's correlation coefficient, or the mutual information (Kensche *et al.*, 2008). However, independent of the actual choice of a measure, the limited focus on pairwise relationships can potentially lead to failing the detection of more complicated functional relationships between groups of proteins (Bowers *et al.*, 2004; Kensche *et al.*, 2008). For this reason, Bowers *et al.* proposed a formalism based on Boolean logic that can model all possible relationships between the members of a triplet of proteins, such as, for example, the dependency of one protein on the joint presence of two other proteins in a genome. In a successive approach, Zhang *et al.* (2006) proposed a framework for the analysis of relationships on quartets of proteins, thus enabling the study of even more complex dependencies. Such relationships can, for example, arise from the existence of pathway variants, where

some of the involved proteins have weaker conservation patterns than the rest of them (Kensche *et al.*, 2008). As a consequence, new methods for functional module mining should be targeting complex relationships between whole groups of proteins instead of isolated pairwise interactions.

Interestingly, the principle of co-occurrence profiling can be applied to different traits of organisms. Apart from detecting pairs of interacting proteins, it was also used to analyze functional relationships between subdomains of proteins (Pagel *et al.*, 2004), and for the identification of protein families whose phylogenetic profiles correlate with the presence/absence patterns of phenotypic traits across different organisms (Liu *et al.*, 2006; Slonim *et al.*, 2006).

3.6.2 Challenges

As described in Section 3.3, co-occurrence based methods implicitly assume the evolutionary cohesiveness of functional modules. However, the evolution of genomes is a complex process, and various factors can disturb the patterns of co-inheritance reflected in the gene content of the genomes of different organisms (Glazko and Mushegian, 2004). Genomes are shaped by processes like gene loss, gene duplication, horizontal gene transfer, and gene genesis, and the contents of gene repertoires change quite dynamically during evolution (Rocha, 2008; Snel *et al.*, 2002). There can even be significant differences between the gene repertoires of strains of the same bacterial species (Achtman and Wagner, 2008). To address these evolutionary mechanisms more adequately, 'model-based phylogenetic profiling' methods infer functional relationships between proteins based on joint gains or losses of genes along the branches of reconstructed evolutionary species trees (Kensche *et al.*, 2008). For instance, Barker and Pagel (2005) developed a maximum likelihood model, and concluded that pairs of proteins with at least two to three correlated events of gain or loss are almost certainly functionally linked.

Phylogenetic relationships between species

Despite the high dynamics of genome evolution, the gene content of closely related species is often highly correlated, meaning that most of the genes will have similar phylogenetic profiles across these species which can lead to false assumptions about their functional relationships (Barker and Pagel, 2005). Species that are very closely related might not have diversified much from each other during evolution, leading to large overlaps of their gene contents, and increased numbers of gene co-occurrence patterns. By consequence, a significant fraction of the observable co-occurrence patterns are spurious with respect to phylogenetic profiling; they do not reflect functional dependencies between genes, and should not be interpreted as indicators for functional relationships.

In line with this, Sun *et al.* (2005), and Jothi *et al.* (2007) reported from their studies that the selection of input genomes plays a critical role for the application of pairwise phylogenetic profiling methods. Their results indicated that a phylogenetically diverse

set of not too closely related reference genomes represents the best choice of input. As of today, most applications of phylogenetic profiling were operating on rather small input sets of genomes, and it has been concluded by Sun *et al.* that the addition of phylogenetically redundant genomes might contribute noise signals to the phylogenetic information. If this is the case, methods for functional module mining should be designed to be robust against noisy input data, especially if they are intended to be used on large datasets of genomes. This is one of the reasons why we decided to use Bayesian probabilistic models for the detection of functional modules from co-occurrence patterns, as this class of models is known to be suited for the analysis of noisy datasets (Friedman, 2004; Wilkinson, 2007).

Notably, a more recent large-scale study by Muley and Ranjan (2012) indicated that the choice of genomes could be less critical than suggested by Jothi *et al.*. The study was based on an input set of 565 prokaryotic genomes, and the authors observed no benefits for the performance of the method from subsampling the genome set. Nevertheless, the phylogenetic relatedness of species can cause problems for the analysis, especially on small datasets. The aforementioned model-based profiling methods try to solve the problem by explicitly considering the evolutionary histories of gene families. But model-based methods are difficult to apply for the analysis of metagenomes, where the phylogenetic origins of DNA sequence fragments are often unknown, and the reconstruction of phylogenetic trees is difficult.

Gene duplications and orthologous groups

A particular challenge for co-occurrence based methods is the sub- or *neo*-functionalization of gene families after gene duplication events (Kensche *et al.*, 2008). Due to specific molecular mechanisms, it is possible that a gene gets duplicated in a genome, leading to two identical copies of the original gene that can perform the same molecular functions. The redundancy of the functions releases one of the copies from selective pressures that might have existed for the original gene. Thus, even if the original gene function was essential for the organism, and thus needs to be maintained unchanged, one of the resulting copies might undergo mutational events that will gradually change its molecular functions (Roth *et al.*, 2007). This is a common mechanism for the differentiation of existing gene families into new sub families, or the genesis of completely novel functions.

Gene duplication events lead to the existence of homologous gene sequences with different functions. Therefore, the detection of homologous sequences across genomes can lead to false assumptions about the presence/absence profiles of gene functions, which is a major challenge for phylogenetic profiling methods (Kensche *et al.*, 2008). To overcome this problem, most methods use the concept of orthology instead of homology to profile the existence of genes across species. An orthologous group of genes is defined as a set of genes from different species that have emerged from the same gene in the last common ancestor of the species (Sonnhammer and Koonin, 2002). The detection of orthologous groups, in principle, needs in-depth phylogenetic considerations but there exist several heuristic strategies for large-scale profilings of these groups, and information on the distribution of orthologous genes can be obtained from databases such as COG (Tatusov *et al.*, 1997), and eggNOG (Jensen *et al.*, 2008).

Pathway variants and non-orthologous gene displacement

Two important factors contributing to the diversity of the composition of functional modules are the existence of species-specific variants of cellular processes (Ye *et al.*, 2005), and evolutionary phenomena like non-orthologous gene displacement. Koonin *et al.* describe gene displacements as the phenomenon that identical functions in cellular processes were found to be encoded by genes with non-orthologous DNA sequences in different organisms. Therefore, single reaction steps of pathways can be realized by different gene families in different organisms. Pathway variants, and gene displacements both disturb the conservation patterns of functional modules across genomes (Snel and Huynen, 2004), and make the discovery of cellular processes from phylogenetic profiles more difficult.

Ambiguous assignments of gene families to processes

According to the 'patchwork hypothesis' about the evolution of metabolic systems, ancient enzymes with broad activity were recruited into diverse metabolic processes, where they were optimized in substrate specificity and reaction efficiency in the course of evolution (Caetano-Anolles *et al.*, 2009). Therefore, some families of orthologous genes are expected to be widely spread over different functional modules, an assumption which is supported by the observation that many proteins are capable of performing different functions depending on the cellular context (which is called 'moonlighting' (Jeffery, 1999)). These two factors complicate the unambiguous assignment of gene families to specific functional modules, as there might be several equally plausible possibilities. For this reason, methods used for the recovery of cellular processes should be flexible enough to allow several possible assignments.

Gold standards for evaluation

Existing databases for metabolic pathways, and protein complexes can serve as a basis for the evaluation of the predictions of genomic context methods (Qi *et al.*, 2006). For example, the proteins of well characterized metabolic pathways were used to define gold

standard sets of true ('intra-pathway') and false ('inter-pathway') functional interactions between proteins, and the same principle was applied to collections of multi-protein complexes. This evaluation scheme was used in several studies, and, in particular, in the context of phylogenetic profiling (see Barker and Pagel (2005); Jothi *et al.* (2007); von Mering *et al.* (2003a); Muley and Ranjan (2013) for examples). While this scheme allows assessing a method's accuracy for the recovery of known pathways, it is less suited to assess the quality of predictions that involve proteins which have not been characterized before. Moreover, the accuracy of the predictions is difficult to measure because the available knowledge of experimentally verified interactions is still very limited (Ferrer *et al.*, 2010). There might be many cases where we cannot exclude the possibility that a seemingly falsely predicted interaction between proteins was indeed correctly identified. To obtain adequate estimates for the specificity of the methods, one would also need reliable datasets of protein pairs that are known to be non-interacting. However, such datasets are currently missing, especially for large-scale applications, and the problem thus remains central in the field (Qi *et al.*, 2006). Finally, the observation that, in many cases, we cannot properly distinguish between false positive and true positive predictions, greatly restricts our capabilities to assess a method's potential to correctly predict novel functional modules.

Summarizing the evaluation schemes of different studies, most frequently the KEGG and EcoCyc databases have served for evaluation purposes. However, even KEGG annotations describe only small portions of the gene contents of bacterial genomes (typically less than 30%, depending on the species; see, for example, table 1 in Date and Marcotte (2003)), and thus represent a limited resource of information. An alternative resource for the evaluation of methods is the assignment of gene products to 17 curated functional categories from the COG database that can be obtained for most genes. Overall, COG categories provide a functional context for many orthologous groups of genes (Tatusov *et al.*, 2000); however, they only provide a low level of detail in the descriptions of functions, and thus can only be used for a coarse evaluation of predicted protein interactions (Wu *et al.*, 2006). Finally, the Gene Ontology (GO)

(Ashburner *et al.*, 2000) also provides information on the functional contexts of proteins; however, in practice, only few studies have used this resource for the evaluation of phylogenetic profiling, probably because this annotation system was originally developed for eukaryotic organisms, and its application in the domain of bacterial and archaeal genomes is lagging behind (Mao *et al.*, 2005).

3.6.3 Previous results obtained with phylogenetic profiling methods

Presently, a plenitude of studies have analyzed the various aspects of phylogenetic profiling, and the main conclusion that can be drawn from the results is that this technique is a versatile and useful tool for the prediction of functional interactions, although its coverage is limited to those cellular systems that evolved in a cohesive and modular fashion (Kensche *et al.*, 2008). The review by Kensche *et al.* lists several examples of successful applications of the methods; for example, Altincicek *et al.* experimentally validated a missing metabolic pathway element whose identity was revealed before by Cunningham *et al.* (2000) based on patterns of gene co-occurrences.

In comparison to other genomic context methods, phylogenetic profiling tends to perform worse (Muley and Ranjan, 2013), as could be expected from its coarse perspective on genomic context. Ferrer *et al.* (2010) found a significant performance advantage of gene neighborhood methods. However, Karimpour-Fard *et al.* (2008) have demonstrated in their study that all GC-methods have strengths and weaknesses, and no single method tends to be superior in all relevant aspects. Thus, the different capabilities of the methods are under debate. In any case, the open challenges for all genomic context methods remain a reduction of the typically high rates of false positive predictions, and a more accurate recovery of metabolic interactions (Muley and Ranjan, 2013).

Pairwise interactions

The pairwise similarity of phylogenetic profiles clearly correlates with the functional similarity of the considered proteins (Date and Marcotte, 2003). Moreover, the distance between profiles of enzymes was shown to be proportional to their pathway distance, that is, the number of intermediate reaction steps that separate enzymes in the metabolic network (Yamada *et al.*, 2006). Pairwise phylogenetic profiling is thus useful to identify functional interactions between proteins. However, it was also observed that some functional interactions cannot be correctly identified by pairwise profiling methods (Bowers *et al.*, 2004; Zhang *et al.*, 2006). In line with that, Schneider *et al.* (2013)

described that many co-occurrence patterns of known interactions are disrupted, in the sense that interaction partners show weak patterns of co-occurrence across genomes. They demonstrated that disrupted patterns are likely caused by multifunctional proteins that violate the assumption that all members of a functional module evolve in a correlated fashion – simply because multifunctional proteins have multiple different possible interaction partners, some of which might not belong to the module. This induces complicated dependencies between interacting proteins that cannot be detected by methods operating on single pairs of proteins (Bowers *et al.*, 2004; Schneider *et al.*, 2013; Zhang *et al.*, 2006).

Recovery of functional modules

Most studies focused on the prediction of pairwise interactions between proteins; however, some researchers have grouped proteins into clusters based on the pairwise similarities of their phylogenetic profiles (Cohen *et al.*, 2012; Date and Marcotte, 2003; Glazko and Mushegian, 2004; Li *et al.*, 2009; Wu *et al.*, 2006; Yamada *et al.*, 2006). Nevertheless, the prediction accuracy with respect to the recovery of single metabolic pathways was evaluated in just a few cases (Muley and Ranjan, 2013).

The most frequently used evaluation scheme for GC-methods pools the proteins from several known pathways, or protein complexes into a joint gold standard set. The class of positive examples then gets defined as the set of protein pairs that participate jointly in any of the considered pathways, or protein complexes ('intra-module pairs'). Conversely, negative examples correspond to proteins from different pathways ('inter-module pairs'). The approach provides insights into the fraction of true positive pairs that are covered by the predictions of a method ('sensitivity'), and it also allows for the assessment of a method's capability to distinguish between different pathways ('specificity'). But in this setting, it is obscured to which extent individual processes are recovered. It should also be noted that this evaluation scheme is restricted to proteins that are annotated in the pathway databases, and therefore ignores proteins of unclear functions.

For example, Muley and Ranjan (2013) used this scheme to evaluate the pairwise

predictions of phylogenetic profiling. Their results, among others, proved the decent accuracy of the method with respect to the recovery of pairwise interactions between proteins from KEGG pathways. Moreover, their study also represents one of the few exceptions of a detailed analysis of individual pathways. Muley and Ranjan showed that phylogenetic profiling works best for housekeeping processes such as DNA replication and repair, protein folding and translation, and cell motility, whereas the majority of metabolic processes are difficult to predict. It should be stressed, however, that their evaluation was based on a small set of nine super-pathways, each one summarizing several pathways from a deeper level of the KEGG hierarchy (Kanehisa *et al.*, 2004), and thus the level of detail of the insights is limited.

With respect to all phylogenetic profiling methods reviewed for this thesis, Li *et al.* were the only authors that explicitly commented on the specificity of their method for the recovery of individual pathways. Strikingly, the specificity of their probabilistic clustering method was below 20% for most of the analyzed pathways. This is, of course, a rather disappointing observation from the perspective of functional module mining, and should not be generalized as strongly indicated by two other studies. The clustering approaches used by Date and Marcotte (2003), and Glazko and Mushegian (2004), which are both based on the application of thresholds to the pairwise similarities of phylogenetic profiles, yielded more promising results. The authors successfully recovered many clusters of proteins that closely mapped to known functional modules of cellular processes, although, in both cases, the false positive rates for individual modules were not provided along with the reports.

Overall, the various results of the clustering approaches proved that phylogenetic profiles provide enough information to recover individual functional modules, but there is a noticeable tendency that the recovered clusters only matched parts of known functional modules (Glazko and Mushegian, 2004; Li *et al.*, 2009; Yamada *et al.*, 2006). The latter observation clearly indicates that the tradeoff between the obtainable rates of true positive and false positive predictions seemed to be a general problem for the methods; otherwise, lowering the thresholds for pairwise similarities should allow to increase the

cluster sizes, and thus the coverage of known functional modules. In line with that, Kensche *et al.*, and others have noted that the fraction of true positive predictions quickly drops when the thresholds for pairwise similarity are relaxed.

It remains remarkable that the majority of studies using clustering approaches in fact addressed the recovery of pathways but omitted an evaluation of the proportions of false positives for individual pathways. Probably, this can be explained by the described challenges of defining an appropriate gold standard for the evaluation of functional module recovery (Section 3.6.2).

The capabilities and limitations of co-occurrence based methods to recover functional modules are in line with the observations concerning the general modularity and evolutionary cohesiveness of the different types of functional modules (Section 3.3). The enzymes of metabolic pathways tend to have flexible evolutionary histories, and metabolic pathways are generally harder to predict than well conserved cellular processes of ancient housekeeping functions. Due to the rapid evolution of gene content in microbial species, phylogenetic patterns of functional modules are complex, and this is one of the reasons why there is a demand for improved techniques for phylogenetic profiling (Glazko and Mushegian, 2004).

Scale of applications and biases

With only a few exceptions, previous applications of phylogenetic profiling represented organism-centric studies, in which the functional interactions of the proteins of a given organism were analyzed by comparing the phylogenetic distributions of their homologs (or orthologs) across a reference set of genomes. The numbers of genomes used for profiling were typically small, and comprised at most a few hundred genomes, which was mainly determined by the availability of genomic sequences at the respective time. Hence, it should be noted that the limited information content of small reference sets might have introduced a bias in the results of some studies. For example, Snitkin *et al.* (2006) reported poor results for phylogenetic profiling on eukaryotic genomes; however, they could only do tests on a set of 23 reference genomes.

3.7 Guilt by association

The 'guilt-by-association principle' states that closely interacting proteins are likely to have related functions (Aravind, 2000). Thus, hypotheses about the possible functions of a candidate protein can be derived from its interaction partners once they were identified (Oliver, 2000). Whereas homology-based methods are used to predict the exact molecular functions of proteins, the guilt-by-association principle mainly guides the process-level annotation of proteins (Huynen *et al.*, 2000).

So-called network-based function prediction methods rely on the information content of protein-protein interaction networks to transfer functions between proteins (Janga *et al.*, 2011). A common approach is to locate the position of a protein of unknown function in the network and to use known functions that are enriched in its local neighborhood as candidates for a functional annotation (Schwikowski *et al.*, 2000). Other approaches rely on the assumption that densely connected subgraphs of the network correspond to functional modules (Brohee and van Helden, 2006). Thus, they restrict the transfer of functions to those regions of the network.

Functional modules can be inferred with or without the explicit construction of protein-protein interaction networks. However, any set of functional interactions between proteins can be represented as a network, where nodes represent the proteins and edges represent their functional links (Janga *et al.*, 2011). Therefore, some strategies of network-based function prediction could be adopted to transfer functions between the proteins of predicted functional modules.

It should be noted that the transfer of functions is difficult and the concept of guilt by association is under debate. Gillis and Pavlidis (2012) have recently argued that the most relevant information for the transfer of functions is concentrated on single edges in interaction networks, which makes the approach error-prone because there is a lot of noise in the networks.

In any case, the guilt-by-association principle is a strong motivation for studying functional modules. Osterman and Overbeek (2003) described several examples that

illustrate how results of genomic context methods were used to identify missing genes of metabolic pathways.

CHAPTER 4

Probabilistic topic models and LDA

The key idea for the development of the methods presented in this thesis was to use probabilistic topics models (PTMs) for the task of module inference (Chapter 5). Topic models are normally used in text mining applications; for example, to identify functional relationships between genes by analyzing the joint occurrences of gene names in abstracts of scientific papers (Aso and Eguchi, 2009; Zheng *et al.*, 2006). The novelty of our approach is the transfer of the concept of semantic text analysis to the inference of functional modules from the co-occurrences of gene families across genomes and metagenomes.

Originally, probabilistic topic models were designed to extract a compact representation of the semantic concepts that are underlying collections of texts. The representation is based on the notion of 'topics', which are groups of words that are commonly associated with a higher semantic meaning (e.g. 'sports', 'music', etc.). We can intuitively interpret 'topics' as the semantic concepts that an author bears in mind, when he starts

to compose text. During the process of writing, documents are getting composed of a series of words, and we will assume that words belong to a vocabulary of fixed size. With this picture in mind, it becomes evident that the way how topics are shaping the contents of documents will be reflected in the co-occurrences of words across many texts. Vice versa, we can exploit the patterns of co-occurring words to infer a model of the interplay of topics (Steyvers and Griffiths, 2007). Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003) is a widely used Bayesian topic model suitable for this purpose.

Latent Dirichlet allocation

The LDA model assumes a well-defined generative process that describes the relationships between the latent topics and the word patterns in the documents (Appendix B). This process can be formulated as a graphical model. In general, a probabilistic graphical model is defined by a joint probability distribution over a set of random variables, which describe relevant aspects of the real-world problem domain (Koller and Friedman, 2009). In case of the LDA model, the variables describe the latent, that is, unknown groupings of words into topics. A key concept of graphical modeling is the assumption of conditional independence between subsets of variables, which allows a simplified factorization of the complex joint distribution, and may significantly reduce the computational costs of learning the model. The term 'graphical model' refers to the fact that the joint distribution can be visualized in a compact diagram, in which the dependencies between variables are made explicit. For instance, Figure B.3 shows the graphical model of LDA in the so-called 'plate notation', which is a specific type of diagram that supports the ease of model interpretability.

Even with the factorized representation of a model's joint probability distribution, the task of parameter inference is computationally demanding for complex models. Exact inference algorithms for computation of the optimal parameter settings exist, but these typically require infeasible computational efforts in practice. Approximate techniques such as Markov Chain Monte Carlo (MCMC) sampling can be used instead,

e.g. Gibbs sampling (Gilks *et al.*, 1999). In practice, these methods are robust and efficient procedures for Bayesian inference of complex probabilistic models.

Once the model's joint probability distribution has been inferred with MCMC methods, we can draw conclusions by analysis of the model. In particular, the posterior probability of latent variables conditioned on observed variables can be analyzed. Considering LDA, MCMC methods are efficient means of deriving estimates of the topics that are underlying a given input collection of texts.

An in-depth explanation of the concept of Bayesian inference and the LDA model is included in the Appendix of this work (Appendix A and Appendix B). For additional details on these topics, the reader is referred to an excellent technical report by Gregor Heinrich (Heinrich, 2009).

Part II

Inference of functional modules

CHAPTER 5

A new topic model based method for functional module inference

The main goal of the PhD project was to develop a machine-learning based framework capable of detecting characteristic evolutionary patterns of cellular processes across the large genomic and metagenomic datasets already available in public databases.

It was already known from the results of phylogenetic profiling that the analysis of gene co-occurrence patterns, in principle, enables a *de novo* discovery of functional modules. In Konietzny *et al.* (2011), we demonstrated the use of probabilistic topic models for large-scale analyses of functional modules encoded in genomes and metagenomes.

5.1 Adaptation of topic models to genomic data

The key to the adaptation of topic models is to establish an analogy between text documents and the genomes of microbial species. We can interpret the gene repertoire of a given genome as an equivalent to a set of words that make up a text document. More precisely, let $d_T = (w_1, \ldots, w_N)$ be a text document that is composed of N words w_i for $i \in [1, N]$. Each word is an instance of a vocabulary term $v \in V$ from a controlled vocabulary V. In an analogous fashion, we can define $d_G = (g_1, \ldots, g_N)$ to represent the genes g_i of a microbial genome that contains N gene sequences.

If we want to represent more than a single genome, we have to make sure that we can refer to genes in a consistent way. For that purpose, we can use a controlled vocabulary (Section 2.5) of functional descriptors (FDs, for short – for example, gene family identifiers, or EC numbers) to consistently annotate gene sequences across various genomes. Figure 5.1 illustrates this principle for gene families or protein domains, respectively: The orange genes (shown as arrows) represent members of a common gene family that occur in different genomes. They can be consistently referred to by the identifier of their family.

We thus need to distinguish[1] between the identifier of the family (vocabulary term), and its corresponding annotations in the genomes (term instances). In other words, if the gene family with the identifier FD is annotated for different genomes, each individual annotation can be addressed as a pair (FD, d), where d is the identifier of the respective genome in the input collection.

Using functional descriptors, we can now substitute the document $d_G = (g_1, \ldots, g_N)$ – representing the genes of a genome – by $d'_G = (FD_1, \ldots, FD_M)$, that is, a new document that is composed of functional descriptors FD_j for $j \in [1, M]$. In general, N and M (the lengths of d_G and d'_G) can differ because we do not assume a one-to-one-correspondence between gene sequences and functional descriptors. It might be the case that some gene

[1] Note the general distinction between terms $v \in V$ of the vocabulary, and so-called term instances which are the occurrences of the vocabulary terms in individual documents (e.g. $w_i = v$ for a word w_i of a specific document).

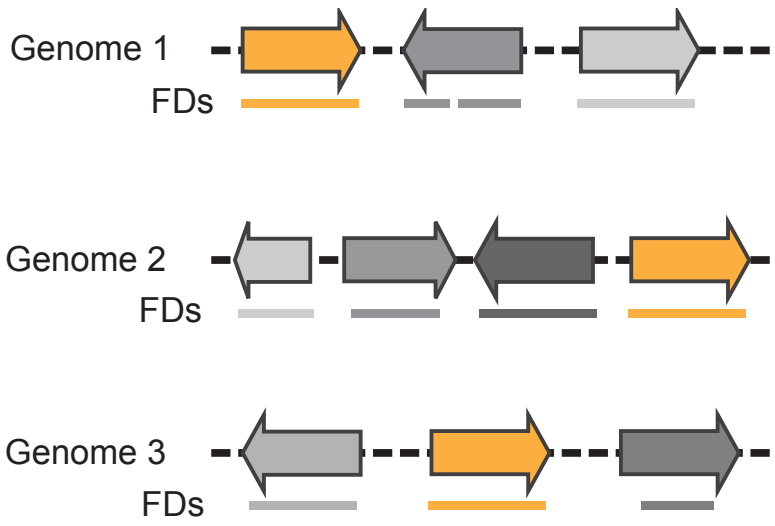


Figure 5.1: Genes can be consistently annotated across genomes with functional descriptors. The arrows with bars underneath shall represent genes with annotated functional descriptors (FDs), such as gene family or protein domain identifiers. The FDs stem from a controlled vocabulary. Note that the actual sizes of the genomes are much larger (indicated by the dashed lines). The *orange* genes are annotated with the same functional descriptor in three different genomes. We assume that they belong to the same gene family.

sequences cannot be annotated, or that a sequence was annotated with more than a single functional descriptor (for example, in the case of protein domain annotations).

The proposed analogy between the text document d_T, which is composed of words from a natural language, and the 'genome document' d'_G is adequate because *i)* we are using controlled vocabularies in both cases, and because *ii)* topic models like LDA do not consider the formal syntax, that is, the inner structure of input documents. LDA is built on the so-called 'bag of words assumption' that considers documents as unstructured containers for words. Essentially, the order of words in a text is assumed to be unimportant, and therefore the document can be summarized by frequency counts for individual words. The gene repertoires of organisms can be summarized in a similar fashion.

Genes in genomes, and words in texts represent basic functional units. In both cases, higher-level meanings emerge from combinations of the basic units; i.e. concerted functional activities of proteins in a microbial organism, and semantic concepts that are expressed in a written piece of text, respectively. Importantly, it should be noted that repeated, or varying combinations of the basic units will provide hints on their relationships if sufficiently large datasets can be analyzed.

With topic models, we can decompose the contents of text documents into distinct topics that represent the underlying semantic concepts. Given a collection of documents, $d_1, \ldots, d_D$, a latent semantic analysis performed with a topic model like LDA results in a set of K discrete probability distributions, $P(v|t)$, over the terms $v \in V$ of the controlled vocabulary. The distributions $P(v|t)$ describe the associations of vocabulary terms with K different latent topic variables t. Each distribution describes a probabilistic cluster (a 'soft cluster') of words with similar occurrence patterns. By assumption, those will reflect the main topics that are discussed in the corpus of texts.

Due to the described analogy which we proposed, we hoped that topic models could be adapted to the analysis of (meta-)genomic input data. In theory, this should allow us to identify the principal functional modules that are encoded in the gene repertoires of a set of microbial organisms. The challenge for topic models like LDA would be to determine the correct associations between functional descriptors (words), and functional modules (topics). This is illustrated by a small example in Figure 5.2, in which LDA needs to find the correct assignment of the orange gene marked with '?' to one of four possible functional modules (named A,B,C,D). The Gibbs sampling inference procedure of LDA will essentially optimize over two different aspects: *1.) The global context*: Instances of the orange functional descriptor were most often assigned to module A. *2.) The local context*: Currently, module A has low probability in the third genome, thus an assignment of the FD annotated for the gene marked with '?' to module A is likely to be incorrect. Obviously, the correct assignment depends on a tradeoff between the two aspects which might be difficult to adjust. Moreover, the probabilities of all the modules in the genomes depend on the current assignments, and might change

from one iteration to the next. Therefore, finding optimal assignments for all modules and functional descriptors is a really difficult problem. At the end of the optimization process, the topic distributions $P(v|t)$ of LDA, which represent the latent functional modules, can be derived from the joint A/B/C/D-assignments of the individual genes across the collection (see Appendix B for more details on the LDA model).

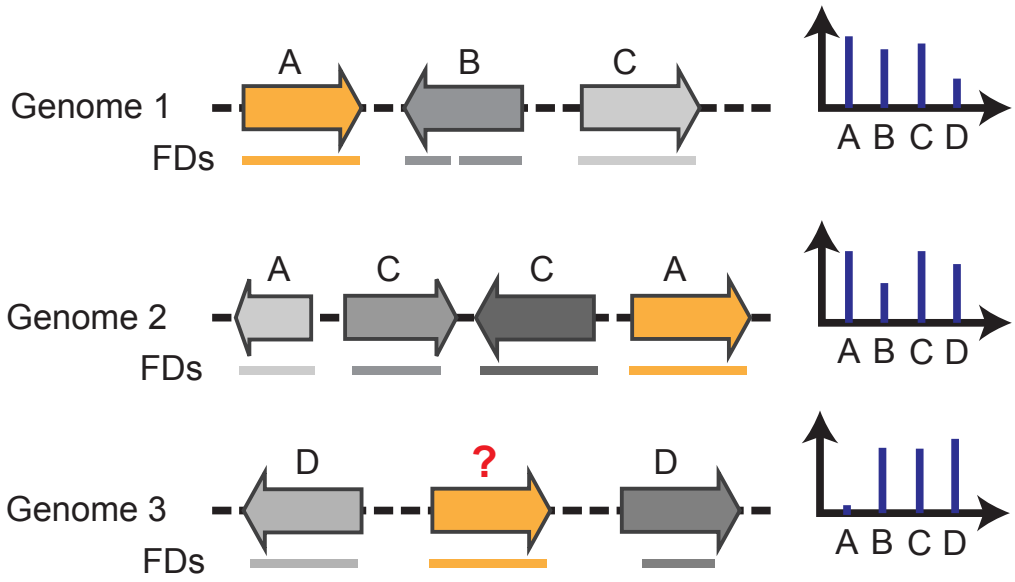


Figure 5.2: We propose the use of topic models to infer assignments of functional descriptors to functional modules. This toy example illustrates the tradeoff between local and global effects of gene-to-module assignments in the optimization procedure of latent Dirichlet allocation (see main text for details). *Orange* marks genes that are annotated with the same functional descriptor. Capital letters denote assignments of FDs to four latent topics (A,B,C,D), representing four functional modules. The histograms on the right illustrate exemplary topic weights for the individual genomes, which locally and globally depend on the current assignments across all genomes. We assume an intermediate state of an optimization process that tries to find optimal assignments for all genes. In the current iteration, the gene marked with '?' needs to be re-assigned to one of the four topics.

5.2 Publication - Konietzny *et al.* 2011

Status	**published**
Journal	BMC Bioinformatics (Impact factor in year of submission: 3.028)
Citation	Sebastian GA Konietzny, Laura Dietz, Alice C McHardy: Inferring functional modules of protein families with probabilistic topic models. *BMC Bioinformatics 2011*, **12**:141
URL	http://www.biomedcentral.com/1471-2105/12/141
Own contribution	75% Designed and performed the experiments Analyzed the data (with co-authors) Wrote the manuscript (with co-authors)

Konietzny et al. BMC Bioinformatics 2011, 12:141
http://www.biomedcentral.com/1471-2105/12/141

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Inferring functional modules of protein families with probabilistic topic models

Sebastian GA Konietzny[1], Laura Dietz[3] and Alice C McHardy[1,2*]

Abstract

Background: Genome and metagenome studies have identified thousands of protein families whose functions are poorly understood and for which techniques for functional characterization provide only partial information. For such proteins, the genome context can give further information about their functional context.

Results: We describe a Bayesian method, based on a probabilistic topic model, which directly identifies functional modules of protein families. The method explores the co-occurrence patterns of protein families across a collection of sequence samples to infer a probabilistic model of arbitrarily-sized functional modules.

Conclusions: We show that our method identifies protein modules - some of which correspond to well-known biological processes - that are tightly interconnected with known functional interactions and are different from the interactions identified by pairwise co-occurrence. The modules are not specific to any given organism and may combine different realizations of a protein complex or pathway within different taxa.

Background

Cells are complex dynamic systems capable of performing a variety of biochemical processes, many of which are of medical or industrial relevance, such as antibiotic biosynthesis or pathways for plant biomass degradation [1]. Despite the large number of sequenced genomes and metagenomes that are becoming available, our knowledge of the biological processes encoded therein is still limited and process-level genome annotation is far from complete [2-4]. Thus, the lack of high quality functional annotation or knowledge of the functional context for the majority of genes in any given genome/metagenome represents one of the biggest obstacles to obtaining quantitative insights into the relevant biological systems [5,6].

The functional units of signal transduction pathways, metabolic or gene regulatory networks are the products of individual genes, and the analysis of biological processes starts with their identification and characterization. A class of methods known as genome context methods are commonly used to infer the functional relationships between genes. One such method is pairwise co-occurrence (or phylogenetic) profiling [7]. This technique is based on the 'guilt by association' principle [8], which states that genes whose products are functionally coupled are likely to co-evolve and show similar evolutionary histories, resulting in conserved co-occurrence patterns across genomes [9]. The phylogenetic profile of a gene defines the organisms in which orthologs can be found, usually encoded as a binary or a real-valued vector with a length corresponding to the number of genomes considered [7,10,11]. A functional linkage between a pair of genes is predicted if their phylogenetic profiles show pairwise similarity. Commonly used similarity or distance measures are the Hamming distance, Pearson's correlation coefficient, the mutual information and the Jaccard coefficient (for a summary, see [11]). Furthermore, functional coupling is frequently seen in genes that are in spatial proximity to each other in the genome [12]. This can be due to their organization in operons, which allows the joint expression and regulation of functionally related genes. Therefore, conserved gene neighborhoods are a strong predictor for functional coupling [12]. Other genome context methods search for gene fusion events [13], similar expression patterns [14] or shared transcription factor binding sites [15]. In particular, gene fusion events, in which two genes with linked functions have been fused into one gene during

* Correspondence: mchardy@mpi-inf.mpg.de
[1]Max Planck Research Group for Computational Genomics and Epidemiology, Max Planck Institute for Informatics, University Campus E1 4, 66123 Saarbrücken, Germany
Full list of author information is available at the end of the article

evolution, provide substantial evidence for functional linkage. An obvious strategy to improve functional linkage prediction is to combine these methods [16]. This approach is realized in the STRING database [17].

Functional module detection

A functional module is defined as a set of proteins that jointly participate in a biological process [18,19]. As such, it is likely to be rich in proteins that are functionally coupled in a pairwise manner. If not all proteins, at least some subsets of a module's proteins are likely to be tightly coupled in their function. Accordingly, the proteins involved may map to densely connected subgraphs of protein-protein interaction networks.

A three step approach for detecting functional modules is common practice: First, genome context methods are used to identify pairwise interactions between proteins. Subsequently, the predicted interactions are combined into a functional linkage graph, in which the nodes represent the proteins, and the weighted edges represent the combined evidence for a functional relationship [16,20,21]. Finally, graph-based clustering techniques are used to identify communities of proteins that are likely to be functionally related [22,23]. Many definitions for communities in graphs exist [24]; however, the detection of functional modules essentially corresponds to identification of highly connected subgraphs [25]. Graph-based clustering and problems related to graph partitioning are often NP-hard, but can be tackled by approximate methods with good (though not optimal) results [24].

Watanabe *et al*. [26] used the Bond Energy Algorithm to find clusters of functionally coupled proteins, based on pairwise co-occurrence patterns, without constructing a graph. Their method uses pairwise distances between gene occurrence profiles, measured with the Hamming distance, to identify (disjoint) groups and is able to detect first-order transitive relationships between proteins. However, biological modules need not be disjoint, in general, and a potential limitation of this approach is the greedy nature of the algorithm, which makes the results sensitive to the order of the input data [26].

Besides the aforementioned unsupervised methods, supervised methods such as support vector machines have been applied to identify the proteins of metabolic and signal transduction pathways [27-29]. Note that the selection of genomes may be a critical factor for context analyses, because of phylogenetically conserved signals in the annotation data and a taxonomic bias in the determined genome sequences [30]. Jothi *et al*. studied the influence of genome selection on the results of co-occurrence profiling and suggested to use phylogenetically diverse, non-redundant sets of genomes [10].

Using genome context information in combination with state-of-the-art machine learning approaches is one of the most promising avenues to make progress in functional inference and has not been greatly explored at present [31]. Here, we demonstrate the utility of this approach for functional context inference. In particular, we use a Bayesian method known as Latent Dirichlet Allocation (LDA), which is based on a probabilistic topic model [32]. Topic models are used in text mining applications to reveal statistical relationships between words in collections of text documents, because it was observed that strong relationships usually correlate well with semantic agreement of words. The LDA model has previously been applied to identify protein relationships from MEDLINE abstracts of scientific articles [33,34] and to identify genes with similar behavior in multiple chemo-genomic experiments with *Saccharomyces cerevisiae* [35]. In contrast to this, our method processes large collections of genome annotations to detect functional modules of biological processes with heterogeneous sizes, which allows both small and large processes to be captured. Furthermore, a Bayesian model like LDA promises a robust performance with respect to common noise present in genome annotations, which greatly vary in quality, and may in part be incomplete or incorporate false functional assignments [36]. We applied our technique to a large collection of microbial genome annotations and compared the results with a state-of-the-art pairwise co-occurrence method. Our method identified a largely distinct set of predictions, many of which are supported by known functional interactions from STRING. The set of inferred modules partially maps to known KEGG pathways and the modules indicate a functional context for many protein families of currently unknown function. Our results thus represent a novel source of functional context assignments for protein families.

Results and discussion

A Bayesian method for functional module inference

Our method uses Latent Dirichlet Allocation (Methods) for inferring functional modules of biological processes as follows: A set of genome annotations serves as the document corpus, with individual genome annotations representing the documents. We define a fixed-sized vocabulary of words based on the gene annotations, such that words correspond to functional descriptors for gene products, as for instance orthologous groups (OGs) of genes [37], FIGfams [38], Pfam terms [39], EC numbers or other commonly used functional identifiers. Genes that are annotated with a certain functional descriptor represent single instances of the respective word. Note that our method treats genome annotations as a 'bag of functional descriptors',

meaning that the order of genes in the genome sequence is not considered.

The collection of functional descriptors, grouped into genome documents, serves as input to LDA, and Gibbs Sampling is used for model inference (Methods). So-called topics represent the latent variables of the LDA model and their values are inferred from the collection of genome annotations (Figure 1). Each inferred LDA topic defines a probability distribution ('topic distribution') over the chosen vocabulary. Functional descriptors with high probabilities show similar co-occurrence patterns within the collection of annotations. According to the 'guilt by association' principle, the inferred topics are likely to represent functional modules - sets of protein domains or orthologous groups of genes that are functionally linked to each other.

We used the k obtained topic distributions to define potential functional modules (PF-modules): Each PF-module is defined by a single topic and comprises the set of functional descriptors selected from the topic distribution by applying a threshold value C. For our experiments, we used C = 0.01. This choice of C was guided by visual inspection of the topic distributions, in accordance with [35]. In this study, identifiers for orthologous groups (OGs) of genes, i.e. COG and NOG terms from the eggNOG database [37], were used as input vocabulary for LDA. As such, the inferred potential functional modules correspond to groups of conserved gene families.

Functional module inference from prokaryotic genomes

We applied our method to 575 prokaryotic genome annotations from the STRING database, and tested it for different settings of k (k = 100 - 500), with three independent runs for each setting. We achieved the best results for k = 200 and k = 400 (see discussion below), and discuss results for k = 200 in more detail. To obtain a reliable estimate of model stability for k = 200, we performed six additional runs and averaged numerical results over the PF-module sets of all nine runs.

The largest PF-module consists of 61 (standard deviation (s.d.) 6.34) OGs; the average module size is 19.4 (s.d. 0.26). Based on the module sizes, we would expect approximately 3,880 OGs to be associated with the

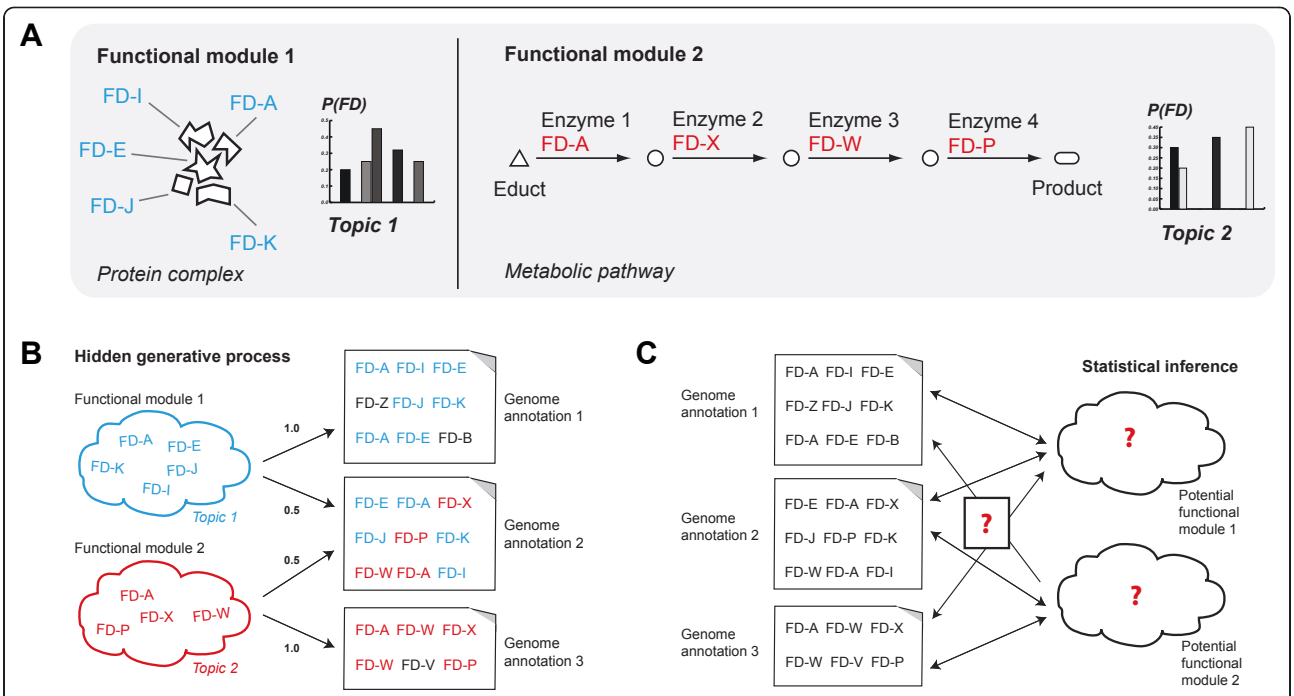


Figure 1 The LDA model assumes a hidden generative process that can be inversed for statistical inference. In our approach, topics are assumed to represent the unknown biological modules that have shaped the contents of genomes. As a simplifying example, the influence of two modules on the contents of three genome annotations is considered. ***Panel A:*** Functional descriptors (*FD* terms) are associated with proteins in the modules, and each module is represented by a probability distribution over *FD* terms. ***Panel B:*** The hidden generative process: Genome annotations are assumed to be generated from weighted mixtures of the probability distributions. The two clouds show the *FD* term set with the highest probabilities for each module. Note that the second genome annotation is equally shaped by both modules, whereas the other two annotations are solely shaped by one module. ***Panel C***: The input data as seen by our method. No *a priori* knowledge about the underlying modules is necessary. The potential functional modules are latent variables of the model that will be inferred from the collection. The identified modules are not necessarily specific to any given microbe, but potentially combine different realizations of a complex or pathway from different organisms.

modules. However, the inferred modules overlap in OG content, as only 1,554.3 (s.d. 38.2) distinct OGs on average are forming the set of modules of a run. In total, 868 OGs were consistently associated with the modules for all nine runs. These are likely to represent a core of conserved gene families with strong co-occurrence patterns in the data.

In accordance to the similar module size distributions over the nine runs, we also observed little variation over runs based on other evaluation criteria, discussed below. In the following, we therefore discuss the results from a randomly chosen, exemplary run with k set to 200. In this run, 198 non-empty PF-modules were identified (Additional file 1, Tables S1-198); for two topic distributions, no OGs exceeded the probability threshold C. Of these 198 modules, 70 particularly stable PF-modules could be tracked over all nine runs (Additional file 1, Tables S1-70). The average size of the 70 stable modules is 15 OGs, and of the 1,532 OGs associated with all 198 modules of the exemplary run, 43.9% are part of these stable modules. Note that the tracking of topic identities over runs follows a greedy heuristic strategy and may underestimate the true number of stable PF-modules (Methods).

We analyzed the functional consistency of the 198 modules in terms of their enrichment in COG functional categories [40]. On average, the most frequent functional category present in a module (with a minimum module size of seven OGs) is associated with 35% of the contained OGs (41.1% for the stable modules). This implies that modules are heterogeneous, but often include a significant portion of OGs from the same functional category. Interestingly, one of the most abundant categories is general function prediction only (histograms in Additional file 2), which contains OGs with insufficiently characterized functions. Thus, placement in a functional module might further indicate the functionalities for these gene families.

We could map 15 of the stable modules (Table 1) and 49 of all modules of the exemplary run (Additional file 3, Table S1) to KEGG pathways, based on six or more matched KO terms in the respective pathway. Overall, the modules contained many interactions annotated in the KEGG database. However, this only explained a part of the PF-modules, so we evaluated the functional coherence of the identified groups by means of additional quantitative measures.

Table 1 Profile of KEGG pathways with at least six matches to one of the stable modules

KEGG name	Modules
ABC transporters	5
Flagellar assembly	1
Porphyrin and chlorophyll metabolism	1
Oxidative phosphorylation	1
Bacterial secretion system	2
Two-component system	2
Phenylalanine, tyrosine and tryptophan biosynthesis	1
Phenylalanine metabolism	1
Starch and sucrose metabolism	1

The modules were identified as being stable across nine independent runs of our method with k = 200.

Functional coherence of modules

For the analyzed genomes, the STRING database provides evidence for pairwise functional couplings of OGs related to different types of functional modules, such as metabolic and signal transduction pathways, as well as protein complexes. The data include predictions from several genome context methods, as well as functional relationships from public protein-protein interaction and metabolic pathway databases. For our evaluation, we exclusively selected pairs of OGs that were sufficiently supported by evidence other than co-occurrence profiling (Methods). Overall, the reference dataset with this restriction comprised 60,880 distinct OG pairs. To evaluate the functional coherence of a module, we considered all possible pairs of OGs within the module, and determined the percentage of these found in the reference set. Module derived OG pairs that are part of the reference are referred to as *verified pairwise functional couplings* (verified-PWF-couplings). A high proportion of verified pairs then indicate the functional coherence of a module.

The average percentage of verified pairwise functional couplings for all 198 modules of the exemplary run is 14.9% (Figure 2), and the individual results for the modules are highly significant according to an estimate based on the hypergeometric distribution (Methods). Given the size of the input vocabulary (10,431 OG terms), the probability of an arbitrary OG pair matching a pair from the reference set by chance is P_{hit} = 0.0011. For an average-sized functional module, which consists of 19 OGs and 171 OG pairs, one would expect to observe less than one match by chance ($E[h]$ = 0.19). Thus, even small numbers of verified couplings are highly significant.

The proteins of a functional module may not directly interact, but be transitively linked to each other instead. Therefore, we also searched for indirect relationships. The OG pairs in the reference set correspond to edges in a functional network defined by high confidence interactions in STRING. We matched the OG pairs of a module against this network. If a pair exists in this reference network, it may either be an isolated edge or an edge connected to other matched edges. Ideally, all OGs of a module are functionally related and form a single

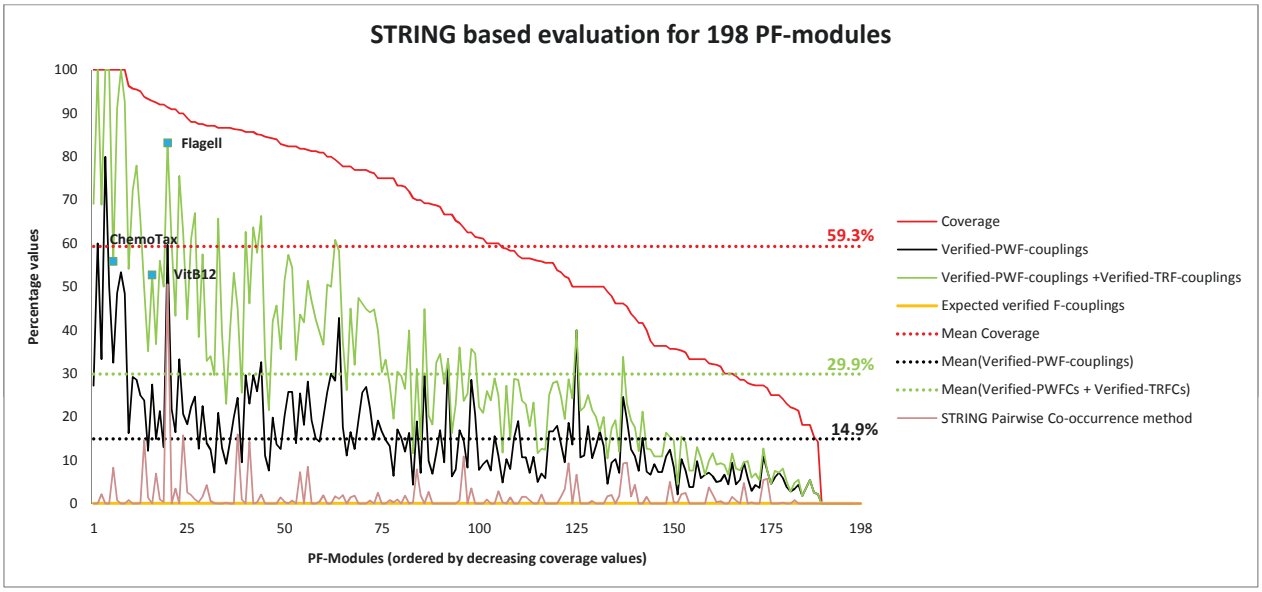


Figure 2 Overview of the STRING-based evaluation of 198 modules. We evaluated 198 PF-modules from a randomly chosen run with k = 200. Coverage values for the modules serve to assess their functional coherence. A coverage value of 100% means that the complete OG set of a module is interconnected within the reference functional network and forms a single cluster therein. If the coverage is 50%, then the same holds for half of the OGs of a module. The plot shows coverage values for all 198 modules of the exemplary run (modules 'ChemoTax', 'Flagell' and 'VitB12' are discussed in detail in the Results section). *'Verified-PWF-couplings'*: The percentage of *verified pairwise functional couplings* with respect to all tested OG pairs of a module. *'Verified-PWF-couplings + Verified-TRF-couplings'*: The percentage of a module's OG pairs that are either verified pairwise couplings or *verified first-order transitive functional couplings*. *'Expected verified F-couplings'*: The expected percentage of verified pairwise couplings to be found by chance for the OG set of a module. For an average-sized module, we expect to obtain less than one ($E[h]$ = 0.19) verified pairwise functional coupling by chance. The dashed lines indicate mean values, and the averaged mean coverage over all nine runs is 57.9% (1.3% s.d.). Finally, we determined the fraction of OG pairs within a module which are verified and have also been predicted by the pairwise co-occurrence method used by STRING.

connected component within the network. This would mean that they are either directly or indirectly functionally linked to each other. We denote the fraction of a module's OGs that are part of the largest connected component in the reference set as the module's *coverage* (Methods; Figure 3). Thus, if a module has coverage of 75%, this means that three-fourths of the module's OGs are either directly or indirectly linked to each other.

In total, we found 132 modules with an average size of 20.8 OGs and coverage of more than 50%. We also evaluated the first-order transitive relationships (*verified-TRF-couplings*), for which the total fraction of verified pairs rose from 14.9% to 29.9% (Figure 2).

We subsequently investigated how the validated pairs of the modules are dispersed within the reference network, i.e. whether they form isolated edges or highly-connected components. To this end, for each module, we determined the number of its connected components within the reference network. More than half of the modules map to only one such component; the average is less than two (1.67) components per module. Thus, the majority of modules represent large clusters within the reference network, with only a few isolated edges.

Modules with matches to KEGG pathways

We found three interesting modules (named 'ChemoTax', 'Flagell' and 'VitB12') among the stable modules in all nine LDA runs. These have a COG category functional enrichment of more than 50%. The first two modules are related to chemotaxis and the flagellar apparatus. 'ChemoTax' consists of 16 OGs and is very rich in signal transduction genes (OGs in the category T, 50%; Additional file 1, Table S1). Strikingly, this module achieves a coverage of 100%. Of the 67 verified OG pairs (39 verified pairwise couplings + 28 verified transitive couplings), only 10 were also detected by co-occurrence profiling. The larger 'Flagell' module consists of 35 OGs and is rich in cell motility genes (category N, 86%; Additional file 1, Table S2). This module has a coverage value of 91%. In the 'Flagell' module, we identified 355 verified pairwise and 140 verified transitive functional couplings, whereas pairwise co-occurrence profiling identified 301 of these 495 couplings. Interestingly, 'ChemoTax' and 'Flagell' share two gene families (*COG0835* and *COG2201*), which are both assigned to the functional categories for signal transduction (category T) and cell motility (category N). 'ChemoTax' and

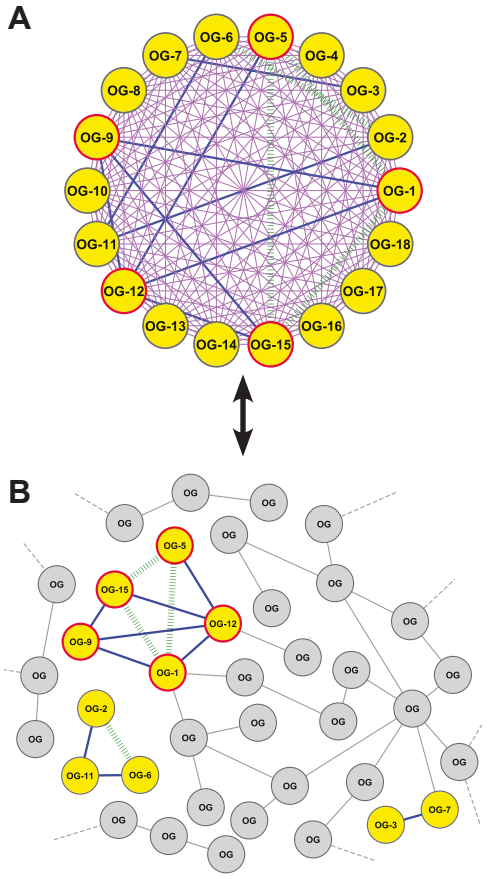


Figure 3 Evaluation scheme for the inferred potential functional modules. ***Panel A:*** A group of 18 OGs, representing an average-sized module. To assess the functional coherence of the module, all possible OG pairs of the module (153 pairs, illustrated as *purple* lines) are matched against a high confidence reference set of OG pairs from STRING. ***Panel B:*** The functional network spanned by the reference OG pairs. Pairwise interactions in the network that are matched by OG pairs of the module are marked as *blue* edges. In contrast, the majority of linkages in the reference network are marked in *gray*, indicating that they are not matched by any pair from the module. In each case where three module OGs are exclusively connected by two blue edges, we presume evidence for a transitive relationship (*green* edges), even though the third edge that completes the triangular relationship is not contained in the reference set. Thus, given the 153 tested pairs, the module yields nine verified pairwise interactions, plus four additional (first-order) transitive interactions. Note that in this case, the module covers three connected components in the network. The five OGs marked with *red* boundaries are part of the largest connected subcomponent, resulting in a coverage value of 5/18 = 27.8%.

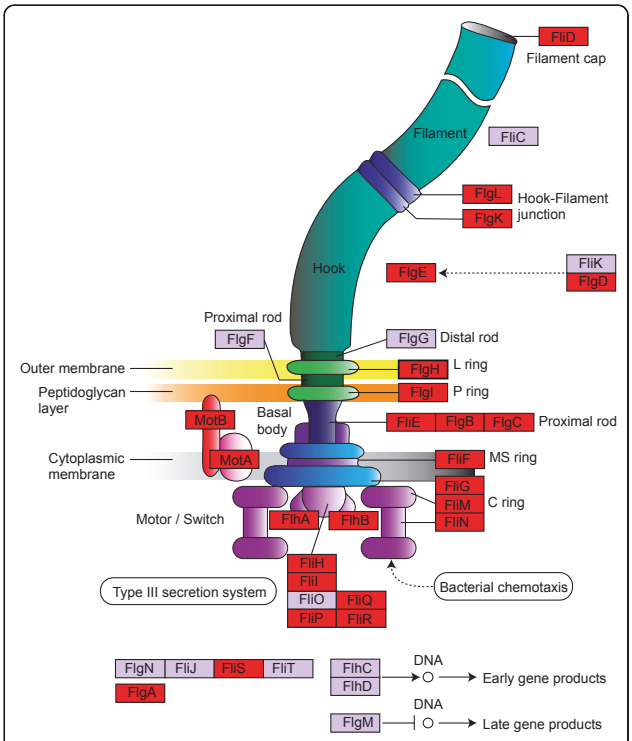


Figure 4 Mapping of the 'Flagell' module to the KEGG map 'flagellar assembly'. OGs of the 'Flagell' module are mapped to the KEGG map. Matched items are highlighted in *red*. This image is adapted from the original KEGG map.

'Flagell' thus may capture different aspects of a biological network related to chemotaxis and cell motility. While 'ChemoTax' does not correspond to a KEGG pathway and mainly consists of the OGs responsible for signal transduction, 'Flagell' comprises structural components of the flagellar apparatus and contains most of the elements of the respective KEGG map (Figure 4). Indeed, the relationship between the flagellar apparatus and chemotaxis is well known [41,42].

Furthermore, we identified a module comprising 32 gene families (Additional file 1, Table S3) that largely maps to the KEGG pathway 'Porphyrin and chlorophyll metabolism' (24 matches, KEGG map in Additional file 4). The module almost completely covers the process of synthesis from precorrin-2 to the vitamin B12 coenzyme, and we therefore refer to this module as 'VitB12'. The module has a coverage of 93.8%, while pairwise co-occurrence only predicted 29.8% of the 245 verified pairwise or transitive functional couplings. It should be noted that 'VitB12' clearly represents a meaningful module, although its 118 OG pairs with direct matches in the reference set cover only 23.8% of the tested pairs. Most supported pairs for this module originate from the 'database' and 'neighborhood' channels of STRING.

Another interesting stable module of 66 OGs that comprises ribosome-related gene families was found for k = 400 in all nine runs (Additional file 5, Table S1; the module was not found in all runs for k = 200; however, a corresponding module found in the exemplary run with k = 200 is presented in Additional file 1, Table S123). The module maps to the KEGG reference pathways 'Ribosome' (29 OGs, KEGG map in Additional file

6) and 'Aminoacyl-tRNA biosynthesis' (13 OGs). The functional enrichment for category translation and ribosomal structure (J) is 73%, and four gene families represent translation factors. Thus, our method is capable of identifying modules with a larger functional context than a single KEGG pathway. Interestingly, the module has a coverage value of 97% and none of the interactions were found by pairwise co-occurrence. Note that this module is not the only module related to the ribosome KEGG pathway. Several other modules map to pathways that are involved in protein biosynthesis. The number of verified pairs for all presented modules was highly significant according to our significance estimate ($P_{mult\_hit} \leq 0.001$).

Functional coherence of the modules in dependence of LDA parameter k

We evaluated the impact of k, the number of topics to be inferred, in a series of experiments with k set to 100, 200, 300, 400 and 500 (Table 2). Each experiment comprised three independent runs. The mean module size increased notably from k = 100 (13.82) to k = 200 (19.47), whereas it remained similar for k = 200, 300, 400 and 500. Interestingly, the functional coverage values only varied slightly for different settings of k, although the number of modules and their sizes steadily increased. This suggests that most of the identified modules for different settings of k were supported by evidence from STRING. The number of stable modules is larger for k = 400 and k = 500 than for runs with smaller values of k, and these modules also contain more distinct OGs. Thus, for larger values of k, a larger part of the STRING reference interaction network is identified. We additionally matched the stable modules from each experiment to the KEGG database to identify pathways with six or more hits to any one of the modules. For k = 200 and k = 400, this resulted in the most diverse profiles with matches to 20 different KEGG pathways. We decided to use a setting of k = 200 for a further detailed analysis, which included six additional runs, because this setting showed good results with respect to identified KEGG pathways and the largest support by known functional interactions; in terms of the fraction of identified stable modules with a coverage of at least 50% (Table 2). Further details of this comparison are discussed in a Supplementary note in Additional file 7.

Comparison with pairwise co-occurrence profiling

We compared our method with the state-of-the-art pairwise co-occurrence profiling method used in STRING. Pairwise co-occurrence identified a small fraction of the verified pairwise or first-order transitive couplings of the modules (Figure 2, Figure 5). The Venn diagram in Figure 5 shows the overlap with the reference set for both methods. Overall, both methods detected a small subset of the 60,880 reference pairs, resulting in recall rates of 8.4% for the modules and 2.4% for pairwise co-occurrence. These rates suggest that, in general, co-occurrence patterns contribute different evidence for functional linkage than other available sources of information about functional linkages. The PF-modules cover a largely distinct set of interactions, which includes 66.1% of the validated predictions of the pairwise method. The 4,174 validated functional couplings exclusive to the PF-modules exceed the overall number of linkages predicted by pairwise profiling. We mapped these 4,174 pairs to KEGG and found matches for 34.6% of them. The most abundant KEGG pathways were 'Ribosome', 'Two-component system' and 'Oxidative phosphorylation'.

For a complete evaluation, the precision, i.e. the fraction of correct assignments of all predicted linkages should be determined. However, this is complicated by the fact that non-existing interactions (according to the reference set) may reflect incomplete knowledge rather than the absence of interaction. Furthermore, solely for the sake of comparison with pairwise co-occurrence, all possible pairs were enumerated as functional linkage candidates within each module, which is

Table 2 Comparison of results for varying numbers of inferred topics

k	Mean # OGs associated with the modules		Mean module size		Mean coverage for k modules		Stable modules with ≥5 OGs and coverage ≥50%
	Average	s.d.	Average	s.d.	Average	s.d.	
100	635	28.6	13.82	0.27	64%	1%	33 (33%)
200	1560	24.9	19.47	0.11	58%	1%	66 (33%)
300	2009	31.1	21.76	0.9	56.7%	0.6%	68 (22.7%)
400	2223	33.9	22.9	0.42	53%	1%	102 (25.5%)
500	2378	7	22.34	0.19	49%	0%	97 (19.4%)

We performed 5 experiments to test different settings of k. The reported numbers for each experiment are averaged mean values of three runs. We used a greedy approach to track module identities across the runs for each setting of k, and refer to the modules identified in all three runs as the stable modules. The last column gives the number of stable modules with sufficient evidence for functional coherence based on the STRING analysis (in parentheses, we denote the fraction of stable modules satisfying the conditions, with respect to k).

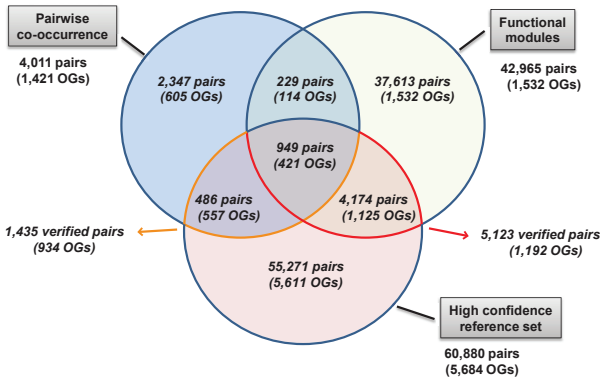


Figure 5 Pairwise predictions deduced from 198 modules compared with predictions of the co-occurrence baseline method. The Venn diagram visualizes the overlaps between the different OG pair sets. The partitions are defined over OG pairs. For each section of the Venn diagram, the number of distinct OG terms defined as parts of the pairs is noted. Note that the OG sets of the single sections are not necessarily disjoint. A large total of 5,123 pairs have been validated for the modules. Additionally, 7,603 first-order transitive relationships could be validated based on the reference set (not included in the Venn diagram). Based on our estimate $E[h]$ for the expected number h of random matches for a single module, we would expect less than 50 matches to the reference set for the 42,695 tested pairs of the modules by chance.

not an assumption warranted by our method and results in a large number of pairwise interactions being tested. With these restrictions being applied, precision values for both methods are 35.8% (pairwise co-occurrence) and 11.9% (functional module inference). This value may serve as an estimate of the lower bound of the actual precision for functional module inference. As pairs of gene families in a module may also interact indirectly with each other, they may not directly match a pair of the reference set. When taking these indirect interactions in the reference set into consideration (verified-TRF-couplings), we found evidence for 7,603 OG pairs that represented further functional relationships implicitly verified by STRING. Combined with the 5,123 directly verified pairs, this corresponds to 29.6% of the tested 42,965 pairs and may serve as a second estimator of precision for the presented functional module inference.

The capacity to identify indirect relationships highlights an important advantage of directly inferring groups of functionally related OGs. Of the identified transitively linked pairs, 828 could be mapped to KEGG pathways. Interestingly, the fraction of verified-TRF-couplings also serves as an indicator for meaningful relationships revealed by the modules that are not directly identified by any of the pairwise operating genome context methods used in STRING. As such OG pairs are not explicitly part of the reference set, none of the prediction methods in STRING provided sufficient evidence for their direct coupling. However, as demonstrated, a transitive relationship exists in the reference.

The modules cover parts of the reference network that tend to be tightly interconnected. Figure 6 visualizes the densely connected core of the functional network defined by the reference set, showing the embedded prediction sets of both methods (a picture of the complete network is provided in Additional file 8). The modules that we have discussed are embedded in this network, showing that our method is capable of identifying meaningful groups of OGs in a dense interaction network.

Conclusions

We proposed and evaluated a new probabilistic method for directly identifying functional modules of gene or protein families using co-occurrence patterns in a collection of annotated sequence samples. In our analysis, we used orthologous groups of genes (OGs), which are considered to be a reliable estimator for isofunctional groups of genes [38,40]. However, one could also use terms such as FIGfams, which incorporate careful manual curation by experts [38], KO terms, Pfam domains, TIGRfam terms or EC numbers [39,43].

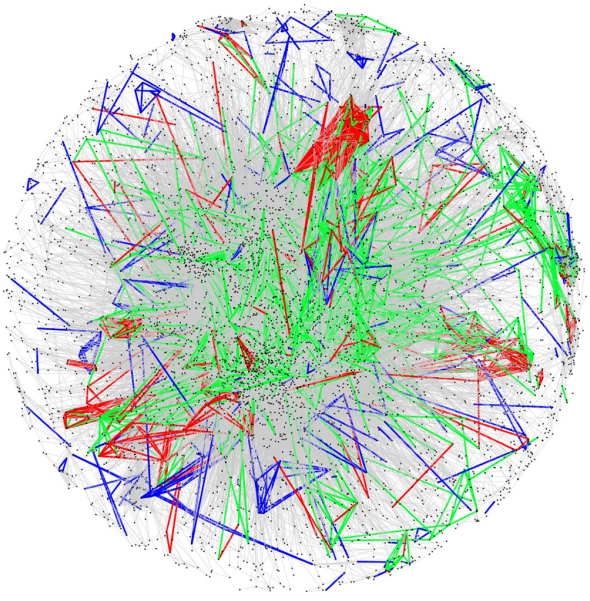


Figure 6 Visualization of the functional network spanned by the OG pairs of the reference set. The figure shows the pairwise functional interactions defined by the reference set as edges between OGs in a network graph. The subset of verified pairwise predictions from the modules is shown in *green*, whereas the subset of verified predictions by pairwise co-occurrence profiling is shown in *blue*. Functional interactions that are predicted by both methods are colored in *red*, and those not detected by any of the methods are shown in *gray*.

The presented method is capable of simultaneously processing a large number of genome or metagenome annotations. We tested our methodology on a comprehensive set of microbial genomes; but certainly a targeted selection of a suitable collection of genome annotations for input will be a major key to the detection of further interesting PF-modules in the future.

Our method returns a soft clustering of functional annotation terms, in which a term can be assigned to multiple modules. This is well suited for the problem of assigning gene families to biological processes, given the multiplicity of roles and functionalities associated with some gene families, which may depend, for instance, on the genomic context [3]. Furthermore, processes may appear in multiple, slightly different variants across genomes. Such process variations arise, for instance, through alternative branches in metabolic pathways, or through reactions that can be realized by structurally different proteins [44,45]. The topic model that we use accounts for this phenomenon adequately. LDA topics are globally defined for the analyzed collection of genomes and therefore generalize over slightly different variants of a process, instead of splitting them into multiple modules. Thus, a functional module may combine multiple isofunctional but non-orthologous gene families, which fulfill similar roles in different organisms.

LDA is an unsupervised method, which requires no *a priori* knowledge about the structures it identifies. This gives us the opportunity to detect previously unknown functional modules. However, this also means that the nature of the biological entities captured by these modules is uncertain. For instance, metabolic and signal transduction pathways, protein complexes or mixtures of these might be the underlying biological signals.

We evaluated the biological significance of the identified modules using functional interactions from STRING, which integrate different sources of biological information about functional modules. As a result, we found that PF-modules cover diverse biological signals, such as protein complexes ('Ribosome' and 'Flagell'), signal transduction components ('ChemoTax') and metabolic pathways ('VitB12'). A convenient property of the method is that the input data are represented as 'bags of gene families' and thus no knowledge of the neighboring genes is required. Therefore, it can also be applied to highly fragmented metagenomes with many short fragments, for which there is currently a shortage of analysis techniques [5,46,47].

In our study, we observed that the identified potential functional modules are significantly enriched with high confidence functional interactions and capture well-known biological processes, such as chemotaxis. Moreover, the majority of the modules' high confidence interactions were not detected by a state-of-the-art pairwise co-occurrence method.

Notably, a great number of newly implied OG interactions, derived from the PF-modules, could not be verified as pairwise interactions, but received reasonable support as first-order transitive relationships from STRING instead. And many of these interactions could also be mapped to KEGG pathways. In summary, this shows that an approach of direct inference of functional modules reveals further information about biological processes. We believe that the direct inference of groups of genes is well suited for the discovery of functional context, as biological processes incorporate many indirect functional couplings between the encoded proteins. For instance, proteins may serve as network hubs that link two or more processes. In this case, proteins from the processes involved will be directly coupled to the hub and only indirectly to the proteins of the other processes. Finally, in many cases, the proteins of a particular process vary between different organisms [44,45]. These scenarios result in proteins that are only transitively linked to each other via other proteins of a process.

The inference procedure of the LDA model is based on a Markov Chain Monte Carlo method. With a limited number of iterations, such methods are non-deterministic, i.e. a series of LDA runs will not produce exactly the same sets of PF-modules. At the chosen stop point of iterations, we observed that the size distribution as well as the degree of functional coherence over the module sets varied only slightly between runs, indicating that there is sufficient convergence of the Markov chain. Our heuristic search strategy allowed us to identify 70 stable modules across the nine runs we performed. However, many of the other PF-modules also occur in more than one run. We investigated the relationship between stability and functional coherence. The average coverage over all 198 PF-modules was 59.3%, while the 128 less stable PF-modules showed an average coverage of 61.1%. Thus, also the less stable modules contribute significantly to the overall estimate of functional coherence.

An interesting future direction for research will be the use of the genome-specific topic-weights to investigate how the gene family content of a functional module varies in individual genomes (Supplementary note in Additional file 7, heatmap in Additional file 9). We found that the OGs of the modules are associated with almost all the COG functional categories and that OGs with no specific functional assignment (category R) are frequent. Therefore, we could use modules to refine gene annotations in the following way: Assuming that a given module shows evidence for being related to a specific biological process and that the module is assigned a high weight for a genome of interest. Then, poorly characterized genes of this genome, whose gene families are

associated with the module, could be tentatively associated with this biological process. Methods for assessing the function of poorly characterized genes from putative interaction partners are currently being investigated [31]. Another interesting research direction will be to analyze the evolution of functional groups, based on the presence or absence of (parts of) a module across taxa.

Prediction methods for functional relationships that rely on conserved genomic context are prone to false positive predictions if pseudo-genes are involved [48,49]. Pseudo-genes are functionally disabled copies of genes, which typically occur in at least 1-5% of gene-like sequences in prokaryotic genomes [50]. We found several PF-modules that integrate different transposable elements. In these cases, pseudo-genes might have had an impact on the inferred modules. LDA, like other co-occurrence techniques, could principally be misguided in cases where multiple copies of an OG reside in the same genome by chance without being retained by selection for a certain functionality, e.g. due to repeats of a genomic sequence.

We compared results for different choices of the number of topics to be inferred, and suggest choosing a setting for k between 200 and 400 for the analyzed set of genomes. Finding an optimal choice for the number of topics corresponds to the problem of model selection in latent class cluster analysis [51]. Blei *et al.* [52] proposed tackling this problem by modeling a hierarchy of topics and embedding LDA into a Hierarchical Dirichlet Process [52,53]. However, the additional level of complexity in these models is likely to cause the inference process to be more time-consuming, and convergence of the process has to be carefully monitored.

In summary, we found that our method allows identification of well-known biological processes, as well as the discovery of new modules supported by high confidence functional interactions. It furthermore places many gene families of currently poorly characterized function within a functional context. The presented technique could thus help to enhance our knowledge of the biological processes governing microbial life and reveal new functional connections for many microbial genes.

Methods

Technical aspects of the LDA model

LDA assumes the existence of a fixed number k of underlying 'topics' that define the essential semantics of the whole text corpus. Each topic t_i (with $i \in \{1,...,k\}$) defines a probability distribution ('topic distribution') over the vocabulary V of words: $P(w|t_i)$ for $w \in V$. Words with high probabilities under a topic distribution are statistically linked to each other, i.e. they have similar co-occurrence patterns with respect to the document collection. This way, each topic defines a specific grouping of words that are thought to be semantically related. However, the topics represent latent variables of the LDA model that need to be inferred from the input data.

LDA uses k multinomial distributions with Dirichlet priors to model the topics. These distributions are globally defined for all documents of the collection, but the model assigns weights to the individual topics for each document. For each document d_j (with $j \in \{1,... D\}$), a set of probabilities $P(t_i|d_j)$ exists for $i \in \{1,... k\}$, which represent the weights. The underlying assumption is that the documents are the result of a hidden generative process, in which the observed word frequencies have been generated from a document-specific, weighted mixture of the topic distributions. This relationship between observed word frequencies and the topic distributions is reflected in $P(w|d_j)$, which is the probability of observing a certain word $w \in V$ as part of document d_j:

$$P(w|d_j) = \sum_{i=1}^{k} P(w|t_i) \cdot P(t_i|d_j)$$

The word content of a single document d_j mainly corresponds to a set of words contributed by the subset of topics with the highest probabilities $P(t_i|d)$. Note that words may have high probabilities in multiple topic distributions. Such ambiguous words are thus related to more than one topic. However, depending on the document in which an instance of a word appears, one may assess the word's correct topic affiliation based on the document-specific weighting of topics. To this end, LDA also offers a probabilistic framework that enables the user to estimate the probabilities for assigning a word instance to certain topics, depending on the document in which it occurs.

For inference of the latent model parameters, Markov Chain Monte Carlo (MCMC) techniques [54], such as Gibbs sampling, can be applied [55]. At the end of the inference process, one obtains the topic distributions and document-specific probability weights for the topics.

Monitoring stability of the inferred model

MCMC sampling techniques efficiently estimate the posterior distribution over model parameters [54]. However, the actual time needed for convergence cannot be estimated precisely, and the efficiency of the sampling depends on the complexity of the model and the analyzed data. To assess the convergence of the inference process, a commonly used approach is to compare the results from a number of runs. We used the symmetrized version of the Kullback-Leibler divergence (KL divergence):

$$KL(p,q) = \frac{1}{2}\left[\left(\sum_{i=1}^{V} p_i \log_2 \frac{p_i}{q_i}\right) + \left(\sum_{i=1}^{V} q_i \log_2 \frac{q_i}{p_i}\right)\right]$$

defined for two distributions, p and q, to assess the stability of the inferred topic distributions across the final results from different runs [56]. Using KL divergence, we performed an all-against-all similarity comparison between the topics of two different runs. Then, topics that showed the smallest distances from each other were mapped between runs with a greedy best first search. Note that this easy to implement algorithm identifies a local optimum which does not necessarily represent a globally optimal solution in terms of the minimal KL divergences between the topics in all formed pairs. A best first search strategy may disregard suboptimal choices of pairs that would allow an improved overall mapping, due to improved mappings of other topics. If a mapping was circularly closed over N different runs (i.e. topic i of run 1 mapped to topic j of run 2, topic j mapped to topic k of run 3, and so on, until topic l of run N mapped back to topic i of run 1), we say the topic behaved consistently across these runs. We refer to topics that we could track consistently across all performed runs as stable topics.

Input data and preprocessing

Genome annotations for 575 prokaryotic genome sequences were downloaded from the STRING database (version 8.2) [17]. The number of distinct OGs (COG/NOG terms) in this dataset is 47,993. We removed all terms appearing in less than 10 genomes, which reduced the vocabulary to 10,431 distinct OGs. This filtering step facilitated evaluation with respect to the computational requirements. In subsequent experiments, we found that removing this restriction does not significantly change the results (data not shown).

Parameter settings for LDA runs and availability of the LDA implementation

Our method uses the LDA implementation available at http://gibbslda.sourceforge.net/, which relies on Gibbs sampling [55]. The LDA model defines two hyperparameters, α and β, which specify the underlying Dirichlet prior distributions. LDA was run with 2,500 iterations and the following parameter settings: k = 200 (the number of topics), α = 0.5, β = 0.01. The experiments were repeated nine times to ensure stability of the results.

Evaluation

Construction of a STRING-based reference set of high confidence interactions

The STRING database provides information about pairwise functional couplings of OGs for the analyzed set of genomes. Supporting evidence for each OG pair can stem from seven sources (channels): 'neighborhood', 'fusion', 'co-occurence', 'coexpression', 'experimental', 'database' and 'textmining'. For each evidence channel, a score quantifies its reliability. These reliability estimates were derived by benchmarking the predictive performance of the individual channels against a reference set of protein associations from the KEGG database [16]. STRING also provides a 'combined score' as an integrated measure of support from all evidence channels [17].

The STRING data (version 8.2) define 6,007,943 distinct pairs of OGs as being functionally coupled, based on a combined evidence score of 0.15 or more. Combined scores in the range of [0.4-0.7] represent a medium level of confidence, whereas the range of [0.7-1.0] denotes a high level of confidence [17]. We followed the procedure described in [17] to combine information from the different channels without co-occurrence information into a combined score, using a script provided by the database maintainers http://bitbucket.org/mkuhn/stringtools/src/tip/prior_correction/discard_channels_cogs.py. Thus, the modified combined scores represent evidence for pairwise functional coupling independent of co-occurrence. This procedure resulted in 2,472,604 remaining pairs, because discarding information from one channel decreases the overall combined scores, and, as a result, some fell below the threshold of 0.15. We used the modified combined scores to define a high confidence reference set of known pairwise functional couplings. Our reference comprises all OG pairs from STRING for which *(i)* both OGs are present in the input vocabulary V and *(ii)* the modified combined score is at least 0.7. The resulting reference set consisted of 60,880 unique OG pairs (Additional file 10).

Assessing the functional coherence of the potential functional modules

For each PF-module, we determined the set S_{PFM} of all possible unique OG pairs. For a PF-module of size l, the number of pairs is $m=|S_{PFM}| = (l\cdot(l\text{-}1))/2$. We then tried to identify as many of these pairs as possible within the reference set of pairs and refer to matches as *verified pairwise functional couplings* (verified-PWF-couplings). Furthermore, for each OG pair of a PF-module without a direct match to the reference set, we searched for a third OG from the respective module with a verified-PWF-coupling to both OGs of the original pair. OG pairs that were validated by this approach are referred to as *verified first-order transitive functional couplings* (verified-TRF-coupling). Additionally, we computed the transitive closure for each module with respect to the functional network spanned by the OG pairs of the reference set. Therefore, we determined the set of verified pairwise functional couplings for the OG set of the respective module, and used this set of pairs as edges between OG nodes to construct an undirected graph. Finally, we determined the connected sub-components of the graph, and used the percentage of the module's OGs that were part of the largest connected component

as an estimate for the module's functional coherence. We refer to this value as the module's *coverage*.

Significance estimation

The data from STRING allow us to verify pairwise OG interactions. We assessed the statistical significance of finding h OG pairs with matches to the reference set of interactions among the pairs of a PF-module. We used the hypergeometric distribution to estimate the probability of observing this result by chance:

Let D be the set of all possible unique pairs of OGs for the input vocabulary V. Given a vocabulary of size $|V|$, the number of pairs in D is $|D| = (|V|\cdot(|V|\text{-}1))/2$. Further, let U be the reference set of OG pairs. $|U|$ denotes the size of the set. Due to the construction rules of the reference set, $U \subseteq D$ holds and therefore a random pair in D will also be part of U with probability $P_{hit} = \frac{|U|}{|D|}$. If we count the observation of a match to the reference set as a success, matching a random OG pair against the reference can be regarded as a Bernoulli experiment with the success probability P_{hit}.

Now, since S_{PFM} represents the set of unique OG pairs of a PF-module, let $m = |S_{PFM}|$ be the size of this set. We used the hypergeometric distribution to assess the probability of achieving h verified-PWF-couplings for S_{PFM} by chance. The subset U of D, comprising the OG pairs defined in the reference, represents the set of possible successes if we randomly draw from the population D. Thus, the probability $P_{hyper}(X = h)$ of observing h successes by chance, given that we draw m times from the population D without replacement is:

$$P_{hyper}(X = h) = \frac{\binom{|U|}{h} \cdot \binom{|D| - |U|}{m - h}}{\binom{|D|}{m}}$$

This gives the expected number of random matches for a set S_{PFM} as $E[h] = m \cdot \frac{|U|}{|D|} = m \cdot P_{hit}$.

Finally, the cumulative distribution function $P(X \leq x)$ of the hypergeometric distribution was used to estimate the probability of observing h or more hits among m randomly formed query pairs. This is $P_{mult\_hit}(h) = P(X \geq h) = 1\text{-}P(X \leq h\text{-}1)$.

Comparison with a state-of-the-art co-occurrence method

The pairwise co-occurrence profiling method of STRING evaluates the mutual information between OG profiles and also accounts for biases in the data caused by phylogenetic relationships between genomes [16]. We determined the predictions of the pairwise profiling method for our comparison as follows: Starting with all OG pairs provided by STRING, we first removed all pairs containing OGs which were not part of the vocabulary V in our analysis. Then, we determined all OG pairs with a score of at least 0.4 from the 'co-occurrence' channel. This score threshold corresponds to the lower bound of the medium confidence interval for STRING scores.

Mapping OG identifiers to KEGG pathway maps

KEGG pathway maps are defined for orthologous groups of genes. However, these maps are based on KEGG-specific identifiers for such groups, known as KO (KEGG Orthology) terms [57]. To map OG identifiers from the eggNOG database to KO identifiers of the KEGG maps, a mapping table from KEGG was used. Note that these mappings are not defined for NOG terms.

Additional material

Additional file 1: A list of 198 potential functional modules. The Supplementary Tables **S1-198** show 198 potential functional modules that were identified in a randomly chosen, exemplary run of the presented method (k = 200). Tables **S1-70** represent the subset of particularly stable modules that could be tracked consistently across nine independent runs of the method.

Additional file 2: Comparison of histograms over COG functional categories. Comparison of two histograms over COG functional categories for (***A***) 70 stable modules and (***B***) modules that could not be tracked across all nine runs.

Additional file 3: Profile of KEGG pathways with at least six matches to one of the 198 modules. Supplementary Table **S1**: List of KEGG pathways with at least six matches of their KO terms to one of the 198 potential functional modules inferred in the exemplary run with k = 200.

Additional file 4: Visualized matches to the KEGG pathway 'Porphyrin and chlorophyll metabolism'. KO terms that are matched by the OGs of the respective potential functional module are highlighted in *pink*.

Additional file 5: 'Ribosome'-related functional module. Supplementary Table **S1**: OGs of the 'Ribosome'-related functional module that was identified in nine runs with k = 400.

Additional file 6: Visualized matches to the KEGG pathway 'Ribosome'. KO terms that are matched by the OGs of the respective potential functional module are highlighted in *pink*.

Additional file 7: Supplementary note. This document includes additional details of the comparison of results for different settings of k, and a discussion on the distribution of the probability weights of the modules across the analyzed genomes.

Additional file 8: Visualization of the functional network spanned by the OG pairs of the reference set. Pairwise functional interactions are defined by the reference set as edges between OGs in a network graph. The subset of verified pairwise predictions from the modules is shown in *green*, whereas the subset of verified predictions by pairwise co-occurrence profiling is shown in *blue*. Functional interactions that are predicted by both methods are colored *red*, and those not detected by any of the methods are shown in *gray*.

Additional file 9: Visualization of the distribution of probability weights of the modules across the analyzed genomes. The unclustered heatmap indicates the strengths of probability weights of the 198 modules across the genomes. Rows represent the genomes, whereas columns represent the weights of the modules. The brighter the color of a cell, the larger is the probability weight for the respective PF-module. We re-scaled the values of each row, using minimum and maximum values, to fit values to the interval [0,1]. A discussion of this heatmap is part of the Supplementary note in Additional file 7.

Additional file 10: Reference set of high confidence pairwise OG interactions. List of high confidence interactions with evidence support values from the individual STRING channels, and modified combined scores.

List of abbreviations

KL divergence: Kullback-Leibler divergence; KO: KEGG Orthology; LDA: Latent Dirichlet Allocation; MCMC: Markov Chain Monte Carlo; OG: Orthologous group (of genes); PF-module: Potential functional module; Verified-PWF-coupling (VPWFC): Verified pairwise functional coupling; Verified-TRF-coupling (VTRFC): Verified (first-order) transitive functional coupling;

Acknowledgements

We would like to thank M. Foster, C. Tusche, L. Steinbrück, K. Patil, I. Sommer, D. Emig, B. Kneissl, D. Stöckl and L. Feuerbach for discussion and comments.

Author details

[1]Max Planck Research Group for Computational Genomics and Epidemiology, Max Planck Institute for Informatics, University Campus E1 4, 66123 Saarbrücken, Germany. [2]Department for Algorithmic Bioinformatics, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany. [3]Department of Databases and Information Systems, Max Planck Institute for Informatics, University Campus E1 4, 66123 Saarbrücken, Germany.

Authors' contributions

SGAK designed the evaluation setup and performed the experiments; LD provided technical advice on Latent Dirichlet Allocation; SGAK and ACM wrote the manuscript; ACM devised the project and gave conceptual advice. All authors read and approved the final manuscript.

Received: 24 September 2010 Accepted: 9 May 2011
Published: 9 May 2011

References

1. Rubin EM: **Genomics of cellulosic biofuels.** *Nature* 2008, **454**:841-845.
2. Osterman A, Overbeek R: **Missing genes in metabolic pathways: a comparative genomics approach.** *Curr Opin Chem Biol* 2003, **7**:238-251.
3. Reed JL, Famili I, Thiele I, Palsson BO: **Towards multidimensional genome annotation.** *Nat Rev Genet* 2006, **7**:130-141.
4. Stein L: **Genome annotation: from sequence to biology.** *Nat Rev Genet* 2001, **2**:493-503.
5. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L, Cieplak P, Miller CS, Li H, Mashiyama ST, Joachimiak MP, van Belle C, Chandonia JM, Soergel DA, Zhai Y, Natarajan K, Lee S, Raphael BJ, Bafna V, Friedman R, Brenner SE, Godzik A, Eisenberg D, Dixon JE, Taylor SS, *et al*: **The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families.** *PLoS Biol* 2007, **5**:e16.
6. **CAFA Challenge: Critical Assessment of Function Annotations.** 2011 [http://biofunctionprediction.org/].
7. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci USA* 1999, **96**:4285-4288.
8. Aravind L: **Guilt by association: contextual information in genome analysis.** *Genome Res* 2000, **10**:1074-1077.
9. Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE: **Co-evolution of proteins with their interaction partners.** *J Mol Biol* 2000, **299**:283-293.
10. Jothi R, Przytycka TM, Aravind L: **Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment.** *BMC Bioinformatics* 2007, **8**:173.
11. Kensche PR, van Noort V, Dutilh BE, Huynen MA: **Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution.** *J R Soc Interface* 2008, **5**:151-170.
12. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci USA* 1999, **96**:2896-2901.
13. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: **Protein interaction maps for complete genomes based on gene fusion events.** *Nature* 1999, **402**:86-90.
14. van Noort V, Snel B, Huynen MA: **Predicting gene function by conserved co-expression.** *Trends Genet* 2003, **19**:238-242.
15. McGuire AM, Church GM: **Predicting regulons and their cis-regulatory motifs by comparative genomics.** *Nucleic Acids Res* 2000, **28**:4523-4530.
16. von Mering C, Zdobnov EM, Tsoka S, Ciccarelli FD, Pereira-Leal JB, Ouzounis CA, Bork P: **Genome evolution reveals biochemical networks and functional modules.** *Proc Natl Acad Sci USA* 2003, **100**:15428-15433.
17. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P: **STRING: known and predicted protein-protein associations, integrated and transferred across organisms.** *Nucleic Acids Res* 2005, **33**:D433-D437.
18. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology.** *Nature* 1999, **402**:C47-C52.
19. Pereira-Leal JB, Enright AJ, Ouzounis CA: **Detection of functional modules from protein interaction networks.** *Proteins* 2004, **54**:49-57.
20. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: **A combined algorithm for genome-wide prediction of protein function.** *Nature* 1999, **402**:83-86.
21. Rentzsch R, Orengo CA: **Protein function prediction - the power of multiplicity.** *Trends Biotechnol* 2009, **27**:210-219.
22. Navlakha S, Schatz MC, Kingsford C: **Revealing biological modules via graph summarization.** *J Comput Biol* 2009, **16**:253-264.
23. Zhang KX, Ouellette BFF: **Pandora, a pathway and network discovery approach based on common biological evidence.** *Bioinformatics* 2010, **26**:529-535.
24. Fortunato S: **Community detection in graphs.** *Phys Rep* 2010, **486**:75-174.
25. Brohee S, van Helden J: **Evaluation of clustering algorithms for protein-protein interaction networks.** *BMC Bioinformatics* 2006, **7**:488.
26. Watanabe RLA, Morett E, Vallejo EE: **Inferring modules of functionally interacting proteins using the Bond Energy Algorithm.** *BMC Bioinformatics* 2008, **9**:285.
27. Bostan B, Greiner R, Szafron D, Lu P: **Predicting homologous signaling pathways using machine learning.** *Bioinformatics* 2009, **25**:2913-2920.
28. Dale JM, Popescu L, Karp PD: **Machine learning methods for metabolic pathway prediction.** *BMC Bioinformatics* 2010, **11**:15.
29. Fröhlich H, Fellmann M, Sültmann H, Poustka A, Beissbarth T: **Predicting pathway membership via domain signatures.** *Bioinformatics* 2008, **24**:2137-2142.
30. Sun J, Xu J, Liu Z, Liu Q, Zhao A, Shi T, Li Y: **Refined phylogenetic profiles method for predicting protein-protein interactions.** *Bioinformatics* 2005, **21**:3409-3415.
31. Janga SC, Diaz-Mejia JJ, Moreno-Hagelsieb G: **Network-based function prediction and interactomics: The case for metabolic enzymes.** *Metab Eng* 2010, **13**:1-10.
32. Blei DM, Ng AY, Jordan MI: **Latent Dirichlet Allocation.** *J Mach Learn Res* 2003, **3**:993-1022.
33. Aso T, Eguchi K: **Predicting protein-protein relationships from literature using latent topics.** *Genome Inform* 2009, **23**:3-12.
34. Zheng B, McLean DC, Lu X: **Identifying biological concepts from a protein-related corpus with a probabilistic topic model.** *BMC Bioinformatics* 2006, **7**:58.
35. Flaherty P, Giaever G, Kumm J, Jordan MI, Arkin AP: **A latent variable model for chemogenomic profiling.** *Bioinformatics* 2005, **21**:3286-3293.
36. Friedberg I: **Automated protein function prediction - the genomic challenge.** *Brief Bioinform* 2006, **7**:225-242.
37. Muller J, Szklarczyk D, Julien P, Letunic I, Roth A, Kuhn M, Powell S, von Mering C, Doerks T, Jensen LJ, Bork P: **eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations.** *Nucleic Acids Res* 2010, **38**:D190-D195.
38. Meyer F, Overbeek R, Rodriguez A: **FIGfams: yet another set of protein families.** *Nucleic Acids Res* 2009, **37**:6643-6654.
39. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer ELL: **The Pfam protein families database.** *Nucleic Acids Res* 2002, **30**:276-280.
40. Tatusov RL, Galperin MY, Natale DA, Koonin EV: **The COG database: a tool for genome-scale analysis of protein functions and evolution.** *Nucleic Acids Res* 2000, **28**:33-36.

41. Pereira M, Parente JA, Bataus LAM, das Dores de Paula Cardoso D, Soares RBA, de Almeida Soares CM: **Chemotaxis and flagellar genes of *Chromobacterium violaceum*.** *Genet Mol Res* 2004, **3**:92-101.
42. Rajagopala SV, Titz B, Goll J, Parrish JR, Wohlbold K, McKevitt MT, Palzkill T, Mori H, Finley RL, Uetz P: **The protein network of bacterial motility.** *Mol Syst Biol* 2007, **3**:128.
43. Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, White O: **TIGRFAMs: a protein family resource for the functional identification of proteins.** *Nucleic Acids Res* 2001, **29**:41-43.
44. Koonin EV, Mushegian AR, Bork P: **Non-orthologous gene displacement.** *Trends Genet* 1996, **12**:334-336.
45. Morett E, Korbel JO, Rajan E, Saab-Rincon G, Olvera L, Olvera M, Schmidt S, Snel B, Bork P: **Systematic discovery of analogous enzymes in thiamin biosynthesis.** *Nat Biotechnol* 2003, **21**:790-795.
46. Harrington ED, Singh AH, Doerks T, Letunic I, von Mering C, Jensen LJ, Raes J, Bork P: **Quantitative assessment of protein function prediction from metagenomics shotgun sequences.** *Proc Natl Acad Sci USA* 2007, **104**:13913-13918.
47. Turnbaugh PJ, Gordon JI: **An invitation to the marriage of metagenomics and metabolomics.** *Cell* 2008, **134**:708-713.
48. Rogozin IB, Makarova KS, Wolf YI, Koonin EV: **Computational approaches for the analysis of gene neighbourhoods in prokaryotic genomes.** *Brief Bioinform* 2004, **5**:131-149.
49. Suhre K, Claverie JM: **FusionDB: a database for in-depth analysis of prokaryotic gene fusion events.** *Nucleic Acids Res* 2004, **32**:D273-276.
50. Liu Y, Harrison PM, Kunin V, Gerstein M: **Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes.** *Genome Biol* 2004, **5**:R64.
51. Vermunt JKMJ: *Latent Class Cluster Analysis* Cambridge University Press, Cambridge; 2002.
52. Blei DM, Griffiths TL, Jordan MI: **The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies.** *J Acm* 2010, **57**:1-30.
53. Teh YW, Jordan MI, Beal MJ, Blei DM: **Hierarchical Dirichlet processes.** *J Am Stat Assoc* 2006, **101**:1566-1581.
54. Gilks WR, Richardson S, Spiegelhalter DJ: *Markov Chain Monte Carlo In Practice* Chapman & Hall, CRC Interdisciplinary Statistics Series; 1999.
55. Griffiths TL, Steyvers M: **Finding scientific topics.** *Proc Natl Acad Sci USA* 2004, **101(Suppl 1)**:5228-5235.
56. Steyvers M, Griffiths T: **Latent semantic analysis: a road to meaning.**Edited by: Landauer T, McNamara D, Dennis S, Kintsch W. Laurence Erlbaum; 2006:.
57. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG.** *Nucleic Acids Res* 2006, **34**:D354-D357.

doi:10.1186/1471-2105-12-141
Cite this article as: Konietzny *et al.*: Inferring functional modules of protein families with probabilistic topic models. *BMC Bioinformatics* 2011 **12**:141.

CHAPTER 6

Related project

In Konietzny *et al.* (2011), we have demonstrated how to infer functional modules based on the generative model of latent Dirichlet allocation (LDA); however, the concept can be generalized to other sorts of topic models as well. This chapter describes some preliminary results obtained with *collocation* LDA (LDA-COL) (Griffiths *et al.*, 2007), which is an alternative choice of an unsupervised topic model that combines the features of LDA with the ability to consider pairs of adjacent words in documents.

Collocation LDA

Spatial proximity of genes in the DNA sequence is an indicator for functional relationships between the gene products (Section 3.5). One could therefore analyze the spatial relationships of genes, and feed the results as additional knowledge into the process of functional module mining. As a related side note, we made use of this principle by

mapping a set of inferred functional modules of protein families to gene clusters for the evaluation of another inference method (Section 8.4).

Following the concepts presented in Konietzny *et al.* (2011), it is possible to adapt the *collocation* LDA model, thus augmenting the topic model-based framework for functional module inference with the ability to detect conserved spatial gene clusters in genomes. LDA-COL combines the core of LDA – that is, a model for the inference of semantic relationships from co-occurrence patterns of words – with a heuristic mechanism for detecting conserved sequential clusters in sequences of words (Griffiths *et al.*, 2007) (Figure 6.1). In contrast to the LDA model, the order of words thus matters, and we now have to drop the 'bag of words assumption'. With respect to functional module mining, this corresponds to representing genomes as ordered lists of protein family identifiers, where the order of terms reflects the positions of the corresponding gene sequences in the DNA.

Preliminary results and conclusions

We tested LDA and *collocation* LDA with comparable parameter sets on the same genomic input sets to compare their results. The comparisons were based on the performance measures described in Konietzny *et al.* (2011), meaning we used the STRING-based coverage criterion, and mapped the protein families of the inferred modules to KEGG pathways and COG functional categories.

The potential functional modules that were derived from the *collocation* model were on average about three to four times larger than the modules obtained with LDA. Moreover, the modules from the LDA-COL model had a higher fraction of validated functional interactions of their gene family members, showing a functional coverage of more than 70% on average. Therefore, the potential modules inferred with LDA-COL seemed to capture larger functional contexts of gene families than LDA-derived modules.

We next analyzed the KEGG pathway profiles of the LDA-COL modules. It turned out that the modules primarily mapped to highly conserved cellular processes like, for example, components of bacterial ribosomes, membrane-bound transport systems ('ABC-

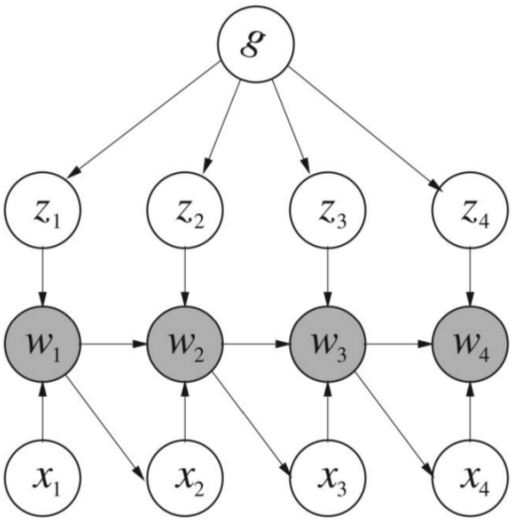


Figure 6.1: Graphical model indicating dependencies among variables in the collocation model. In the *collocation* model, new word instances w_i can be generated from two different types of probability distributions. The first corresponds to the topic distributions in the normal LDA model (represented by the latent z_i variables in the image), whereas the second type is a distribution over words that depends on the previous word in the sequence, and reflects the frequencies of observed word pairs involving the two respective vocabulary items. When a new word gets generated, the choice of the underlying distribution depends on a Bernoulli random variable x_i, and can be imagined as being dependent on the outcome of a coin flip. (Image source: Griffiths *et al.* (2007))

transporters'), and processes of energy turnover such as 'ATP-synthesis'. In addition, carbohydrate and amino acid metabolism processes were matched. The distribution of COG categories for the gene families of the modules was mainly in line with that. The categories [E] ('Amino acid transport and metabolism'), [C] ('Energy production and conversion'), [J] ('Translation, ribosomal structure'), [P] ('Inorganic ion transport'), [H] ('Coenzyme transport and metabolism'), and [G] ('Carbohydrate transport') made up about 70% of all gene families in the modules.

In summary, we clearly obtained better results with LDA-COL, but there were strong indicators that the observed improvements were due to an increased detection of functional modules which are highly conserved in the genomes. This can be explained because conserved cellular processes often correspond to conserved operon and gene

cluster structures (Section 3.3). Since LDA-COL specifically targets collocations of genes, this likely explains why we see a bias towards well conserved functional modules in the results. The types of functional modules for which we see the biggest improvements of LDA-COL thus coincide with the previously observed preferences of phylogenetic profiling methods. Phylogenetic profiling builds on the assumption of evolutionary cohesiveness of functional modules, and therefore works particularly well for the detection of highly conserved cellular processes (Section 3.3). Therefore, the tendency of LDA-COL to detect well conserved cellular processes makes it behave more similar to the classical phylogenetic profiling approach than LDA. However, a final proof of its potential benefits for the detection of functional modules that are evolutionary flexible will need further research. It should be kept in mind that evolutionary flexible functional modules represent particularly interesting biochemical functions (Campillos *et al.*, 2006).

CHAPTER 7

Summary of Part II

We proposed a new framework for functional module inference based on probabilistic topic models. The two key ideas are *i)* learning a topic model on a collection of genomes, and *ii)* interpreting the inferred topic probability distributions as potential functional modules.

In detail, we have developed an unsupervised method, in which the LDA model describes a set of K multinomial distributions that correspond to probabilistic clusters of functionally interacting protein families (Konietzny *et al.* (2011)). The K distributions are latent variables of the model which can be inferred with Gibbs sampling from a collection of protein family identifiers annotated for a set of genomes. The inference process is essentially based on learning from the co-occurrence patterns of protein families across the genomes. For that purpose, the genomes are represented as text documents ('bags of protein family identifiers') in the input of the method. We have demonstrated that the inferred distributions can be converted into sets of protein

families which correspond to potential functional modules, and that the modules capture diverse biological processes, including protein complexes, metabolic pathways, as well as signal transduction cascades. By assumptions of the underlying LDA model, the discovered functional modules represent dominant patterns in the analyzed genomes. This is because the model postulates a single set of topic distributions that are globally defined for all input documents. Thus, the content of a genome gets influenced by a certain subset of the modules, and most modules have shaped larger fractions of the genomes.

Inferring the latent distributions of the topic model, that is, finding the optimal assignments of protein families to potential functional modules is a computationally challenging task that requires approximate Bayesian inference methods such as Gibbs sampling. The results from Gibbs sampling represent estimates of the true distributions, and their variance depends on the convergence rate of the underlying Markov chain. We have accounted for the non-deterministic aspects of the Gibbs sampling process, and demonstrated the stability of the results for several repetitions of the analysis performed on the same input.

We have used KEGG pathway maps and the functional interaction network of the STRING database to map the protein family content of the individual modules to known functional interactions. In this way, we could show that the inferred modules were not randomly formed clusters of unrelated protein families, but instead represented biologically meaningful entities, which showed a high average degree of functional coherence of their contained protein families. With the proposed evaluation scheme, we were also able to validate potential modules that did not map to known processes, and thus represent novel discoveries that are interesting for further studies by biologists. Our results proved the validity of the assumption that functional modules can be inferred from the co-occurrence patterns of protein families. We thus confirmed previous results obtained with phylogenetic profiling methods (Section 3.6). However, in contrast to these methods, the LDA-based method directly targets entire groups of co-occurring elements instead of single pairs. As biological processes may include direct and indirect

interactions of proteins, the group-oriented approach certainly has benefits. In line with that, Schneider *et al.* (2013) recently also came to the conclusion that pairwise-operating methods may miss important functional linkages:

> "These results demonstrate that pairwise analysis is not sufficient to explain the complex relationship between phylogenetic profiles and protein interactions, as there is a large fraction of interacting pairs with no obvious profile pattern." (...) "We found that by considering triplets of proteins, of which one protein is multifunctional, a large fraction of disturbed co-occurrence patterns can be explained." (...) "However, there are likely other interaction scenarios, possibly even higher-order than just triplets, that could explain even more of the seemingly unexpected profile pair relationships."

In our study, the LDA-based approach predicted a large proportion of (transitive) functional interactions that were not predicted by the pairwise genomic context method integrated in the STRING database. We have compared the pairwise functional interactions implied by the modules with the pairwise predictions of the phylogenetic profiling method. In retrospect, an alternative approach would have been to use hierarchical clustering on the predictions of pairwise phylogenetic profiling to compare the modules with the clusters instead. This could have led to more detailed insights about the comparison between LDA and classical phylogenetic profiling. However, our main focus was on testing the hypothesis that the discovered potential functional modules were covering biologically relevant groups of interacting proteins, especially for modules that could not be mapped to known biological processes. This was a challenging task. Moreover, as described in Section 3.6.3, only a few clustering approaches on phylogenetic profiles had been validated at that time, and it wasn't clear whether their results could have served as a reference or not. In any case, it should be noted that there are fundamental differences between LDA and classical phylogenetic profiling. Although the LDA approach is implicitly based on detecting co-occurrence signals in the data, the way this is incorporated into the LDA model is completely different from classical

phylogenetic profiling (Figure 5.2).

Part III

Detection of phenotype-defining genomic elements

CHAPTER 8

Two new methods for solving a biotechnological challenge

8.1 Microbial lignocellulose degradation

Part II of this thesis described a new topic model-based framework for functional module inference. It enables a large-scale analysis of the dominant functional modules that are encoded in a set of microbial genomes.

However, biotechnological applications often need a different perspective on genomic elements. In this setting, the focus is on identifying a small set of protein families or functional modules that are related to specific phenotypic traits of organisms, and therefore represent promising candidates for further research. As wet lab experiments are typically more expensive than computational analyses, being able to perform *in silico* an effective pre-selection of candidate molecules for further laboratory screenings

is important.

Moreover, the characterization of genomic elements that are key factors in phenotype-related processes would help to identify novel microbial species with related molecular activities. Such key factors could be included into computational models to predict an organism's phenotype directly from its genome, without the need for time-consuming and expensive type-screenings in wet lab experiments. This is especially useful for metagenome studies, where we only have partial genome sequences from the members of a microbial community. Thus, and this is important – good prediction methods enable us to characterize the individual fractions ('community members') of a metagenomic binning ensemble.

Microbial lignocellulose degradation is a typical example for the described setting. It represents a complex microbial phenotype, and a process of high biotechnological relevance. Lignocellulose, an integral part of plant cell walls, is an important resource for the production of biofuels (Figure 8.1, Kumar *et al.* (2008)). In a world with limited, rapidly vanishing resources of fossil fuels such as oil and gas, biofuels are important as a renewable energy resource that also provides the benefits of low greenhouse-gas emissions, and a good compatibility with the existing petroleum-based industrial infrastructures (Kohse-Hinghaus *et al.*, 2010).

Biochemical companies already target the genes of lignocellulose-degrading organisms in an attempt to identify unknown biocatalysts, but the current knowledge of the involved biological processes is still limited in many ways (Rubin, 2008; Wilson, 2011). A reason for this is the intricate composite structure of plant cell walls (Figure 8.2), which consist of several intertwined polysaccharides such as cellulose, hemicellulose, and pectins (Figure 8.3).

Metagenomic studies have a strong potential to provide further insights, as it is known that many cellulolytic organisms cannot be cultured in the lab and remain uncharacterized so far (Duan and Feng, 2010). Therefore, there is a clear need for methods that may detect protein families associated with lignocellulose degradation from genomes and metagenomes.

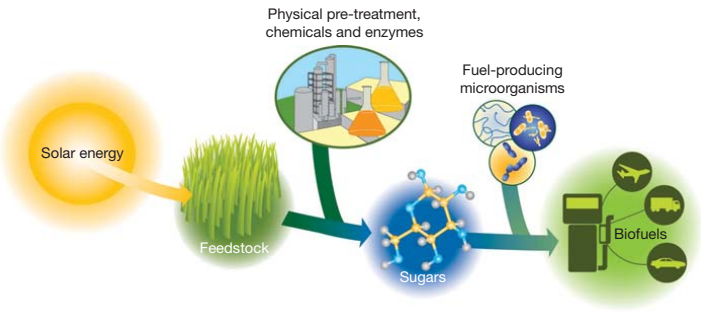


Figure 1 | Biology of bioconversion of solar energy into biofuels. Solar energy is collected by plants via photosynthesis and stored as lignocellulose. Decomposition of the cellulosic material into simple 5- and 6-carbon sugars is achieved by physical and chemical pre-treatment, followed by exposure to enzymes from biomass-degrading organisms. The simple sugars can be subsequently converted into fuels by microorganisms.

Figure 8.1: Bioconversion of solar energy into biofuels. Solar energy gets converted into plant biomass through photosynthesis in cells. It is possible to decompose lignocellulose into its component polysaccharides by using mechanical and chemical procedures. However, common processes typically require the use of high mechanical pressures, high temperatures, and aggressive chemicals. Cellulose can then be converted into biofuels via an intermediate conversion into ethanol. (Image source: Rubin (2008))

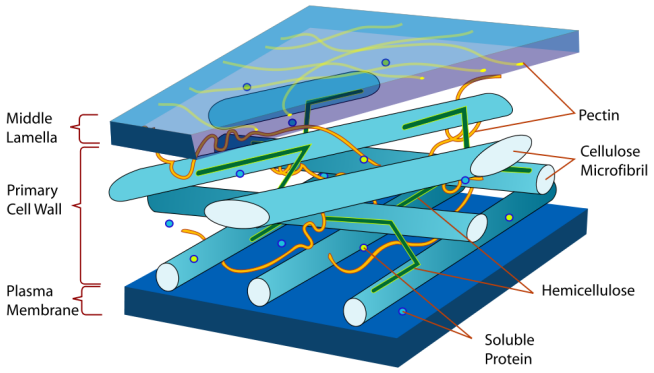


Figure 8.2: The composite structure of the plant cell wall. Lignocellulose makes up the major part of plant cell walls. A variety of complex glycan compounds, such as cellulose, hemicellulose, and pectins are glued together by a three-dimensional polymer, lignin. (Image source [1]: www.wikipedia.org)

[1] Image URL: http://commons.wikimedia.org/wiki/File:Plant_cell_wall_diagram.svg (accessed 27/12/2014)

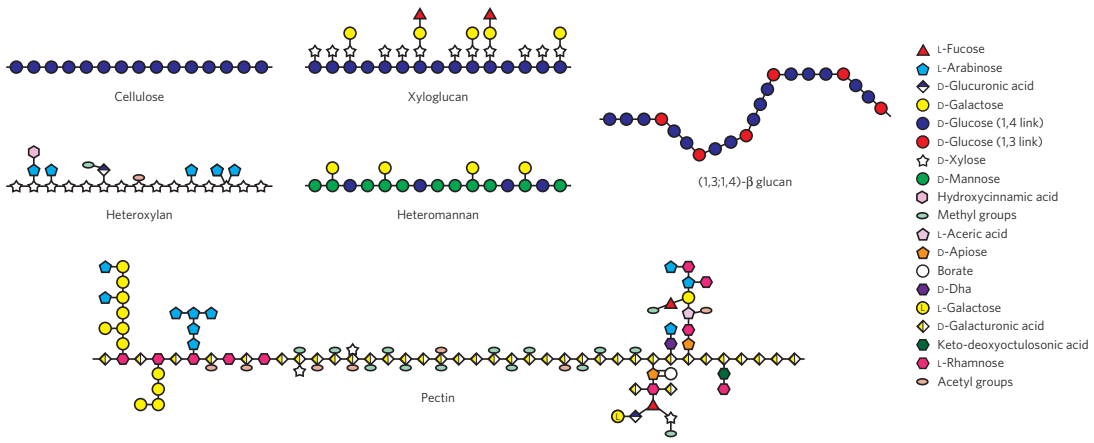


Figure 8.3: Molecular structures of plant cell wall components. The diagram illustrates the composition of the main components of plant cell walls. The degradation of plant cell walls and its main component lignocellulose thus requires a variety of concerted biological subprocesses, that is, functional modules. (Image source: Burton *et al.* (2010))

8.2 Attribute ranking schemes

Phenotype-related genomic elements can be identified with supervised attribute ranking (or selection) schemes. Such methods are typically based on binary classifiers that distinguish phenotype-positive from phenotype-negative genomes[2]. They take as training input the genomic elements ('attributes') of a labeled set of phenotype-positive and phenotype-negative genomes, and get optimized to distinguish between the two phenotype classes. In attribute ranking schemes, the input attributes get scored by their influence on the classification decision, which allows to output a ranking of the attributes. Such a ranking corresponds to a list of genomic elements ordered by their estimated relevance for the observed phenotype.

Lingner *et al.* (2010), and Kastenmüller *et al.* (2009) – among others – had already used attribute ranking schemes to identify phenotype-related genomic elements. But there was still a need for new and improved methods, especially for the *de novo* inference of phenotype-defining functional modules, and the analysis of metagenomic

[2] A phenotype-positive (phenotype-negative) genome belongs to an organism that is known to possess (not possess) the respective phenotype.

data sets. As described in Section 8.1, both aspects play a critical role in the study of microbial lignocellulose degradation. We developed two new methods for the discovery of phenotype-defining elements in genomes and metagenomes; one *family-centric* approach that targets single protein families (Section 8.3), and a *pathway-centric* method that targets entire groups of functionally interacting protein families in the context of functional modules (Section 8.4).

8.3 Publication - Weimann *et al.* 2013

Status	**published**
Journal	Biotechnology for Biofuels (Impact factor in year of submission: 5.552)
Citation	Aaron Weimann, Yulia Trukhina, Phillip B Pope, Sebastian GA Konietzny, Alice C McHardy: De novo prediction of the genomic components and capabilities for microbial plant biomass degradation from (meta-)genomes *Biotechnology for Biofuels* 2013, **6**:24
URL	http://www.biotechnologyforbiofuels.com/content/6/1/24
Own contribution	10% Computation of CAZy and Pfam annotations (with A. Weimann) Identification of lignocellulose-degrading species from the literature for training and testing of the methods (with co-authors)

Weimann et al. Biotechnology for Biofuels 2013, 6:24
http://www.biotechnologyforbiofuels.com/content/6/1/24

Biotechnology for Biofuels

RESEARCH Open Access

De novo prediction of the genomic components and capabilities for microbial plant biomass degradation from (meta-)genomes

Aaron Weimann[1,3†], Yulia Trukhina[1,3†], Phillip B Pope[2], Sebastian GA Konietzny[1,3] and Alice C McHardy[1,3*]

Abstract

Background: Understanding the biological mechanisms used by microorganisms for plant biomass degradation is of considerable biotechnological interest. Despite of the growing number of sequenced (meta)genomes of plant biomass-degrading microbes, there is currently no technique for the systematic determination of the genomic components of this process from these data.

Results: We describe a computational method for the discovery of the protein domains and CAZy families involved in microbial plant biomass degradation. Our method furthermore accurately predicts the capability to degrade plant biomass for microbial species from their genome sequences. Application to a large, manually curated data set of microbial degraders and non-degraders identified gene families of enzymes known by physiological and biochemical tests to be implicated in cellulose degradation, such as GH5 and GH6. Additionally, genes of enzymes that degrade other plant polysaccharides, such as hemicellulose, pectins and oligosaccharides, were found, as well as gene families which have not previously been related to the process. For draft genomes reconstructed from a cow rumen metagenome our method predicted Bacteroidetes-affiliated species and a relative to a known plant biomass degrader to be plant biomass degraders. This was supported by the presence of genes encoding enzymatically active glycoside hydrolases in these genomes.

Conclusions: Our results show the potential of the method for generating novel insights into microbial plant biomass degradation from (meta-)genome data, where there is an increasing production of genome assemblages for uncultured microbes.

Background

Lignocellulosic biomass is the primary component of all plants and one of the most abundant organic compounds on earth. It is a renewable, geographically distributed and a source of sugars, which can subsequently be converted into biofuels with low greenhouse gas emissions, such as ethanol. Chemically, it primarily consists of cellulose, hemicellulose and lignin. Saccharification - the process of degrading lignocellulose into the individual component sugars - is of considerable biotechnological interest. Several mechanical and chemical procedures for saccharification have been established; however, all are relatively expensive, slow and inefficient [1]. An alternative approach is realized in nature by various microorganisms, which use enzyme-driven lignocellulose degradation to generate sugars as sources of carbon and energy. The search for novel enzymes allowing an efficient breakdown of plant biomass has therefore attracted considerable interest [2-5]. In particular, the discovery of novel cellulases for saccharification is considered crucial in this context [6]. However, the complexity of the underlying biological mechanisms and the lack of robust enzymes that can be economically produced in larger quantities currently still prevent industrial application.

For some lignocellulose-degrading species, carbohydrate-active enzymes (CAZymes) and protein domains implicated in lignocellulose degradation are well known. Many of

* Correspondence: alice.mchardy@uni-duesseldorf.de
†Equal contributors
[1]Max-Planck Research Group for Computational Genomics and Epidemiology, Max-Planck Institute for Informatics, University Campus E1 4, Saarbrücken 66123, Germany
[3]Department of Algorithmic Bioinformatics, Heinrich Heine University Düsseldorf, Düsseldorf 40225, Germany
Full list of author information is available at the end of the article

these have been recognized by physiological and biochemical tests as being relevant for the biochemical process of cellulose degradation itself, such as the enzymes of the glycoside hydrolase (GH) families GH6 and GH9 and the endoglucanase-containing family GH5. Two well-studied paradigms are currently known for microbial cellulose degradation: The 'free-enzyme system' is realized in most aerobic microbes and entails secretion of a set of cellulases to the outside of the cell. In anaerobic microorganisms large multi-enzyme complexes, known as cellulosomes, are assembled on the cell surface and catalyze degradation. In both cases, the complete hydrolysis of cellulose requires endoglucanases (GH5 and GH9), which are believed to target non-crystalline regions, and exo-acting cellobiohydrolases, which attack crystalline structures from either the reducing (GH7 and GH48) or non-reducing (GH6) end of the beta-glucan chain. However, in the genomes of some plant biomass-degrading species, homologs of such enzymes have not been found. Recent genome analyses of the lignocellulose-degrading microorganisms, such as the aerobe *Cytophaga hutchinsonii* [7], the anaerobe *Fibrobacter succinogenes* [8,9] and the extreme thermophile anaerobe *Dictyoglomus turgidum* [10] have revealed only GH5 and GH9 endoglucanases. Genes encoding exo-acting cellobiohydrolases (GH6 and GH48) and cellulosome structures (dockerins and cohesins) are absent.

Metagenomics offers the possibility of studying the genetic material of difficult-to-culture (i.e. uncultured) species within microbial communities with the capability to degrade plant biomass. Recent metagenome studies of the gut microbiomes of the wood-degrading higher termites (*Nasutitermes*), the Australian Tammar wallaby (*Macropus eugenii*) [11,12] and two studies of the cow rumen metagenome [13,14] have revealed new insights into the mechanisms of cellulose degradation in uncultured organisms and microbial communities. Microbial communities of different herbivores have been shown to be dominated by lineages affiliated to the Bacteroidetes and Firmicutes, of which different Bacteroidetes lineages exhibited endoglucanse activity [11,15]. Notably, exo-acting families and cellulosomal structures have a low representation or are entirely absent from gut metagenomes sequenced to date. Thus, current knowledge about genes and pathways involved in plant biomass degradation in different species, particularly uncultured microbial ones, is still incomplete.

We describe a method for the *de novo* discovery of protein domains and CAZy families associated with microbial plant biomass degradation from genome and metagenome sequences. It uses protein domain and gene family annotations as input and identifies those domains or gene families, which in concert are most distinctive for the lignocellulose degraders. Among the gene and protein domains identified with our method were known key genes of plant biomass degradation. Additionally, it identified several novel protein domains and gene families as being relevant for the process. These might represent novel leads towards elucidating the mechanisms of plant biomass degradation for the currently less well understood microbial species. Our method furthermore can be used to identify plant biomass-degrading species from the genomes of cultured or uncultured microbes. Application to draft genomes assembled from the metagenome of a switchgrass-adherent microbial community in cow rumen predicted genomes from several Bacteroidales lineages which encode active glycoside hydrolases and a relative to a known plant biomass degrader to represent lignocellulose degraders.

In technical terms, our method selects the most informative features from an ensemble of L1-regularized L2-loss linear Support Vector Machine (SVM) classifiers, trained to distinguish genomes of cellulose-degrading species from non-degrading species based on protein family content. Protein domain annotations are available in public databases and new protein sequences can be rapidly annotated with Hidden Markov Models (HMMs) or - somewhat slower - with BLAST searches of one protein versus the NCBI-nr database [16]. Co-occurrence of protein families in the biomass-degrading fraction of samples and an absence of these families within the non-degrading fraction allows the classifier to link these proteins to biomass degradation *without* requiring sequence homology to known proteins involved in lignocellulose degradation. Classification with SVMs has been previously used successfully for phenotype prediction from genetic variations in genomic data. In Beerenwinkel *et al.* [17], support vector regression models were used for predicting phenotypic drug resistance from genotypes. SVM classification was used by Yosef *et al.* [18] for predicting plasma lipid levels in baboons based on single nucleotide polymorphism data. In Someya *et al.* [19], SVMs were used to predict carbohydrate-binding proteins from amino acid sequences. The SVM [20,21] is a discriminative learning method that infers, in a supervised fashion, the relationship between input features (such as the distribution of conserved gene clusters or single nucleotide polymorphisms across a set of sequence samples) and a target variable, such as a certain phenotype, from labeled training data. The inferred function is subsequently used to predict the value of this target variable for new data points. This type of method makes no *a priori* assumptions about the problem domain. SVMs can be applied to datasets with millions of input features and have good generalization abilities, in that models inferred from small amounts of training data show good predictive accuracy on novel data. The use of models that include an L1-regularization term favors solutions in which few

features are required for accurate prediction. There are several reasons why sparseness is desirable: the high dimensionality of many real datasets results in great challenges for processing. Many features in these datasets are usually non-informative or noisy, and a sparse classifier can lead to a faster prediction. In some applications, like ours, a small set of relevant features is desirable because it allows direct interpretation of the results.

Results

We trained an ensemble of SVM classifiers to distinguish between plant biomass-degrading and non-degrading microorganisms based on either Pfam domain or CAZY gene family annotations (see Methods section for the training and evaluation of the SVM classification ensemble). We used a manually curated data set of 104 microbial (meta-)genome sequence samples for this purpose, which included 19 genomes and 3 metagenomes of lignocellulose degraders and 82 genomes of non-degraders (Figure 1, Figure 2, Additional file 1: Table S1). Fungi are known to use several enzymes for plant biomass degradation for which the corresponding genes are not found in prokaryotic genomes and vice versa, while other genes are shared by prokaryotic and eukaryotic degraders. To investigate similarities and differences detectable with our method, we included the genome of lignocellulose degrading fungus *Postia placenta* into our analysis. After training, we identified the most distinctive protein domains and CAZy families of plant biomass degraders from the resulting models. We compared these protein domains and gene families with known plant biomass degradation genes. We furthermore applied our method to identify plant biomass degraders among 15 draft genomes from the metagenome of a microbial community adherent to switch grass in cow rumen.

Distinctive Pfam domains of microbial plant biomass degraders

For the training of a classifier which distinguishes between plant biomass-degrading and non-degrading microorganisms we used Pfam annotations of 101 microbial genomes and two metagenomes. This included metagenomes of microbial communities from the gut of a wood-degrading higher termite and from the foregut of the Australian Tammar Wallaby as examples for plant biomass-degrading communities. Furthermore, 19 genomes of microbial lignocellulose degraders were included - of the phyla Firmicutes (7 isolate genome sequences), Actinobacteria (5), Proteobacteria (3), Bacteroidetes (1), Fibrobacteres (1), Dictyoglomi (1) and Basidiomycota (1). Eighty-two microbial genomes annotated to not possess the capability to degrade lignocellulose were used as examples of non-lignocellulose-degrading microbial species (Additional file 1: Table S1).

We assessed the value of information about the presence or absence of protein domains for distinguishing lignocellulose degraders from non-degraders. With the respective classifier, $eSVM_{bPFAM}$, each microbial (meta-) genome sequence was represented by a feature vector with the features indicating the presence or absence of Pfam domains (see Methods). The nested cross-validation macro-accuracy of $eSVM_{bPFAM}$ in distinguishing plant biomass-degrading from non-degrading microorganisms was 0.91. This corresponds to 94% (97 of 103) of the (meta-)genome sequences being classified correctly. Only three of the 21 cellulose-degrading samples and three of the non-degraders were misclassified (Table 1, Table 2). Among these were four Actinobacteria and one genome affiliated with the Basidiomycota and Theromotogae each.

We identified the Pfam domains with the greatest importance for assignment to the lignocellulose-degrading class by $eSVM_{bPFAM}$ (Figure 1; see Methods for the feature selection algorithm). Among these are several protein domains known to be relevant for plant biomass degradation. One of them is the GH5 family, which is present in all of the plant biomass-degrading samples. Almost all activities determined within this family are relevant to plant biomass degradation. Because of its functional diversity, a subfamily classification of the GH5 family was recently proposed [24]. The carbohydrate-binding modules CBM_6 and CBM_4_9 were also selected. Both families are Type B carbohydrate-binding modules (CBMs), which exhibit a wide range of specificities, recognizing single glycan chains comprising hemicellulose (xylans, mannans, galactans and glucans of mixed linkages) and/or non-crystalline cellulose [25]. Type A CBMs (e.g. CBM2 and CBM3), which are more commonly associated with binding to insoluble, highly crystalline cellulose, were not identified as relevant by $eSVM_{bPFAM}$. Furthermore, numerous enzymes that degrade non-cellulosic plant structural polysaccharides were identified, including those that attack the backbone and side chains of hemicellulosic polysaccharides. Examples include the GH10 xylanases and GH26 mannanases. Additionally, enzymes that generally display specificity for oligosaccharides were selected, including GH39 β-xylosidases and GH3 enzymes.

We subsequently trained a classifier - $eSVM_{fPFAM}$ - with a weighted representation of Pfam domain frequencies for the same data set. The macro-accuracy of $eSVM_{fPFAM}$ was 0.84 (Table 2); lower than that of the $eSVM_{bPFAM}$; with nine misclassified samples (4 Actinobacteria, 2 Bacteroidetes, 1 Basidiomycota, 1 Thermotogae phyla and the Tammar Wallaby metagenome). Again, we determined the most relevant protein domains for identifying a plant biomass-degrading sequence sample from the models by

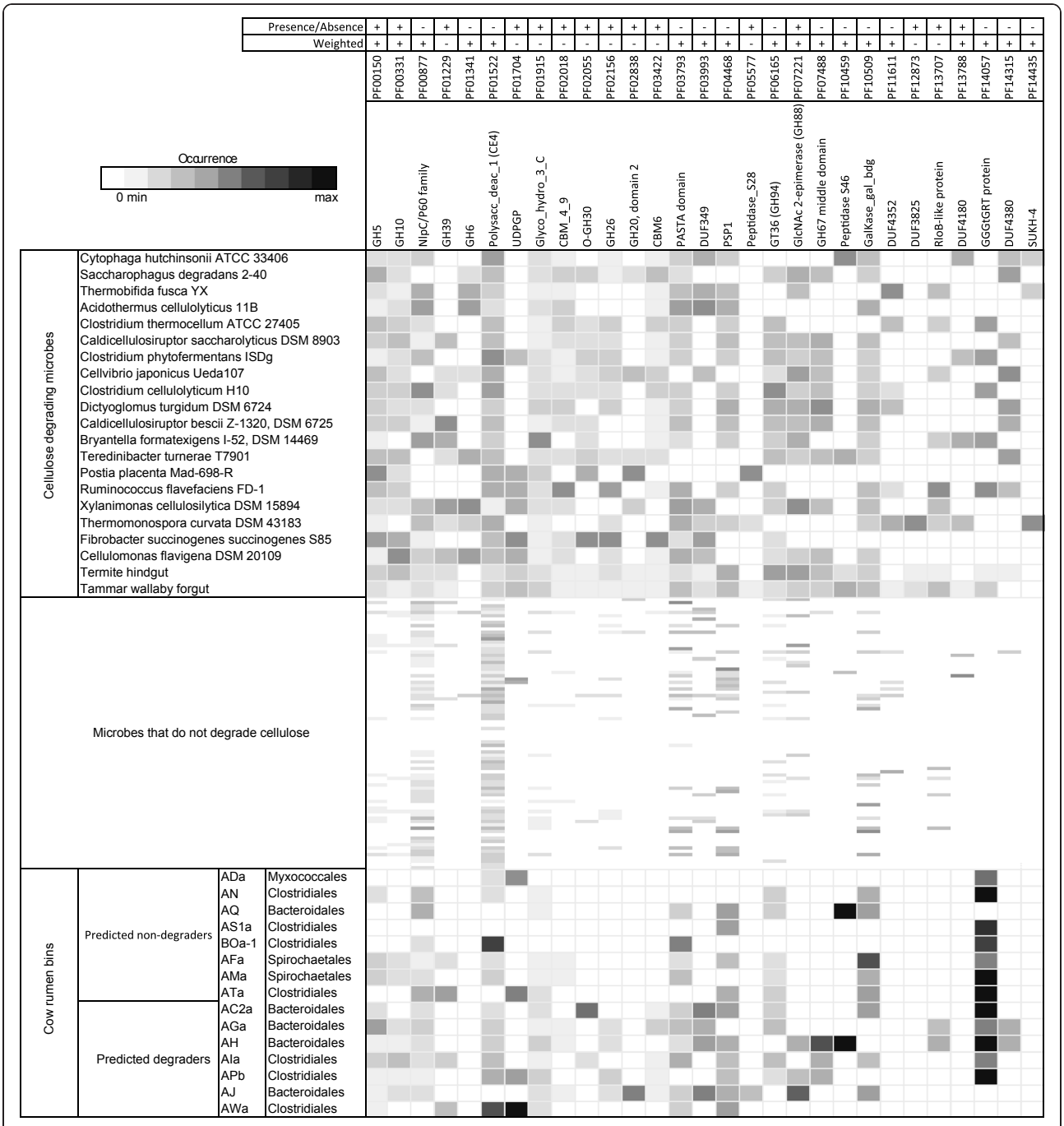


Figure 1 Frequencies of the selected Pfam families in the individual genomes and metagenomes. The data for each entry are rescaled by the total number of Pfam domains annotated to the microbial genome or metagenome. The color scale from grey to black indicates domain families that are present in low to high amounts, respectively. White indicates absent protein domains. The signs "+" and "-" indicate whether a protein domain was chosen in the respective experiment.

feature selection. Among the most important protein families were, as before, GH5, GH10 and GH88 (PF07221: N-acylglucosamine 2-epimerase) (Figure 1). GH6, GH67 and CE4 acetyl xylan esterases ("accessory enzymes" that contribute towards complete hydrolysis of xylan) were only relevant for prediction with the $eSVM_{fPFAM}$ classifier. Additionally, both models specified protein domains not commonly associated with plant biomass degradation as being relevant for assignment, such as the lipoproteins DUF4352 and PF00877 (NlpC/P60 family) and binding domains PF10509 (galactose-binding signature domain) and PF03793 (PASTA domain) (Figure 1).

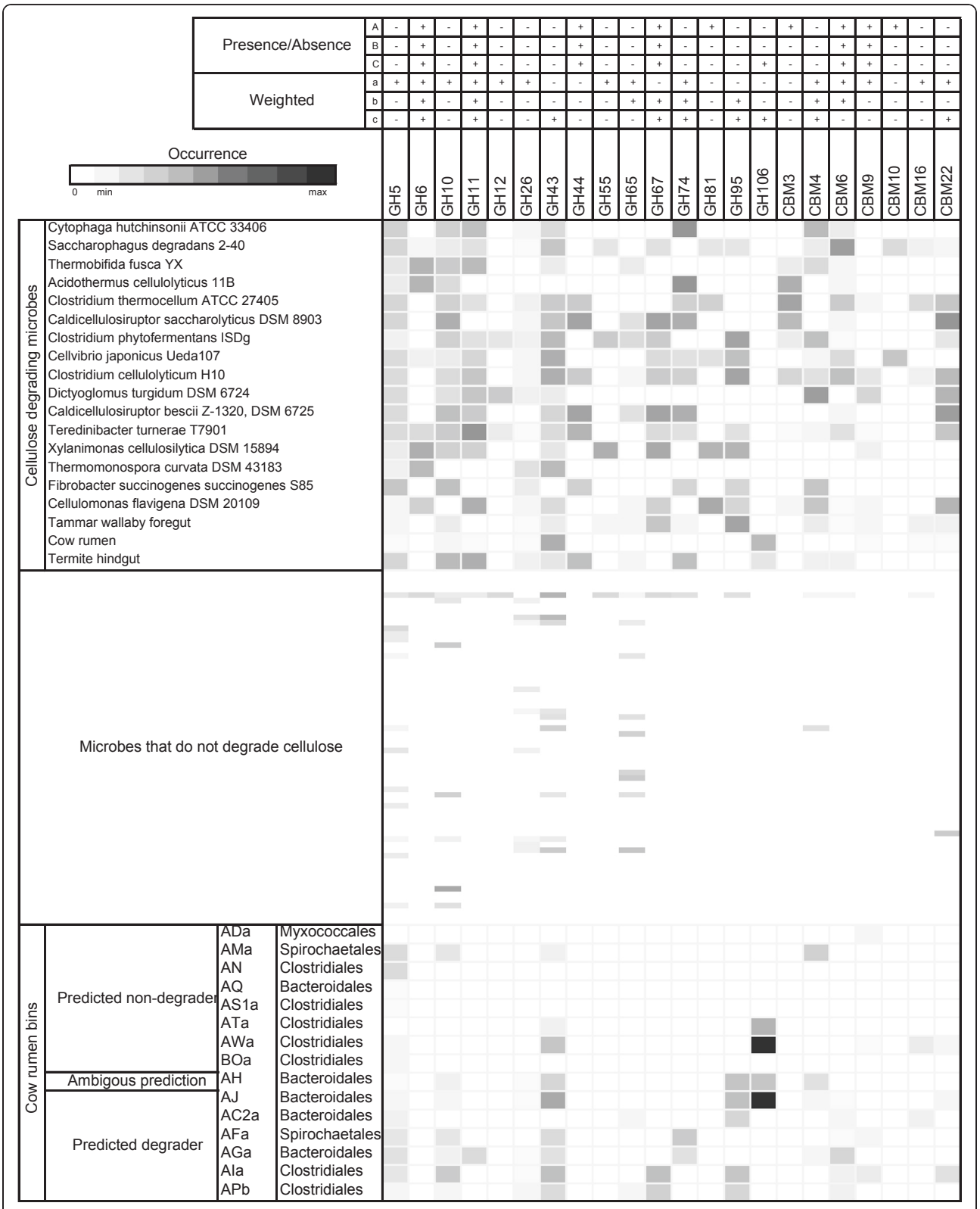


Figure 2 Frequencies of selected glycoside hydrolase (GH) families and carbohydrate binding modules (CBMs) in the (meta-) genome sequences. The data for each entry are rescaled by the total number of GH and CBM domains annotated to the microbial genome or metagenome. The coloring from black to grey indicates domains that are present in high to low amounts, respectively. White indicates absent domain families ("A", "a", "B", "b", "C", "c" as described in Table 1).

Table 1 Misclassified species in the SVM analyses

	$eSVM_{bPFAM}$	$eSVM_{CAZY\_B}$
False negatives	*Postia placenta Mad-698-R*	*Thermomonospora curvata DSM 43183*
	Xylanimonas cellulosilytica DSM 15894	
	Thermomonospora curvata DSM 43183	
False positives	*Actinosynnema mirum 101*	*Actinosynnema mirum 101*
	Arthrobacter aurescens TC1	
	Thermotoga lettingae TMO	

Shown are species which were misclassified with the $eSVM_{CAZY\_B}$ and the $eSVM_{bPFAM}$ classifiers. Contrary to previous beliefs [22], recent literature indicates in agreement with our predictions that *T. curvata* is a non-degrader. Furthermore, recent evidence supports that *A. mirum* is a lignocellulose degrader, which has not been previously described [23].

Distinctive CAZy families of microbial plant biomass degraders

We searched for distinctive CAZy families of microbial plant biomass degraders with our method. CAZy families include glycoside hydrolases (GH), carbohydrate-binding modules (CBM), glycosyltransferases (GT), polysaccharide lyases (PL) and carbohydrate esterases (CE). The annotations from the CAZy database comprised 64 genomes of non-lignocellulose-degrading species and 16 genomes of lignocellulose-degraders. There were no CAZy annotations available for the remaining genomes. In addition, we included the metagenomes of the gut microbiomes of the Tammar wallaby (TW), the wood-degrading higher termite and of the cow rumen microbiome (Additional file 1: Table S1). We evaluated the value of information about the presence or absence of CAZy domains, or of their relative frequencies for identification of lignocellulose-degrading microbial (meta-)genomes in the following experiments:

1) By training of the classifiers $eSVM_{CAZY\_A}$ (presence/absence) and $eSVM_{CAZY\_a}$ (counts), based on genome annotations with all CAZy families.
2) By training of the classifiers $eSVM_{CAZY\_B}$ (presence/absence) and $eSVM_{CAZY\_b}$ (counts), based on the annotations of the genomes and the TW sample with all CAZy families, except for the GT family members, which were not annotated for the TW sample.
3) By training of the classifiers $eSVM_{CAZY\_C}$ (presence/absence) and $eSVM_{CAZY\_c}$ (counts) with the entire data set based on GH family and CBM annotations, as these were the only ones available for the three metagenomes.

The macro-accuracy of these classifiers ranged from 0.87 to 0.96, similar to the Pfam-domain-based models (Table 2). Notably, almost exclusively Actinobacteria were misclassified by the $eSVM_{CAZY}$ classifiers, except for the Firmicute *Caldicellulosiruptor saccharolyticus*. The best classification results were obtained with the presence-absence information for all CAZy families except for the GT families of the microbial genomes and the TW sample. In this setting ($eSVM_{CAZY\_B}$) only two species (*Thermomonospora curvata* and *Actinosynnema mirum*) were misclassified (Table 1). These species remained misclassified with all six classifiers.

Using feature selection, we determined the CAZy families from the six $eSVM_{CAZy}$ classifiers that are most relevant for identifying microbial cellulose-degraders. Many of these GH families and CBMs are present in all (meta-) genomes (Figure 2). This analysis identified further gene families known to be relevant for plant biomass degradation. Among them are cellulase-containing families (GH5, GH6, GH12, GH44, GH74), hemicellulase-containing families (GH10, GH11, GH26, GH55, GH81, GH115), families with known oligosaccharide/side-chain-degrading activities (GH43, GH65, GH67, GH95) and several CBMs (CBM3, -4, -6, -9, -10, -16, -22, -56). Several of these (GH6, GH11, GH44, GH67, GH74, CBM4, CBM6, CBM9) were consistently identified by at least half of the six classifiers as distinctive for plant biomass degraders. These might be considered signature genes of the plant biomass-degrading microorganisms we analyzed. Additionally, several GT, PL and CE domains were identified as relevant ($eSVM_{CAZY\_A}$: PL1, PL11 and CE5, "$eSVM_{CAZY\_B}$: CE5; $eSVM_{CAZY\_a}$: GT39, PL1 and CE2, $eSVM_{CAZY\_b}$: none). These CAZy families, as well as GH115 and CBM56, are not included in Figure 2, as they are not annotated for all sequences.

Identification of plant biomass degraders from a cow rumen metagenome

We used our method to predict the plant biomass-degrading capabilities for 15 draft genomes of uncultured microbes reconstructed from the metagenome of a microbial community adherent to switchgrass in cow rumen [14] (see Methods for the classification with an ensemble of SVM classifiers). The draft genomes represent genomes with more than 50% of the sequence reconstructed by taxonomic binning of the metagenome

Table 2 Accuracy of classifying microbes as lignocellulose-degraders or non-degraders

	Presence/absence of Pfam domains	Weighted Pfam domain representation	Presence/absence CAZy family representation			Weighted CAZy family representation		
			A	B	C	a	b	c
nCV macro-accuracy	0.91	0.84	0.90	0.96	0.94	0.91	0.93	0.87
nCV recall	0.86	0.73	0.81	0.94	0.90	0.88	0.88	0.79
nCV true negative rate	0.96	0.96	0.98	0.98	0.98	0.95	0.98	0.95

L1-regularized SVMs were trained with Pfam domain or CAZY family (meta-)genome annotations. Capital letters denote classifiers trained based on the presence or absence of CAZy families and small letters indicate classifiers trained based on the relative abundances of CAZy families in annotations. Abbreviations "A", "a"," B", "b", "C", "c" denote the following: Classifiers "A","a" were trained with annotations of all CAZy families for 16 microbial genomes; Classifiers "B","b" were trained with annotations for all CAZy families, except for the GT family members (which were not annotated for the Tammar Wallaby metagenome), for 16 genomes and the TW metagenome of plant biomass degraders; Classifiers "C","c" were trained with annotations for the GH families and CBMs for the 16 microbial genomes and three metagenomes of plant biomass degraders, as only these were annotated for the metagenomes. All CAZy-based classifiers were trained with available annotations for 64 genomes of non-biomass degraders. The Pfam-based classifiers were trained with 21 (meta-)genomes of biomass-degraders and 82 microbial genomes of non-degraders. For more details on the experimental set-up and the evaluation measures shown see the Methods section on performance evaluation.

sample. The microbial community adherent to switchgrass is likely to be enriched with plant biomass degraders, as it was found to differ from the rumen fluid community in its taxonomic composition and degradation of switch grass after incubation in cow rumen had occurred. For identification of plant biomass-degrading microbes, we classified each draft genome individually with the $eSVM_{bPFAM}$ and $eSVM_{CAZY\_B}$ models, which had the highest macro-accuracy based on Pfam domain or CAZy family annotations, respectively. The $eSVM_{bPFAM}$ classifier assigned seven of the draft genomes to plant biomass degraders (Table 3). One of these, genome *APb*, was found by 16S rRNA analysis to be related to the fibrolytic species *Butyrivibrio fibrisolvens*. Four others (*AC2a*, *AGa*, *AJ* and *AH*) are of the order of Bacteroidales, and include all but one draft genomes affiliated to the Bacteroidales. The 6th and 7th predicted degrader, represented by genome *AIa* and *AWa*, belong to the Clostridiales, like genome *APb*. The $eSVM_{CAZY\_B}$ classifier also assigned five of these genomes to the plant biomass degraders. Additionally it classified genome *AH* as plant biomass-degrading, while being ambiguous in the assignment of *AFa* (Table 3). To validate these predictions, we searched the draft genomes for genes encoding 51 enzymatically active glycoside hydrolases characterized from the same rumen dataset (for the results of these experiments see Figure three in Hess *et al.* [14]). Genomes *AGa*, *AC2a*, *AJ* and *AIa* were all linked to different enzymes of varying specificities (Table 3). *AC2a* was linked to cellulose degradation, specifically to a carboxymethyl cellulose (CMC)-degrading GH5 endoglucanase as well as GH9 enzyme capable of degrading insoluble cellulosic substrates such as Avicel®. *AIa* demonstrated capabilities towards xylan and soluble cellulosic substrates with affiliations to four GH10 xylanases. Both *AGa* and *AJ* demonstrated broader substrate versatility and were linked to enzymes with capabilities towards cellulosic substrates CMC and Avicel® (GH5, GH9 and GH26), hemicellulosic substrates lichenan (β-1,3, β-1,4 β-glucan) and xylan (GH5, GH9 and GH10), as well as the natural feedstocks miscanthus and switchgrass (GH5 and GH9). Importantly, no carbohydrate-active enzymes were affiliated to draft genomes that were predicted to not possess plant biomass-degrading capabilities (Table 3). Overall, assignments were largely consistent between the two classifiers and supporting evidence for the capability to degrade plant biomass was found for five of the predicted degraders.

Timing experiments

Our method uses annotations with Pfam domains or CAZy families as input. Generating these by similarity-searches with profile HMMs rather than with BLAST provides a better scalability for next-generation sequencing data sets. HMM databases such as dbCAN contain a representation of entire protein families rather than of individual gene family members, which largely decreases the number of entries one has to compare against. For example, searching the ORFs of the *Fibrobacter succinogenes* genome [26] for similarities to CAZy families with the dbCAN HMM models took 23 seconds on an Intel® Xeon® 1.6 GHz CPU. In comparison, searching for similarities to CAZy families by BLASTing the same set of ORFs against all sequences with CAZy family annotation of the NCBI non-redundant protein database (downloaded from http://csbl.bmb.uga.edu/dbCAN/ on April 19th 2011) on the same machine required approximately 1 hour and 55 minutes, a difference of two orders of magnitude. Because of their better scalability and also because they are well-established for identifying protein domains or gene families [27-29], we recommend the use of HMM-based similarities and annotations as input to our method.

Discussion

We investigated the value of information about the presence-or-absence of CAZy families and Pfam protein domains, as well as information about their relative abundances, for the identification of lignocellulose degraders. Classifiers trained with CAZy family or Pfam

Table 3 Prediction of the plant biomass degradation capabilities for 15 draft genomes

	AC2a	AGa	Ala-2	AJ	APb	AFa	AH	AWa	ADa	AMa	AN	AQ	AS1	ATa	BOa
$eSVM_{CAZY\_B}$	++	++	++	+	++	++	0	- -	- -	- -	- -	- -	- -	- -	- -
$eSVM_{bPFAM}$	++	++	++	++	++	-	++	+	- -	-	- -	- -	- -	-	- -
CMC	GH5 (TW-33)	GH5 (TW-40) GH5 (MH-2)	GH10 (TW-34)	GH5 (TW-39) GH26 (TW-10) GH10 (TW-8)											
XYL		GH10 (TW-25)	GH10 (TW-30) GH10 (TW-31) GH10 (TW-37)	GH10 (TW-8)											
SWG		GH5 (TW-40) GH5 (MH-2)													
MIS	GH9 (TW-64)	GH5 (TW-40) GH5 (MH-2) GH9 (TW-50)		GH5 (TW-39)											
AVI	GH9 (TW-64)	GH5 (TW-40) GH5 (MH-2) GH9 (TW-50)		GH5 (TW-39)											
LIC		GH5 (TW-40) GH5 (MH-2) GH9 (TW-50)		GH5 (TW-39)											

Genome reconstructions from the metagenome of a microbial community adherent to switchgrass in the cow rumen were obtained by taxonomic binning of assembled sequences in the original study. Symbols depict the prediction outcome of a voting committee of the 5 $eSVM_{CAZY\_B}$ and the $eSVM_{bPFAM}$ classifiers with the best macro-accuracy (see text for the description of the classifiers). ++: genome classified as plant biomass degrader by all classifiers; +: genome classified as plant biomass degrader by 4 out of 5 classifiers; 0: ambiguous prediction; -: genome classified as not plant biomass degrader by 4 out of 5 classifiers; --: genome classified as not plant biomass degrader by all classifiers. For every draft genome, the presence of genes encoding glycoside hydrolases with verified enzymatic activity for different substrates in this study [14] is indicated. The genome and substrate names correspond to those of Figure 3 and Table S6 of the study.
Hydrolytic activity detected on:
(CMC) 1% (w/v) carboxymethyl cellulose agar.
(XYL) 1% (w/v) Xylan.
(SWG) 1% (w/v) IL-Switchgrass.
(MIS) 1% (w/v) IL-Miscanthus.
(AVI) 1% (w/v) IL-Avicel.

domain annotations allowed an accurate identification of plant biomass degraders and determined similar domains and CAZy families as being most distinctive. Many of these are recognized by physiological and biochemical tests as being relevant for the biochemical process of cellulose degradation itself, such as GH6, members of the GH5 family and to a lesser extent GH44 and GH74. In contrast to widely accepted paradigms for microbial cellulose degradation, recent genome analysis of cellulolytic bacteria has identified examples (i.e. *Fibrobacter*) where there is an absence of genes encoding exo-acting cellobiohydrolases (GH6 and GH48) and cellulosome structures [30]. In addition, these exo-acting families and cellulosomal structures have had a low representation or are entirely absent from sequenced gut metagenomes. Our method also finds the exo-acting cellobiohydrolases GH7 and GH48 to be less important. GH7 represents fungal enzymes, so its absence makes sense; however, the lower importance assigned to GH48 is interesting. The role of GH48 is believed to be of high importance, although recent research has raised questions. Olson *et al.* [31] have found that a complete solubilization of crystalline cellulose can occur in *Clostridium thermocellum* without the expression of GH48, albeit at significantly lower rates. Furthermore, genome analysis of cellulose-degrading microbes *Cellvibrio japonicus* [32] and *Saccharophagus degradans* [33] have determined the presence of only non-reducing end enzymes (GH6) and an absence of a reducing end cellobiohydrolase (GH48), suggesting that the latter are not essential for all cellulolytic enzyme systems.

While we have focused on cellulose degradation, our method has also identified enzymes that degrade other plant polysaccharides as being relevant, such as hemicellulose (GH10, GH11, GH12, GH26, GH55, GH81, CE4), pectins (PL1, GH88 and GH43), oligosaccharides (GH3, GH30, GH39, GH43, GH65, GH95) and the side-chains attached to noncellulosic polysaccharides (GH67, GH88, GH106). This was expected, since many cellulose-degrading microbes produce a repertoire of different glycoside hydrolases, lyases and esterases (see, for example, [32,33]) that target the numerous linkages that are present within different plant polysaccharides, which often exist in tight cross-linked forms within the plant cell wall. The results from our method add further weight to this. The observation of numerous CBMs being relevant in the CAZy analysis also agrees with previous findings that many different CBM-GH combinations are possible in bacteria. Moreover, recent reports have demonstrated that the targeting actions of CBMs have strong proximity effects within cell wall structures, i.e. CBMs directed to a cell wall polysaccharide (e.g. cellulose) other than the target substrate of their appended glycoside hydrolase (e.g. xylanase) can promote enzyme action against the target substrate (e.g. xylan) within the cell wall [34]. This provides explanations as to why cellulose-directed CBMs are appended to many non-cellulase cell wall hydrolases.

Several Pfam domains of unknown function (DUFs) or protein domains which have not previously been associated with cellulose degradation are predicted as being relevant. These include transferases (PF01704) and several putative lipoproteins (DUF4352), some of which have predicted binding properties (NlpC/P60 family: PF00877, PASTA domain: PF03793). The functions of these domains in relation to cellulose degradation are not known, but possibilities include binding to cellulose, binding to other components of the cellulolytic machinery or interaction with the cell surface.

Another result of our study are the classifiers for identifying microbial lignocellulose-degraders from genomes of cultured and uncultured microbial species reconstructed from metagenomes. Classification of draft genomes reconstructed from switchgrass-adherent microbes from cow rumen with the most accurate classifiers predicted six or seven of these to represent plant biomass-degrading microbes, including a close relative to the fibrolytic species *Butyrivibrio fibrisolvens*. Cross-referencing of all draft genomes against a catalogue of enzymatically active glycoside hydrolases provided a degree of method validation and was in majority agreement with our predictions. Four genomes (*AGa*, *AC2a*, *AJ* and *AIa*) predicted positive were linked to cellulolytic and/or hemicellulolytic enzymes, and importantly no genomes that were predicted negative were linked to carbohydrate-active enzymes from that catalogue of enzymatically active enzymes. Also, no connections to carbohydrate-active enzymes from that catalogue were observed for the three genomes (*AFa*,*AH* and *AWa*) where ambiguous predictions were made. As both draft genomes as well as the catalogue of carbohydrate active enzymes in cow rumen are incomplete, in addition to our training data not covering all plant-biomass-degrading taxa, such ambiguous assignments might be better resolvable with more information in the future.

We trained a previous version of our classifier with the genome of *Methanosarcina barkeri fusaro* incorrectly labeled as a plant biomass degrader, according to information provided by IMG. In cross-validation experiments, our method correctly assigned *M. barkeri* to be a non-plant biomass-degrading species. We labeled *Thermonospora curvata* as a plant biomass degrader and *Actinosynnema mirum* as non-degrader according to information from the literature (see Additional file 1: Table S1). Both were misassigned by all classifiers in the cross-validation experiments. However, in a recent work by Anderson *et al.* [23] it was shown that in cellulose activity assays *A. mirum* could degrade various cellulose substrates. In the same study, *T. curvata* did not show cellulolytic activity against

any of these substrates, contrary to previous beliefs [22]. The authors found out that the cellulolytic *T. curvata* strain was in fact a *T. fusca* strain. Thus, our method could correctly assign both strains despite of the incorrect phenotypic labeling. The genome of *Postia placenta*, the only fungal plant biomass degrader of our data set was misassigned in the Pfam-based SVM analyses. Fungi possess cellulases not found in prokaryotic species [35] and might employ a different mechanism for plant biomass degradation [36,37]. Indeed, in our data set, *Postia placenta* is annotated with the cellulase-containing GH5 family and xylanase GH10, but the hemicellulase family GH26 does not occur. Furthermore, the (hemi-)cellulose binding CBM domains CBM6 and CBM_4_9, which were identified as being relevant for assignment to lignocellulose degraders with the $eSVM_{bPFAM}$ classifier, are absent. All of the latter ones, GH26, CBM6 and especially CBM4 and CBM9, occur very rarely in eukaryotic genome annotations, according to the CAZy database.

Conclusions

We have developed a computational technique for the identification of Pfam protein domains and CAZy families that are distinctive for microbial plant biomass degradation from (meta-)genome sequences and for predicting whether a (draft) genome of cultured or uncultured microorganisms encodes a plant biomass-degrading organism. Our method is based on feature selection from an ensemble of linear L1-regularized SVMs. It is sufficiently accurate to detect errors in phenotype assignments of microbial genomes. However, some microbial species remained misclassified in our analysis, which indicates that further distinctive genes and pathways for plant biomass degradation are currently poorly represented in the data and could therefore not be identified.

To identify a lignocellulose degrader from the currently available data, the presence of a few domains, many of which are already known, is sufficient. The identification of several protein domains which have so far not been associated with microbial plant biomass degradation in the Pfam-based SVM analyses as being relevant may warrant further scrutiny. A difficulty in our study was to generate a sufficiently large and correctly annotated dataset to reach reliable conclusions. This means that the results could probably be further improved in the future, as more sequences and information on plant biomass degraders become available. The method will probably also be suitable for identifying relevant gene and protein families of other phenotypes.

The prediction and subsequent validation of three Bacteroidales genomes to represent cellulose-degrading species demonstrates the value of our technique for the identification of plant biomass degraders from draft genomes from complex microbial communities, where there is an increasing production of genome assemblages for uncultured microbes. These to our knowledge represent the first cellulolytic Bacteroidetes-affiliated lineages described from herbivore gut environments. This finding has the potential to influence future cellulolytic activity investigations within rumen microbiomes, which has for the greater part been attributed to the metabolic capabilities of species affiliated to the bacterial phyla Firmicutes and Fibrobacteres.

Methods

Annotation

We annotated all protein coding sequences of microbial genomes and metagenomes with Pfam protein domains (Pfam-A 26.0) and Carbohydrate-Active Enzymes (CAZymes) [28,38]. The CAZy database contains information on families of structurally related catalytic modules and carbohydrate binding modules (CBMs) or (functional) domains of enzymes that degrade, modify or create glycosidic bonds. HMMs for the Pfam domains were downloaded from the Pfam database. Microbial and metagenomic protein sequences were retrieved from IMG 3.4 and IMG/M 3.3 [39,40]. HMMER 3 [41] with gathering thresholds was used to annotate the samples with Pfam domains. Each Pfam family has a manually defined gathering threshold for the bit score that was set in such a way that there were no false-positives detected. For annotation of protein sequences with CAZy families, the available annotations from the database were used. For annotations not available in the database, HMMs for the CAZy families were downloaded from dbCAN (http://csbl.bmb.uga.edu/dbcan) [42]. To be considered a valid annotation, matches to Pfam and dbCAN protein domain HMMs in the protein sequences were required to be supported by an e-value of at least 1e-02 and a bit score of at least 25. Additionally, we excluded matches to dbCAN HMMs with an alignment longer than 100 bp that did not exceed an e-value of 1e-04. Multiple matches of one and the same protein sequence against a single Pfam or dbCAN HMM exceeding the thresholds were counted as one annotation.

Phenotype annotation of lignocellulose-degrading and non-degrading microbes

We defined genomes and metagenomes as originating from either lignocellulose-degrading or non-lignocellulose-degrading microbial species based on information provided by IMG/M and in the literature. For every microbial genome and metagenome, we downloaded the genome publication and further available articles (Additional file 1: Table S1). We did not consider genomes for which no publications were available. For cellulose-degrading species annotated in IMG, we verified these assignments based on these publications. We used text search to

identify the keywords "cellulose", "cellulase", "carbon source", "plant cell wall" or "polysaccharide" in the publications for non-cellulose-degrading species. We subsequently read all articles that contained these keywords in detail to classify the respective organism as either cellulose-degrading or non-degrading. Genomes that could not be unambiguously classified in this manner were excluded from our study.

Classification with an ensemble of support vector machine classifiers

The SVM is a supervised learning method that can be used for data classification [20,21]. Here, we use an L1-regularized L2-loss SVM, which solves the following optimization problem for a set of instance-label pairs $(\vec{x_i}, y_i)$, $\vec{x_i} \in R^n$, $y_i \in \{-1, +1\}$, $i = 1, \ldots, l$:

$$\min_{\vec{w}} || \vec{w} ||_1 + C \sum_{i=1}^{l} \left(\max(0, 1 - y_i \vec{w}^T \vec{x_i}) \right)^2, \quad (1)$$

where $C \geq 0$ is a penalty parameter. This choice of the classifier and regularization term results in sparse models, where non-zero components of the weight vector $\vec{w}$ are important for discrimination between the classes [43]. SVM classification was performed using the LIBLINEAR package [44]. The components of $\vec{x_i}$ are either binary valued and represent the presence or absence of protein domains, or continuous-valued and represent the frequency of a particular protein domain or gene family relative to the total number of annotations. All features were normalized by dividing by the sum of all vector entries and subsequently scaled, such that the value of each feature was within the range [0,1]. The label +1 was assigned to genomes and metagenomes of plant biomass-degrading microorganisms, the label -1 to all sequences from non-degrading ones. Classification of the draft genomes assembled from the fiber-adherent microbial community from cow rumen was performed with a voting committee of multiple models with different settings for the penalty parameter C that performed comparably well. A majority vote of the 5 most accurate models was used here obtained in a single cross-validation run with different settings of the penalty parameter C.

Performance evaluation

The assignment accuracy of a classifier was determined with a standard nested cross-validation (nCV) setup [45]. In nCV, an outer cross-validation loop is organized according to the leave-one-out principle: In each step, one data point is left out. In an inner loop, the optimal parameters for the model (here, the penalty parameter C) are sought, in a second cross-validation experiment with the remaining data points. For determination of the best setting for the penalty parameter C, values for $C = 10^x$, x = -3.0, -2.5, -2.25, . . ., 0 were tried. Values of the parameter C larger than 1 were not tested extensively, as we found that they resulted in models with similar accuracies. This is in agreement with the Liblinear tutorial in the appendix of [44] which states that once the parameter C exceeds a certain value, the obtained models have a similar accuracy. The SVM with the penalty parameter setting yielding the best assignment accuracy was used to predict the class membership of the left out data point. The class membership predictions for all data points were used to determine the assignment accuracy of the classifier, based on their agreement with the correct assignments. For this purpose, the result of each leave-one-out experiment was classified as either a true positive (TP - correctly predicted lignocellulose degraders), true negative (TN - correctly predicted non-degraders), false positive (FP - non-degraders predicted to be degraders) or a false negative assignment (FN - degraders predicted to be non-degraders). The recall of the positive class and the true negative rate of the classifier were calculated according to the following equations:

$$\text{Re}call = \frac{TP}{TP + FN} \quad (2)$$

$$True\ negative\ rate = \frac{TN}{TN + FP} \quad (3)$$

The average of the recall and the true negative rate, the macro-accuracy, was used as the assignment accuracy to assess the overall performance:

$$MACC = \frac{\text{Re}call + True\ negative\ rate}{2} \quad (4)$$

Subsequently, we identified the settings for the penalty parameter C with the best macro-accuracy by leave-one-out cross-validation. The parameter settings resulting in the most accurate models were used to each train a separate model on the entire data set. Prediction of the five best models were combined to form a voting committee and used for the classification of novel sequence samples such as the partial genome reconstructions from the cow rumen metagenome of switch-grass adherent microbes (see Additional file 2: Table S2 for an evaluation and meta-parameter settings of these ensembles of classifiers).

Feature selection

An SVM model can be represented by a sparse weight vector $\vec{w}$. The positive and negative components of $\vec{w}$, the 'feature weights', specify the relative importance of the protein domains or CAZy families for discrimination between plant biomass-degrading and non-plant

biomass-degrading microorganisms. To determine the most distinctive features for the positive class (that is, the lignocellulose degraders), we selected all features that received a positive weight in weight vectors of the majority of the five most accurate models. This ensemble of models was also used for classification of the cow rumen draft genomes of uncultured microbes (see Classification with a SVM).

Additional files

Additional file 1: Table S1. Isolate strains and metagenome samples used in this study. The signs "+" and "-" indicate availability of CAZy or Pfam annotation data. The symbol * marks strains for which we provide another reference than the genome publication characterizing the metabolic capacities of the respective strain.

Additional file 2: Table S2: Evaluation and meta-parameter settings of the ensembles of classifiers. The ensembles were used for feature selection and phenotype classification of the (draft) genomes and metagenomes. The macro-accuracy for each model for a discrete set of values for the parameter *C* was calculated in cross-validation experiments. The five best models were selected based on macro-accuracy. The mean of the exponentially transformed parameter *C* and the mean macro-accuracy for these five models are shown for all trained classifiers. For details on the different ensemble classifiers, see the Results section in the manuscript.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
AW, YT, PBP and ACM designed the study, interpreted the results and wrote the manuscript. AW and YT conducted the experiments under the supervision of ACM. SGAK and AW computed the CAZy family and protein domain annotations. All authors read and approved the final manuscript.

Acknowledgements
YT, AW and ACM were supported by the Max Planck society and Heinrich Heine University Düsseldorf. PBP gratefully acknowledges support from the Research Council of Norway and the Bilateralt Forskningssamarbeid - Prosjektetablering (BILAT) program. The authors are grateful to Angela Rennwanz who helped downloading the articles for the microbial genomes used in our analysis.

Author details
[1]Max-Planck Research Group for Computational Genomics and Epidemiology, Max-Planck Institute for Informatics, University Campus E1 4, Saarbrücken 66123, Germany. [2]Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Post Office Box 5003, Ås 1432, Norway. [3]Department of Algorithmic Bioinformatics, Heinrich Heine University Düsseldorf, Düsseldorf 40225, Germany.

Received: 16 August 2012 Accepted: 12 February 2013
Published: 15 February 2013

References
1. Rubin EM: **Genomics of cellulosic biofuels.** *Nature* 2008, **454:**841–845.
2. Kaylen M, Van Dyne DL, Choi YS, Blasé M: **Economic feasibility of producing ethanol from lignocellulosic feedstocks.** *Biores Technol* 2000, **72:**19–32.
3. Lee J: **Biological conversion of lignocellulosic biomass to ethanol.** *J Biotechnol* 1997, **56:**1–24.
4. Wheals AE, Basso LC, Alves DMG, Amorim HV: **Fuel ethanol after 25 years.** *TIBTECH* 1999, **17:**482–487.
5. Mitchell WJ: **Physiology of carbohydrate to solvent conversion by clostridia.** *Adv Microb Physiol* 1998, **39:**31–130.
6. Himmel ME, Ding SY, Johnson DK, Adney WS, Nimlos MR, Brady JW, Foust TD: **Biomass recalcitrance: engineering plants and enzymes for biofuels production.** *Science* 2007, **315:**804–807.
7. Xie G, Bruce DC, Challacombe JF, Chertkov O, Detter JC, Gilna P, Han CS, Lucas S, Misra M, Myers GL, *et al*: **Genome sequence of the cellulolytic gliding bacterium cytophaga hutchinsonii.** *Appl Environ Microbiol* 2007, **73:**3536–3546.
8. Brumm P, Mead D, Boyum J, Drinkwater C, Gowda K, Stevenson D, Weimer P: **Functional annotation of fibrobacter succinogenes S85 carbohydrate active enzymes.** *Appl Biochem Biotechnol* 2010, doi:10.1007/s12010-010-9070-5.
9. Morrison M, Pope PB, Denman SE, McSweeney CS: **Plant biomass degradation by gut microbiomes: more of the same or something new?** *Curr Opin Biotech* 2009, **20:**358–363.
10. Brumm P, Hermanson S, Hochstein B, Boyum J, Hermersmann N, Gowda K, Mead D: **Mining *Dictyoglomus turgidum* for enzymatically active carbohydrases.** *Appl Biochem Biotechnol* 2010, doi:10.1007/s12010-010-9029-6.
11. Pope PB, Denman SE, Jones M, Tringe SG, Barry K, Malfatti SA, McHardy AC, Cheng J-F, Hugenholtz P, McSweeney CS, Morrison M: **Adaptation to herbivory by the Tammar wallaby includes bacterial and glycoside hydrolase profiles different to other herbivores.** *Proc Natl Acad Sci USA* 2010, **107:**14793–14798.
12. Warnecke F, Luginbuhl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, Cayouette M, McHardy AC, Djordjevic G, Aboushadi N, *et al*: **Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite.** *Nature* 2007, **450:**560–565.
13. Brulc JM, Antonopoulos DA, Berg Miller ME, Wilson MK, Yannarell AC, Dinsdale EA, Edwards RE, Frank ED, Emerson JB, Wacklin P, *et al*: **Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases.** *Proc Natl Acad Sci USA* 1948, **2009:**106.
14. Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T, *et al*: **Metagenomic discovery of biomass-degrading genes and genomes from cow rumen.** *Science* 2011, **331:**463–467.
15. Pope PB, Mackenzie AK, Gregor I, Smith W, Sundset MA, McHardy AC, Morrison M, Eijsink VGH: **Metagenomics of the svalbard reindeer rumen microbiome reveals abundance of polysaccharide utilization loci.** *PLoS One* 2012, doi:10.1371/journal.pone.0038571.
16. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, *et al*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2012, **40:**D13–D25.
17. Beerenwinkel N, Dumer M, Oette M, Korn K, Hoffmann D, Kaiser R, Lengauer T, Selbig J, Walter H: **Geno2Pheno: estimating phenotypic drug resistance from HIV-1 genotypes.** *Nucleic Acids Res* 2003, **31:**3850–3855.
18. Yosef N, Gramm J, Wang Q-F, Noble WS, Karp RM, Sharan R: **Prediction of phenotype information from genotype data.** *Commun Inf Syst* 2010, **10:**99–114.
19. Someya S, Kakuta M, Morita M, Sumikoshi K, Cao W, Ge Z, Hirose O, Nakamura S, Terada T, Shimizu K: **Prediction of carbohydrate-binding proteins from sequences using support vector machines.** *Adv Bioinformatics* 2010, doi:10.1155/2010/289301.
20. Cortes C, Vapnik V: **Support-vector networks.** *Mach Learn* 1995, **20:**273–297.
21. Boser B, Guyon I, Vapnik V: **A training algorithm for optimal margin classifiers.** In *Fifth Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. Pittsburgh: ACM; 1992:144–152.
22. Chertkov O, Sikorski J, Nolan M, Lapidus A, Lucas S, Del Rio TG, Tice H, Cheng J-F, Goodwin L, Pitluck S, *et al*: **Complete genome sequence of Thermomonospora curvata type strain (B9).** *Stand Genomic Sci* 2011, **4:**13–22.
23. Anderson I, Abt B, Lykidis A, Klenk HP, Kyrpides N, Ivanova N: **Genomics of aerobic cellulose utilization systems in actinobacteria.** *PLoS One* 2012, **7:**e39331.
24. Aspeborg H, Coutinho PM, Wang Y, Brumer H 3rd, Henrissat B: **Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5).** *BMC Evol Biol* 2012, **12:**186.
25. Boraston AB, Bolam DN, Gilbert HJ, Davies GJ: **Carbohydrate-binding modules: fine-tuning polysaccharide recognition.** *Biochem J* 2004, **15:**769–781.
26. Suen G, Weimer PJ, Stevenson DM, Aylward FO, Boyum J, Deneke J, Drinkwater C, Ivanova NN, Mikhailova N, Chertkov O, *et al*: **The complete genome sequence of fibrobacter succinogenes S85 reveals a cellulolytic and metabolic specialist.** *PLoS One* 2011, **6:**e18814.

27. Schultz J, Copley RR, Doerks T, Ponting CP, Bork P: **SMART: a web-based tool for the study of genetically mobile domains.** *Nucleic Acids Res* 2000, **28:**231–234.
28. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, *et al*: **The Pfam protein families database.** *Nucleic Acids Res* 2012, **40:**D290–D301.
29. Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, White O: **TIGRFAMs: a protein family resource for the functional identification of proteins.** *Nucleic Acids Res* 2001, **29:**41–43.
30. Wilson DB: **Three microbial strategies for plant cell wall degradation.** *Ann N Y Acad Sci* 2008, **1125:**289–297.
31. Olson DG, Tripathi SA, Giannone RJ, Lo J, Caiazza NC, Hogsett DA, Hettich RL, Guss AM, Dubrovsky G, Lynd LR: **Deletion of the Cel48S cellulase from Clostridium thermocellum.** *Proc Natl Acad Sci USA* 2010, doi:10.1073/pnas.1003584107.
32. DeBoy RT, Mongodin EF, Fouts DE, Tailford LE, Khouri H, Emerson JB, Mohamoud Y, Watkins K, Henrissat B, Gilbert HJ, Nelson KE: **Insights into plant cell wall degradation from the genome sequence of the soil bacterium Cellvibrio japonicus.** *J Bacteriol* 2008, **190:**5455–5463.
33. Taylor LE, Henrissat B, Coutinho PM, Ekborg NA, Hutcheson SW, Weiner RM: **Complete cellulase system in the marine bacterium Saccharophagus degradans strain 2-40 T.** *J Bacteriol* 2006, **188:**3849–3861.
34. Hervé C, Rogowski A, Blake AW, Marcus SE, Gilbert HJ, Knox JP: **Carbohydrate-binding modules promote the enzymatic deconstruction of intact plant cell walls by targeting and proximity effects.** *Proc Natl Acad Sci USA* 2010, **107:**15293–15298.
35. Duan CJ, Feng JX: **Mining metagenomes for novel cellulase genes.** *Biotechnol Lett* 2010, **32:**1765–1775.
36. Wilson DB: **Evidence for a novel mechanism of microbial cellulose degradation.** *Cellulose* 2009, **16:**723–727.
37. Lynd LR, Weimer PJ, van Zyl WH, Pretorius IS: **Microbial cellulose utilization: fundamentals and biotechnology.** *Microbiol Mol Biol Rev* 2002, **66:**506–577.
38. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B: **The carbohydrate-active EnZymes database (CAZy): an expert resource for glycogenomics.** *Nucleic Acids Res* 2009, **37:**D233–D238.
39. Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, Grechkin Y, Ratner A, Jacob B, Pati A, Huntemann M, *et al*: **IMG/M: the integrated metagenome data management and comparative analysis system.** *Nucleic Acids Res* 2012, **40:**D123–D129.
40. Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, *et al*: **IMG: the integrated microbial genomes database and comparative analysis system.** *Nucleic Acids Res* 2012, **40:**D115–D122.
41. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching.** *Nucleic Acids Res* 2011, **39:**W29–W37.
42. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y: **dbCAN: a web resource for automated carbohydrate-active enzyme annotation.** *Nucleic Acids Res* 2012, doi:10.1093/nar/gks479.
43. Yaun G-X, Chang K-W, Hsieh C-J, Lin C-J: **A comparison of optimization methods for large-scale L1-regularized linear classification.** *J Mach Learn Res* 2010, **11:**3183–3234.
44. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ: **LIBLINEAR: a library for large linear classification.** *J Mach Learn Res* 2008, **9:**1871–1874.
45. Ruschhaupt M, Huber W, Poustka A, Mansmann U: **A compendium to ensure computational reproducibility in high-dimensional classification tasks.** *Stat Appl Genet Mol Biol* 2004, **3:**Article 37.

doi:10.1186/1754-6834-6-24
Cite this article as: Weimann *et al.*: De novo prediction of the genomic components and capabilities for microbial plant biomass degradation from (meta-)genomes. *Biotechnology for Biofuels* 2013 **6**:24.

8.4 Publication - Konietzny *et al.* 2014

Status	**published**
Journal	Biotechnology for Biofuels (Impact factor in year of submission: 6.221)
Citation	Sebastian GA Konietzny, Phillip B Pope, Aaron Weimann, Alice C McHardy: Inference of phenotype-defining functional modules of protein families for microbial plant biomass degraders. *Biotechnology for Biofuels 2014*, **7**:124
URL	http://www.biotechnologyforbiofuels.com/content/7/1/124
Own contribution	75% Designed and performed the experiments (with co-authors) Analyzed the data (with co-authors) Wrote the manuscript (with co-authors)

Konietzny et al. Biotechnology for Biofuels 2014, 7:124
http://www.biotechnologyforbiofuels.com/content/7/1/124

Biotechnology for Biofuels

RESEARCH ARTICLE Open Access

Inference of phenotype-defining functional modules of protein families for microbial plant biomass degraders

Sebastian GA Konietzny[1,3], Phillip B Pope[2], Aaron Weimann[3] and Alice C McHardy[1,3*]

Abstract

Background: Efficient industrial processes for converting plant lignocellulosic materials into biofuels are a key to global efforts to come up with alternative energy sources to fossil fuels. Novel cellulolytic enzymes have been discovered in microbial genomes and metagenomes of microbial communities. However, the identification of relevant genes without known homologs, and the elucidation of the lignocellulolytic pathways and protein complexes for different microorganisms remain challenging.

Results: We describe a new computational method for the targeted discovery of functional modules of plant biomass-degrading protein families, based on their co-occurrence patterns across genomes and metagenome datasets, and the strength of association of these modules with the genomes of known degraders. From approximately 6.4 million family annotations for 2,884 microbial genomes, and 332 taxonomic bins from 18 metagenomes, we identified 5 functional modules that are distinctive for plant biomass degraders, which we term "plant biomass degradation modules" (PDMs). These modules incorporate protein families involved in the degradation of cellulose, hemicelluloses, and pectins, structural components of the cellulosome, and additional families with potential functions in plant biomass degradation. The PDMs were linked to 81 gene clusters in genomes of known lignocellulose degraders, including previously described clusters of lignocellulolytic genes. On average, 70% of the families of each PDM were found to map to gene clusters in known degraders, which served as an additional confirmation of their functional relationships. The presence of a PDM in a genome or taxonomic metagenome bin furthermore allowed us to accurately predict the ability of any particular organism to degrade plant biomass. For 15 draft genomes of a cow rumen metagenome, we used cross-referencing to confirmed cellulolytic enzymes to validate that the PDMs identified plant biomass degraders within a complex microbial community.

Conclusions: Functional modules of protein families that are involved in different aspects of plant cell wall degradation can be inferred from co-occurrence patterns across (meta-)genomes with a probabilistic topic model. PDMs represent a new resource of protein families and candidate genes implicated in microbial plant biomass degradation. They can also be used to predict the plant biomass degradation ability for a genome or taxonomic bin. The method is also suitable for characterizing other microbial phenotypes.

Keywords: Latent Dirichlet allocation, LDA, Probabilistic topic models, (Ligno)cellulose degradation, Plant biomass degradation, Phenotype-based identification of functional modules, Pectin degradation, Feature ranking, Polysaccharide utilization loci, Gene clusters

* Correspondence: mchardy@hhu.de
[1]Max-Planck Research Group for Computational Genomics and Epidemiology, Max-Planck Institute for Informatics, University Campus E1 4, Saarbrücken 66123, Germany
[3]Department of Algorithmic Bioinformatics, Heinrich Heine University Düsseldorf, Düsseldorf 40225, Germany
Full list of author information is available at the end of the article

Background

Lignocellulose is an integral part of plant cell walls. Its high energy content and renewability make it a promising alternative energy resource, particularly for the production of biofuels [1,2]. However, the current industrial methods of degrading recalcitrant plant cell wall material remain inefficient [3], which has created great interest in lignocellulolytic microbial organisms [4], because these represent a promising source of potential enzymes for improving industrial degradation processes [4,5]. Plant cell walls consist of cellulose and hemicelluloses (for example, xylan, xyloglucan, β-glucan), which are cross-linked by lignin, and pectins [6,7]. Cellulose is a macromolecule of β-(1,4)-linked D-glucose molecules. Xylans and β-glucans are homopolysaccharides composed of either xylose or β-1,3, β-1,4-linked D-glucose, respectively, and are commonly found in plant cell walls of grasses. Xyloglucan is a hemicellulose occurring in the plant cell wall of flowering plants, and consists of a glucose homopolysaccharide backbone with xylose side chains, which are occasionally linked to galactose and fucose residues. Pectin is a heteropolysaccharide that represents a major component of the middle lamella of plant cell walls, while lignin is a strongly cross-linked polymer of various aromatic compounds. Degradation of plant material requires the concerted action of different carbohydrate-binding modules (CBMs) and catalytic enzymes, such as cellulases, xylanases, pectin lyases and peroxidases [8-10]. The CAZy database [11] distinguishes four important subclasses of carbohydrate-active enzymes (CAZymes): glycoside hydrolases (GHs), glycosyltransferases (GTs), polysaccharide lyases (PLs), and carbohydrate esterases (CEs). However, cellulolytic enzymes can also be multifunctional, and combine several CAZy families in a modular architecture [12].

Microorganisms use different strategies to degrade recalcitrant plant material. The free enzyme and the cellulosome strategies are the strategies most widely used by known microbial plant biomass degraders [12,13]. The free enzyme paradigm is frequently used by aerobic bacteria, and involves the secretion of cellulolytic enzymes to degrade lignocellulose in the external medium. The cellulosome strategy has so far been described only for anaerobic bacteria [13], and involves large protein complexes (the cellulosomes) that incorporate cellulolytic enzymes, as well as CBMs for localized lignocellulose degradation [14]. The cellulosome includes a scaffoldin backbone to which cellulases and hemicellulases attach via cohesin–dockerin interactions. The corresponding (hemi)cellulases contain the dockerin domains, one or more catalytic domains (for example, GHs), and non-catalytic CBMs [14]. More recently, two additional strategies for (hemi)cellulose degradation have been outlined. The first strategy is the Sus-like protein system, which relies on mechanisms that are similar to the starch utilization (Sus) system in *Bacteroides thetaiotaomicron* [15,16]. These mechanisms are mediated by enzymes located in the outer membrane [17]. The second strategy involves the oxidative cleavage of cellulose by copper mono-oxygenases, a mechanism that increases the efficiency of the hydrolytic enzymes [18].

Certain cellulolytic organisms, such as *Fibrobacter succinogenes* and *Cytophaga hutchinsonii*, do not seem to use any of the currently known mechanisms [13]. Additional insights into microbial degradation processes have been generated by studies of microbial communities using metagenomics. This has led to the identification of thousands of putative carbohydrate-active genes [19,20] and of several novel genes encoding proteins with cellulolytic activities from uncultured organisms [21-23]. Overall, more than 1,000 cellulase genes have been discovered by genomic and functional screens [24]; however, important details about their degradation mechanisms remain unresolved [13,25]. Therefore, the discovery of novel protein families that are involved in plant biomass degradation is an ongoing effort.

The CAZymes Analysis Toolkit (CAT) can be used to recognize carbohydrate-active enzymes [26]. CAT deduces its prediction rules from the frequencies of modular proteins with Pfam and CAZy assignments in the CAZy database, thus its application to newly emerging sequences from metagenomes is likely to be limited because it is restricted to protein families that already have correspondences in the database. An alternative approach for determining the protein families that participate in a particular process but have no homologs with known activities is to use computational methods that assign families to a functional context. Depending on the granularity of the context, this approach allows narrowing down of the set of possible functions for an uncharacterized protein family. Applied to thousands of families on a large scale, this allows the *de novo* discovery of phenotype-defining protein families, genes, or entire functional modules [27].

Several methods for ranking genes or pathways by their assumed relevance for a certain phenotype have been described [28-37]. These methods measure the association of individual protein families [28], known pathways [29] or single nucleotide polymorphisms [30] with the presence or absence of phenotypes across a set of genomes. In some instances, the search space is limited to proteins in predicted operon structures [31] or to pairs of functionally coupled proteins [32].

We have previously described a family-centric method for the identification of protein families involved in lignocellulose degradation [28]. This method uses an ensemble of linear L1-regularized support vector machine (SVM) classifiers trained with the genome annotations of known lignocellulose-degrading and non-lignocellulose-degrading species. Similar methods use ranking approaches

that are followed by a clustering step, whereby phenotype-associated families are grouped into modules based on their co-occurrence patterns across organisms, which are likely to indicate functional dependencies [33,34]. However, we suggest that the order of steps should be reversed, that is, functional dependencies between families should be detected first. This is because family-centric ranking methods may fail to detect moonlighting proteins [35] that are active in multiple processes, which could reduce the global correlation of their absence/presence profiles with the ability of the organisms to perform the target process.

By contrast, pathway-centric methods search for sets of functionally coupled protein families related to a specific phenotype. These methods use prior information about pathways from, for example, the Kyoto Encyclopedia of Genes and Genomes (KEGG) [36] or BioPath [29] databases in the form of organism-specific enzyme reaction networks based on enzyme classification (EC) numbers. The Network Instance-Based Biased Subgraph Search (NIBBS) searches for phenotype-associated edges in order to identify phenotype-related enzyme reactions in a KEGG-based network [37]. Similarly, MetaPath identifies subgraphs of a KEGG-derived network by assessing the statistical support of phenotype associations for every edge [36]. To date, there has been no application of pathway-centric methods to the study of lignocellulose degradation. Moreover, because of their focus on well-defined reaction networks, these methods have limitations for the analysis of metagenome samples, which often allow only partial metabolic reconstructions. Furthermore, species from newly sequenced microbial communities are likely to have a metabolism that is distinct from the metabolisms of well-studied model species, and the latter have been the basis for most of the currently described reaction networks. We are not aware of a pathway-centric method for inferring phenotype-associated functional modules that is applicable to metagenomes, and does not require prior knowledge about the underlying enzyme reaction networks or the target pathways. However, such a method would represent an important addition to computational metagenome analysis methods [38]. A possible solution could be the use of the aforementioned family-centric methods that cluster families into modules after determining their associations with the phenotype of interest.

An indication of the functional context for a protein family can be obtained by clustering families by their co-occurrences across genomes [39,40]. We have previously used latent Dirichlet allocation (LDA) [41], a Bayesian method, to infer 200 potential functional modules from 575 prokaryotic genomes [42]. The modules represent sets of co-occurring protein families that are likely to be involved in a common biological process, and cover a broad range of biochemical activities, including several known protein complexes, metabolic pathways, and parts of signal transduction processes. Overall, the modules show significant functional coherence, as indicated by a comparison with high-confidence protein–protein interactions from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database [43]. Here, we describe a new method based on LDA for determining the functional modules associated with microbial plant biomass degradation. This method detects relevant functional modules by the strength of their associations with the plant biomass degradation phenotype.

In the current study, we processed a large dataset of nearly 3,000 sequenced bacterial and archaeal genomes and taxonomic bins of 18 metagenomes. Based on the abundance estimates reported by Medie *et al*. [44] and Berlemont *et al*. [45], the relative abundance of species possessing plant biomass degradation capabilities within the sequenced genomes could exceed 20 to 25%; however, to date, only a small set of species have been confirmed to possess such capabilities [4]. With our method, genomes of both known and unknown degraders could be included in the inference process, and be used to identify distinct sets of protein families that are specific for microbial plant biomass degraders. The use of metagenome data allows us to incorporate information from environmental communities into the inference process. We identified five functional modules for plant biomass degradation, which we call “plant biomass degradation modules” (PDMs). The PDMs found included many protein families that are known to be involved in plant biomass degradation, and a substantial number of families that have not previously been linked to microbial plant biomass degradation. To verify the relevance of these newly identified PDMs and candidate families, we searched for gene clusters including the families of the PDMs. Several of the identified clusters are known to be active in the degradation of lignocellulose. Furthermore, the PDMs had a predictive value for identifying plant biomass degraders from the genomes of sequenced isolates or of plant biomass-degrading microbial communities.

Results and discussion

We generated approximately 6.4 million protein annotations with Pfam and CAZy families for 2,884 bacterial and archaeal genomes from the Integrated Microbial Genomes database (IMG) and 332 taxonomic bins from 18 metagenomes (see Methods). We then used a two-step approach to identify functional modules that are distinctive for microbial lignocellulose degraders. First, the set of protein family annotations was processed with LDA, and 400 potential functional modules were inferred, with each corresponding to a set of Pfam and/or

CAZy families (Figure 1, steps 1 and 2). The modules were learned in an unsupervised fashion without consideration of the phenotypes of the organisms, as described previously [42]. In the second step, we ranked the 400 functional modules according to their strength of association with the genomes of plant biomass degraders across a subset of the genomes consisting of 38 known lignocellulose degraders and 82 non-degraders (Figure 1, step 3). For this, we defined genome-specific module weights, which corresponded to the fraction of the protein families of a module that were annotated for a certain genome or taxonomic bin (completeness scores). Functional modules were considered to be present in a genome or bin if their completeness score reached a certain threshold. For each module, we determined the best setting for this threshold, corresponding to the one that

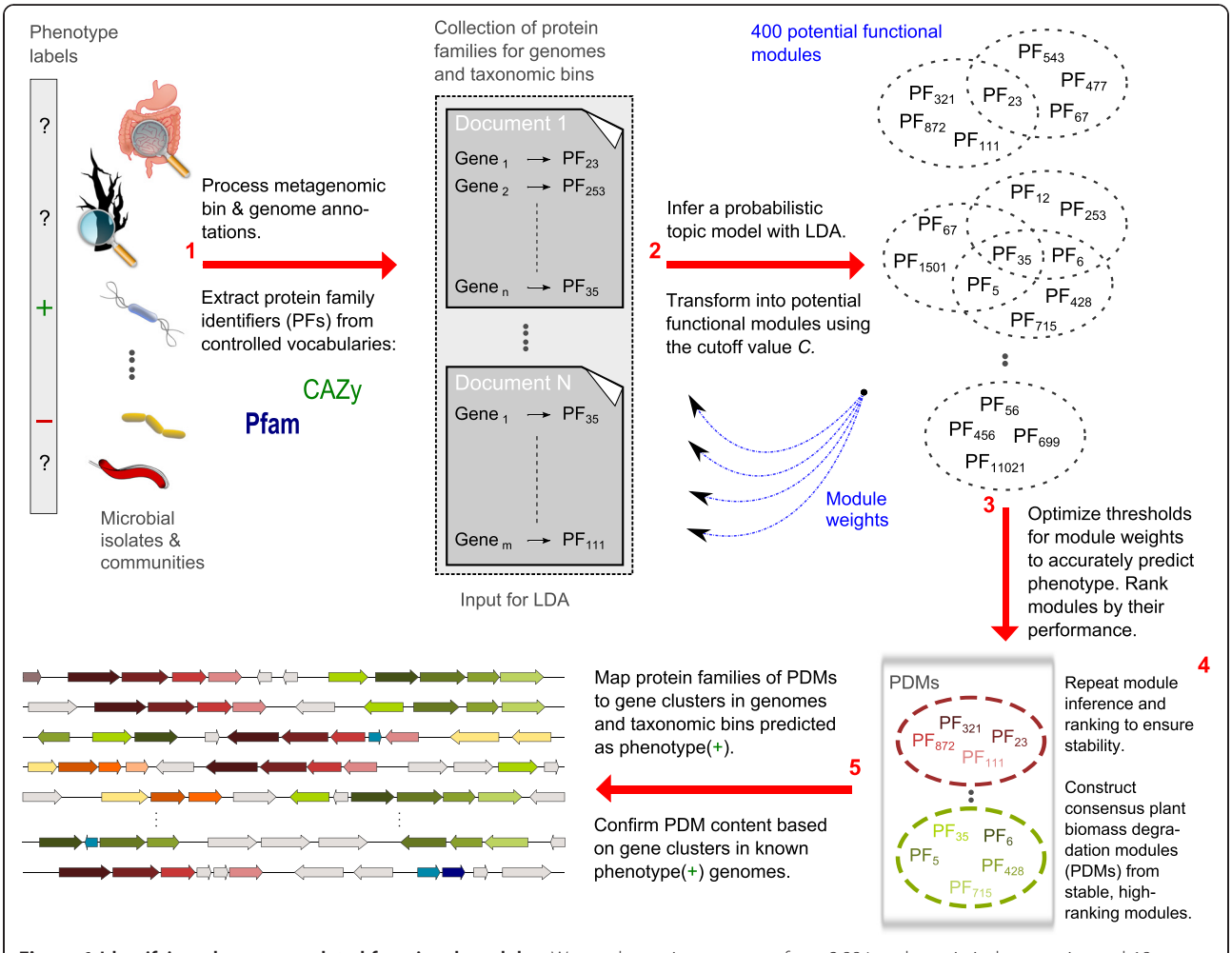


Figure 1 Identifying phenotype-related functional modules. We used protein sequences from 2,884 prokaryotic isolate species and 18 microbial communities, some of which are known to be involved in lignocellulose degradation. Known lignocellulose degradation abilities are indicated by phenotype labels (positive/negative: +/–). For the metagenomes, we considered only protein-coding sequences with predicted taxonomic origins assigned by a taxonomic binning method (*PhyloPythia* or *PhyloPythiaS*). We used HMMER to assign protein family annotations from Pfam and CAZy to all input sequences, and summarized the set of (meta-)genome annotations as a document collection for LDA **(1)**. Each document represented a single genome or a metagenome bin, and was composed of protein family identifiers from a controlled vocabulary (Pfam, CAZy). We then inferred a probabilistic topic model **(2)**. The topic variables of the model can be interpreted as potential functional modules, that is, sets of functionally coupled protein families [42]. We obtained 400 modules with diverse biochemical functions. Next, we defined genome-specific weights of the modules, and used these weights in conjunction with the phenotype labels to rank the modules according to their estimated relevance for the phenotype of lignocellulose degradation **(3)**. As weights, we used the fraction of protein families in a module that were present in a certain genome or metagenome bin (completeness scores). We identified stable, high-ranking modules from independent repetitions of the analysis, and constructed consensus modules, which we named "plant biomass degradation modules" (PDMs) **(4)**. These PDMs were found to cover different aspects of plant biomass degradation, such as degradation of cellulose, hemicellulose, and pectin. Moreover, the weights of the PDMs could be used to predict the biomass degradation abilities of organisms, and we were able to identify specific gene clusters in the input set of (meta-)genomes that reflected the protein family content of individual modules **(5)**. The clusters thus provided evidence for the functional coherence of the modules by gene neighborhood.

optimally separated the genomes of degraders and non-degraders according to the F-measure (the weighted harmonic mean of precision and recall (see Methods)). The modules with the largest F-scores were strongly associated with the genomes of lignocellulose degraders, as indicated by an average F-score of 87.45% for the top 10 modules.

Identification of stable PDMs

The implementation of the LDA model that we used was based on Gibbs sampling, a Markov chain Monte Carlo (MCMC) method that efficiently estimates parameters for complex models such as LDA. In agreement with the recommended procedures for MCMC sampling [46], we repeated the analysis multiple times (18 LDA runs) to ensure the stability of the results. We thus repeated the two central steps of our method, that is, the inference of modules and their subsequent ranking by phenotype association (Figure 1, steps 2 and 3), 18 times to identify stable, high-ranking modules. We summarized the information from stable, high-ranking modules found in different runs by constructing consensus modules that contained all the protein families found in similar modules in at least nine LDA runs (Figure 1, step 4; see Methods).

We identified five consensus modules (M1 to M5), which we referred to as PDMs (Table 1; see Additional file 1: Tables S1A-5A). We mapped the CAZy families of these PDMs to essential activities in the degradation of plant cell wall material, based on their EC numbers (Table 2). All PDMs included protein families with cellulase or hemicellulase activities, which supports the relevance of these modules for plant biomass degradation. M1 to M5 were functionally distinct, with only a moderate overlap (12.6%) of their protein family content, including the broadly defined families GH5 and GH43 [47]. Isofunctional Pfam and CAZy terms, such as PF00150 and GH5, were grouped together into the same PDMs in most cases.

Table 1 Functional characterization of the consensus plant biomass degradation modules M1 to M5

Module[a,b]	Description
M1	Lignocellulose degradation (cellulose and hemicellulose degradation)
M2	Xylan binding and xyloglucan degradation (hemicellulose degradation)
M3	Pectin degradation
M4	Degradation of glycan compounds
M5	Structural components of the cellulosome-based degradation paradigm (dockerin and cohesin)

[a]We characterized each module based on the set of protein families contained within it.
[b]Additional file 1 shows each consensus module as a list of Pfam/CAZy terms (Tables S1A-5A).

The modules also included 20 Pfam families without a commonly known link to plant biomass degradation, such as domains of unknown function (DUF), ricin-type β-trefoil lectin-like domains, and GDSL-like lipase/acylhydrolase (Table 3; see Additional file 2: Section 1). Some of these domains could encode novel functions that are important for plant biomass degradation.

Gene clusters with PDM protein families

To confirm a functional context for the protein families assigned to the same PDM, we searched for gene clusters annotated with multiple families of a PDM in the 38 genomes of known degraders, as the proximity of genes within a genome indicates a shared functional context [50,51] (Figure 1, step 5). For each PDM, we identified gene clusters of four or more neighboring genes, with intergenic distances of 2 kb or less between consecutive genes. Overall, 81 gene clusters were found for the 5 PDMs, which represented 51 distinct, non-overlapping clusters. The average distance between the genes of these clusters was 340 bp. On average, 70.7% of the family content of each PDM could be mapped to gene clusters in known degraders. Some of the gene clusters discovered have been described previously as being active in lignocellulose degradation (see following sections), whereas the novel gene clusters are candidates for further experimental investigation. Notably, more than half (55%) of the possibly relevant protein families in Table 3 appeared in at least one gene cluster identified in known degrading species, which supports their potential role in the degradation process.

Assessment of the potential of the PDMs to predict unknown lignocellulose degraders

The completeness of a PDM in a genome was predictive for the ability of an organism to degrade lignocellulosic plant biomass. We determined the predictive value for each PDM by standard evaluation protocols in leave-one-out (LOO) and 10-fold cross-validation experiments (see Methods). In these experiments, genomes from the learning set of 120 known lignocellulose degraders and non-degraders were successively left out of the process of determining the completeness threshold. Subsequently, PDMs were predicted to be present in the omitted genomes if their completeness score for the genome was greater than or equal to the inferred threshold. This procedure was used to assess the generalization error of a PDM-based classifier to avoid overly optimistic performance estimates [52,53]. We observed high F-scores for the PDMs in the LOO setting (82.1 to 96.2%) and lower bounds for the cross-validation estimates of prediction accuracy between 76.57% and 91.69% (Table 4).

Table 2 CAZymes with key functions for plant cell wall degradation in the plant biomass degradation modules

Type	Subtype	Module				
		M1	M2	M3	M4	M5
Cellulases [8]	Endoglucanase (EC 3.2.1.4)	GH5, GH9	–	GH5	GH5	GH124
	Cellobiohydrolase (EC 3.2.1.91)	GH5, GH9	–	GH5	GH5	–
	β-glucosidase (EC 3.2.1.21)	GH9	GH30	–	GH3	–
Hemicellulases [8,9]	Endo-1,4-β-xylanase (1,4-β-d-xylan xylanohydrolases, EC 3.2.1.8)	GH5, GH10, GH43	–	GH5, GH43	GH5, GH43	–
	β-xylosidase (1,4-β-d-xylan xylohydrolase, EC 3.2.1.37)	GH43	GH30	GH43	GH3, GH43	–
	α-arabinofuranosidase (EC 3.2.1.55)	GH43	–	GH43	GH3, GH43	–
	α-glucuronidase (EC 3.2.1.139)	–	–	–	–	–
	Acetyl xylan esterase (EC 3.1.1.72)	–	–	CE6, CE7, CE12	–	–
	Ferulic acid esterase (EC 3.1.1.73)	–	–	–	–	–
	Xyloglucanase (EC 3.2.1.151); xyloglucosyltransferase (EC 2.4.1.207)	–	GH16	–	–	–
Carbohydrate-binding modules [48,49]	Targeting cellulose	CBM4	–	–	–	CBM3
	Targeting xylan	CBM4, CBM6, CBM35	CBM6, CBM13, CBM35			CBM36
Cellulosomes [14]	Structural components	–	–	–	–	Cohesin, dockerin
Pectinolytic enzymes [10]	Pectin methyl esterase (EC 3.1.1.11)	–	–	CE8	–	–
	Endopolygalacturonase (EC 3.2.1.15); exopolygalacturonase (EC 3.2.1.67)	–	–	GH28	–	–
	Endopolygalacturonase lyase (EC 4.2.2.2); exopolygalacturonase lyase (EC 4.2.2.9)	–	–	PL1, PL9	–	–
	Pectin lyase (EC 4.2.2.10)	–	–	PL1	–	–

The top-ranking PDMs, M1 and M2, predicted the ability to degrade lignocellulose with cross-validation accuracies of more than 93%. Four genomes were misclassified by both M1 and M2 (see Additional file 4: Figure S1; see Additional file 5: Tables S1 and S2): *Bryantella formatexigens* (false negative (FN)), *Xylanimonas cellulosilytica* (FN), *Thermonospora curvata* 43183 (FN), and *Actinosynnema mirum* (false-positive (FP)). Interestingly, *A. mirum* and *T. curvata* might have been mischaracterized previously [55], supporting the predictions by the two PDMs (see Additional file 2: Section 2). All PDMs showed a precision of more than 82% for lignocellulose degraders, with few occurrences predicted for non-degraders. M3 and M5 were found only in a subset of the known degraders, with the lowest recall being 57.9% (Table 4), suggesting that these modules might represent specific aspects of degradation strategies. However, looking at the presence/absence profiles of the PDMs across the degrading species, none of the PDMs showed an exclusive association with a known degradation paradigm (Figure 2).

Protein families of the PDMs

The highest-scoring PDM M1 (F-measure 96.2%) incorporated various key families for the degradation of cellulose and hemicelluloses (Table 2): GH5, GH9, GH10, GH26, GH43, and CBM6 [47]. The GH5 and GH9 families together cover three classes of important cellulases [8]: endoglucanases, cellobiohydrolases, and β-glucosidases. Both are large families of cellulases that have been studied in many lignocellulolytic organisms (see Additional file 2: Section 3). In addition to their cellulase activities, some members of these families are also hemicellulases with characterized activity on β-glucans, xyloglucans, and heteroxylans [11]. The GH10 and GH43 families include xylanases and arabinases. M1 was present almost exclusively in lignocellulose-degrading bacteria (97.2% precision) and in almost all of them (92.1% recall). Similarly, also the individual modules used for creating the M1 consensus PDM showed strong associations with plant biomass degradation: M1 was always among the three best-ranking modules, and was the top-ranked module in 14 of 18 LDA runs.

Table 3 Protein families of the modules M1 to M5 with potential functions in plant biomass degradation

Module	Family ID	Description
M1	PF13472[a]	GDSL-like lipase/acylhydrolase family
	PF00756[a]	Putative esterase
M2	PF14200[a,b]	Ricin-type β-trefoil lectin domain-like
	PF00652[a,b]	Ricin-type β-trefoil lectin domain
	PF00754[a]	F5/8 type C domain
	PF00041[a]	Fibronectin type III domain
	PF02311	AraC-like ligand-binding domain
	PF13483	Beta-lactamase superfamily domain
M3	PF03629[a,b]	DUF303
	PF00657[a]	GDSL-like lipase/acylhydrolase
	PF13472[a]	GDSL-like lipase/acylhydrolase family
	PF13229[a]	Right-handed beta helix region
M4	PF14310[a]	Fibronectin type III-like domain
	PF07859	Alpha/beta hydrolase fold
	PF00135	Carboxylesterase family
	PF13802	Galactose mutarotase-like
M5	PF13186[a,b]	DUF4008
	PF05593	RHS repeat
	PF07591	Pretoxin Hint domain
	PF07238	PilZ domain
	PF13403	Hint domain

DUF, domain of unknown function; GDSL, a motif in the amino acid sequences of the members of this protein family; PDM, plant biomass degradation module.
[a]Protein families appearing in the gene clusters identified by mapping the PDMs to the phenotype-positive genomes.
[b]Some potential functions of the families PF14200, PF00652, PF03629, and PF13186 are discussed in the context of the respective PDMs (see Results section; see Additional file 2: Section 1).
The table lists protein families of the plant biomass degradation modules (PDMs) that had no commonly known functions in plant biomass degradation. Every second family (55%) occurred in the gene clusters that were identified based on the PDMs. Note that family PF13472 was part of M1 and M3.

M2 (F-measure 94.1%) contained families that bind and degrade xylan, xyloglucan, and β-glucan (Table 2), such as GH30 (β-xylosidases), GH16 (β-glucanases, xyloglucanases) [9], CBM61 (which is often found with GH16), and the fucose-binding module CBM47. In addition, M2 included the xylan-binding domains CBM6, CBM35, and PF02018, which were also present in hemicellulolytic gene clusters with M2 families of *Clostridium cellulolyticum* (Figure 3B) and *F. succinogenes* (see Additional file 6: Figure S1). In *Streptomyces lividans,* several small gene clusters of two or three genes with M2 member families might be linked to a xylan-binding mechanism involving CBM13 (also known as the ricin superfamily or R-type lectins) [56]. CBM13 and two ricin-type β-trefoil lectin domains (PF14200 and PF00652 in Table 3) belonged to M2 and occurred in the clusters. Interestingly, the two different functional aspects of M2 (xyloglucan degradation and xylan binding) were reflected by a split of the M2 module into two modules in some LDA runs.

M3 (F-measure 89.6%) included cellulose-degrading, hemicellulose-degrading, and multiple pectinolytic enzymes, such as pectin methyl esterase (CE8), pectin lyases PL1, PL9, and PF12708 (PL3), and endopolygalacturonase (GH28) (Table 2). M3 also included GH106 (α-L-rhamnosidase), which catalyzes the release of L-rhamnose from pectin (rhamnogalacturonan) molecules, and GH105, an unsaturated rhamnogalacturonyl hydrolase. Moreover, three acetyl xylan esterases (CE6, CE7, and CE12) were assigned to M3, along with the uncharacterized domain PF03629 (DUF303), which may be an acetyl xylan esterase-related enzyme (InterPro accession: IPR005181). As CE12 has both acetyl xylan esterase (EC 3.1.1.72) and pectin acetylesterase (EC 3.1.1.-) activities assigned in CAZy, the other esterase families are possibly also relevant for pectin degradation. Overall, the presence of multiple families involved in cellulose, hemicellulose, and pectin degradation confirmed the relevance of M3 for plant biomass degradation.

Module M4 (F-measure 82.5%) contained the GH5, GH43, GH2, and GH3 families, as well as some associated Pfam domains, such as a GH2 sugar-binding domain (PF02837) and the *C*- and *N*-terminal domains of GH3 (PF01915, PF00933). M4 also included GH35 and GH42, which are both β-galactosidases, and three members of a superfamily of α-galactosidases. D-galactose is an abundant component of the side chains of pectin, heteromannan, and xyloglucan [7]. Activities in the degradation of pectins have been described for several β-galactosidases from plants [58]. Furthermore, M4 seemed to be linked to xyloglucan degradation in *Bacteroides cellulosilyticus* and *Cellvibrio japonicus* (see Additional file 2: Section 4). In conclusion, M4 comprised functionally diverse glycan degradation families, in line with the heterogeneous nature of hemicellulose polysaccharides and their widely varying constituent sugars [7].

M5 (F-measure 82.1%) included structural components of the cellulosome complex (cohesin and dockerin domains), the endoglucanase family GH124, and CBMs targeting cellulose (CBM3) and hemicellulose (CBM36). CBM3 is frequently found as a domain of cellulosome scaffoldin proteins [14]. The S-layer homology domain (PF00395), which anchors cellulosomes to the bacterial cell surface [14], was not associated with M5. It was consistently grouped into modules without significant scores in our rankings, indicating that the S-layer homology domain could perform other functions in non-degraders. M5 included five more Pfam domains of unknown relevance which are interesting candidates for novel functional activities (Table 3). PF13186, a domain of unknown function in our dataset, was annotated for the gene Cthe_3076 in *Clostridium thermocellum,* which lies directly upstream of a gene cluster (Cthe_3077–3080) that is responsible for

Table 4 Association with lignocellulose degradation based on different performance measures for the consensus PDMs M1 to M5

	Module				
	M1	M2	M3	M4	M5
Set of recurring modules (18 repetitions of analyses)					
Number of modules in set	18	18	18	18	16
Average $F_{0.5}$-score in rankings, %	95.2 ± 1.7	92.5 ± 1.1	88.9 ± 2.1	85.8 ± 1.3	84.9 ± 5.3
Average rank	1.3 ± 0.57	2.4 ± 0.61	4.2 ± 1.5	6 ± 1.6	7.5 ± 3.4
Consensus PDM					
Size	18	22	23	25	13
Weight threshold used for classification, %	66.67	50.00	73.91	72.00	38.46
Performance evaluation					
LOO $F_{0.5}$-score, %	96.2	94.1	89.6	82.5	82.1
LOO recall, %	92.1	84.2	63.2	84.2	57.9
LOO precision, %	97.2	97.0	100.0	82.1	91.7
CV accuracy, %	96.7	93.8	87.7	89.6	84.3
Estimated 95% confidence interval for CV accuracy	[91.69, 99.08]	[87.82, 97.35]	[80.42, 92.96]	[82.68, 94.42]	[76.57, 90.32]
CV-MAC, %	95.4	91.3	81.2	88.2	77.2

CV, cross-validation; LOO, leave-one-out; MAC, macro-accuracy; PDM, plant biomass degradation module.
Each consensus PDM represents a set of recurring modules from 18 independent repetitions of our analysis (Figure 1), and contains all families that occurred in at least nine of these modules. The recurring modules used to build the PDMs were identified by finding modules having minimal pairwise distances from each other (see Methods). We reported the average rank and average F-score of these module sets ($F_{0.5}$ puts stronger emphasis on precision; that is, it weights recall as half as strongly as precision [54]; see Additional file 3: Section 3). "Size" gives the number of Pfam and/or CAZy families that are contained in a PDM. We computed recall, precision, and the F-measure scores for the individual PDMs in LOO validation. In addition, accuracies and estimated confidence intervals for 10-fold cross-validation (CV) were used to assess the generalization error more accurately. Following our previous study [28], we also computed the cross-validation macro-accuracy (CV-MAC) as the average of the true-positive (TP) and true-negative (TN) rates.

the structural organization of the cellulosome [59]. However, PF13186 was also annotated for non-degrading genomes (Figure 4), and has been characterized as an iron-sulfur cluster-binding domain in a recently updated version of the Pfam database.

For the complete protein family sets of the consensus PDMs, see Additional file 1 (Tables S1A-5A), which also lists PDM families that were found in fewer than 9 similar modules of the 18 LDA runs, and were thus not included in the consensus PDMs (Tables S1B-5B).

Absence of the cellulase families GH6 and GH48

Interestingly, none of the PDMs contained the cellulase families GH6 or GH48. Both of these play an important role in cellulose degradation in some bacteria, but are not universally found in known lignocellulose degraders. They were not identified in *F. succinogenes*, *C. hutchinsonii*, or several gut and rumen metagenomes with lignocellulose-degrading capabilities [17,20,24,60]. In line with these findings, we found GH6 and GH48 to be annotated in less than 5% of the samples of our input collection, and there was only a single GH6 annotation (no GH48) in the metagenome bins. This rarity in our dataset caused weak co-occurrence signals, and is probably the cause why neither of these families were assigned to the stable, high-ranking modules (see Additional file 2: Section 5; see Additional file 7: Figures S1 and S2 (heat maps visualizing the co-occurrences of GH6 and GH48 with the protein families of M1)).

Despite this, GH48 was among the top 50 protein families of 10 functional modules used to derive the M5 consensus module. This association with M5 is in line with the fact that many bacterial cellulosomes include proteins from the GH48 family [61]. However, the probabilities for GH48 were less than the threshold value $C = 0.01$ that we required for inclusion into modules. This is also evident from a weaker co-occurrence of GH48 with the M5 protein families in lignocellulose degraders (Figure 4). Similar to GH48, most of the other members of the top 50 protein families of the M5 topics co-occurred in the phenotype-positive genomes with cohesin and dockerin annotations (see Additional file 8: Figures S1 and S2 (heat maps)). This could be indicative of functional links with cellulosomes, as for GH48. However, the weak associations of these families with M5 suggest that they are not exclusively related to the cellulosome-based paradigm. Nevertheless, their potential relevance for plant biomass degradation was indicated when we applied a less stringent cutoff to the topic distributions. With $C = 0.005$, the size of the M5 consensus module increased from 13 to 34 protein families, including GH48. Despite the substantial increase in the number of families, the F-score (80%) for the M5 module decreased by only 2%

M1	M2	M3	M4	M5	Cellulosome	Free enzymes	Oxygen requirement	Name	Reference
					1	0	Anaerobe	*Butyrivibrio fibrisolvens 16/4*	Doi, Kosugi (2004)
					1	0	Anaerobe	*Clostridium papyrosolvens DSM 2782*	Doi, Kosugi (2004)
					1	0	Anaerobe	*Clostridium thermocellum ATCC 27405*	Doi, Kosugi (2004)
					1	0	Anaerobe	*Ruminococcus flavefaciens FD-1*	Doi, Kosugi (2004)
					1	0	Anaerobe	*Clostridium cellulovorans 743B, ATCC 35296*	Doi, Kosugi (2004)
					1	0	Anaerobe	*Ruminococcus albus 7*	Suen, Stevenson et al. (2011)
					1	0	Anaerobe	*Acetivibrio cellulolyticus CD2, DSM 1870*	Doi, Kosugi (2004)
					1	0	Anaerobe	*Clostridium cellulolyticum H10*	Doi, Kosugi (2004)
					1	0	Anaerobe	*Ruminococcus albus 8*	Doi, Kosugi (2004)
					1	0	Anaerobe	*Clostridium acetobutylicum ATCC 824*	Doi, Kosugi (2004)
					0	1	Anaerobe	*Caldicellulosiruptor hydrothermalis 108*	Blumer-Schuette, Lewis et al. (2010)
					0	1	Anaerobe	*Caldicellulosiruptor bescii Z-1320, DSM 6725*	Blumer-Schuette, Lewis et al. (2010)
					0	1	Anaerobe	*Caldicellulosiruptor obsidiansis OB47*	Blumer-Schuette, Lewis et al. (2010)
					0	1	Anaerobe	*Caldicellulosiruptor kristjanssonii 177R1B, DSM 12137*	Blumer-Schuette, Lewis et al. (2010)
					0	1	Anaerobe	*Caldicellulosiruptor lactoaceticus 6A, DSM 9545*	Blumer-Schuette, Lewis et al. (2010)
					0	1	Anaerobe	*Caldicellulosiruptor saccharolyticus DSM 8903*	Blumer-Schuette, Lewis et al. (2010)
					0	1	Anaerobe	*Spirochaeta thermophila DSM 6192*	Bergquist, Gibbs et al. (1999)
					0	1	Anaerobe	*Caldicellulosiruptor kronotskyensis 2002*	Blumer-Schuette, Lewis (2010)
					0	1	Anaerobe	*Caldicellulosiruptor owensensis OL*	Blumer-Schuette, Lewis (2010)
					0	1	Aerobe	*Xylanimonas cellulosilytica DSM 15894*	Anderson, Abt et al. (2012)
					0	1	Aerobe	*Cellvibrio japonicus Ueda107*	Himmel, Xu et al. (2010)
					0	1	Aerobe	*Thermobifida fusca YX*	Lynd, Weimer et al. (2002)
					0	1	Aerobe	*Thermomonospora curvata DSM 43183*	Anderson, Abt et al. (2012)
					0	1	Aerobe	*Acidothermus cellulolyticus 11B*	Lynd, Weimer et al. (2002)
					0	1	Aerobe	*Cellulomonas flavigena 134, DSM 20109*	Lynd, Weimer et al. (2002)
					0	1	Aerobe	*Saccharophagus degradans 2-40*	Himmel, Xu et al. (2010)
					0	1	Aerobe	*Sorangium cellulosum So ce 56*	Himmel, Xu et al. (2010)
					0	0	Anaerobe	*Fibrobacter succinogenes subsp. succinogenes S85*	Wilson (2011)
					0	0	Aerobe	*Cytophaga hutchinsonii ATCC 33406*	Wilson (2011)
					0	0	Anaerobe	*Dictyoglomus turgidum DSM 6724*	Brumm, Hermanson et al. (2011)
					0	?	Anaerobe	*Clostridium phytofermentans ISDg*	Tolonen, Chilaka et al. (2009)
					?	?	Anaerobe	*Bryantella formatexigens I-52, DSM 14469*	not described
					?	?	Anaerobe	*Cellulosilyticum lentocellum RHM5, DSM 5427*	not described
					?	?	Anaerobe	*Eubacterium cellulosolvens 6*	not described
					?	?	Aerobe	*Amycolatopsis mediterranei U32*	not described
					?	?	Aerobe	*Streptomyces lividans TK24*	not described
					?	?	Aerobe	*Teredinibacter turnerae T7901*	not described
					?	?	Anaerobe	*Bacteroides cellulosilyticus DSM 14838*	not described

Figure 2 Occurrences of plant biomass degradation modules (PDMs) in organisms using different degradation paradigms. The predicted occurrences of the PDMs M1 to M5 in the genomes of 38 known lignocellulose degraders are indicated by different colors. Each PDM was predicted to be present or absent from a genome, depending on its genome-specific weight, that is, the degree of completeness for its protein families. Two major cellulose degradation paradigms – the free enzyme and cellulosome-based strategies – were assigned to the organisms according to the literature. Assignments can be ambiguous; for example, *Clostridium thermocellum* seems to be able to use mixed strategies [47]. No PDM was exclusively associated with these two paradigms, including M5, which, in addition to the cohesin and dockerin domains of cellulosomes, also included non-cellulosomal protein families (Table 3).

(see Additional file 9: Table S1), indicating that the additional families were relevant for the distinction between degrading and non-degrading species. Thus, even the weakly associated families seem to be predictive for plant biomass degradation. We decided to use the stringent cut-off value $C = 0.01$, as in our previous study, to infer smaller functional modules that could be more easily interpreted in terms of the functional contexts that they represented.

Another family with rare occurrences in the input set was GH44 (endoglucanases and xyloglucanases [62]), which appeared in less than 2% of our data samples, and was not grouped into any module. This family does not seem to be essential for all lignocellulose degraders, as its catalytic activities are also covered by the CAZy families GH5, GH9, and GH16 (Table 2) [11]. Overall, the observed rarity of GH6, GH44, and GH48 might indicate that they are non-universal across lignocellulose-degrading species. However, it might also be possible that more remote homologs exist that were not identified with the current Pfam and CAZy models.

PDMs mapping to known gene clusters of essential lignocellulose degradation genes

The gene clusters in known degrader genomes that were identified based on the protein families of the individual PDMs included well-characterized clusters of lignocellulolytic genes. For example, the modules M1 and M5 mapped to the *cip-cel* operon and the *xyl-doc* gene cluster in *C. cellulolyticum* H10 (Figure 3). *Cip-cel* encodes genes that are essential for cellulose degradation; *xyl-doc* encodes hemicellulose degradation genes [57]. The genes from both clusters have a multi-domain architecture with catalytic and carbohydrate-binding domains [57]. Within M1, GH5, GH9, and CBM4 occurred in *cip-cel,* while CBM6, CBM35, GH10, GH43, PF00756, and PF02018 have been annotated for *xyl-doc*. Genes from both clusters also included the cohesin and dockerin domains, which reflects the cellulosome-based degradation paradigm used by *C. cellulolyticum* H10.

Interestingly, LDA assigned the cohesin and dockerin domains to the M5 module, despite their co-occurrence

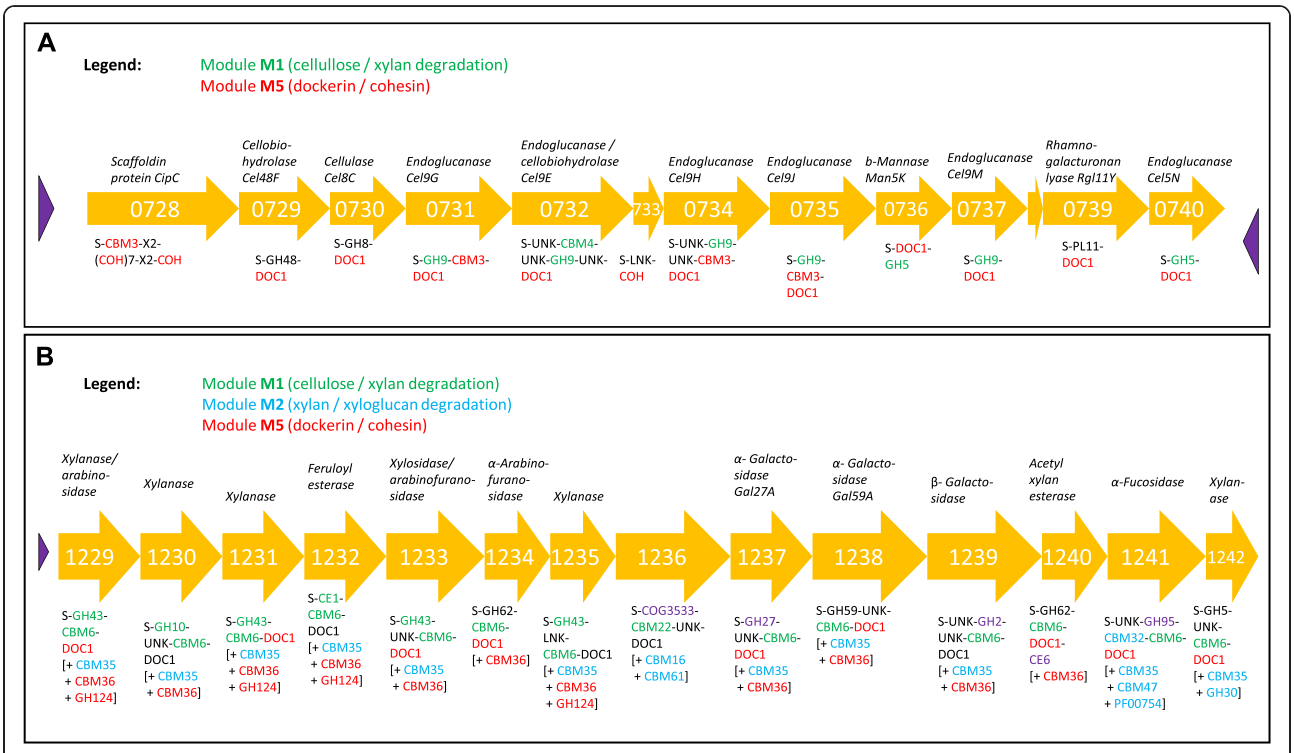


Figure 3 Plant biomass degradation modules (PDMs) mapping to the *cel-cip* and *xyl-doc* gene clusters in *Clostridium cellulolyticum* H10. Blouzard *et al.* described two clusters of genes that are involved in cellulose and hemicellulose degradation [57], and we adopted their domain architecture in this study. Abbreviations used for the carbohydrate-binding module (CBM) and glycoside hydrolase (GH) architecture are S, signal sequence; DOC1, dockerin type-I module; COH, cohesin type-I module; LNK, linker sequence; UNK, unknown function. We marked additional predicted domains as part of our in-house annotation sets using [+family X]. Some dockerin annotations were filtered out by our bit score criterion. **(A)** Genes from the *cel-cip* operon (Ccel_0728 to Ccel_0740) are essential for the cellulose degradation ability of the organism *C. cellulolyticum* H10, which uses the cellulosome strategy. The cluster includes multiple protein families of the PDMs M1 and M5. Although the consensus modules of M1 and M5 did not directly include the two endoglucanase families GH8 and GH48, associations between M1 and GH8, and between M5 and GH48 existed (probability values ≥ 0.005 in the respective topic probability distributions). **(B)** Genes from the 32-kb *xyl-doc* gene cluster (Ccel_1229 to Ccel_1242) encode functionalities for hemicellulose degradation. The cluster includes multiple protein families of the PDMs M1, M2, and M5, which together cover most of the cluster. Some additional protein families originate from M3 and M4 (purple). We assumed the following correspondences: CE1 ~ PF00756 (esterase); CBM22 ~ PF02018, and COG3533 (an uncharacterized protein in bacteria) ~ PF07944 (a putative glycosyl hydrolase of unknown function, DUF1680). The *xyl-doc* cluster contains a xylosidase/arabinofuranosidase gene (Ccel_1233), which is characterized as a putative β-xylosidase in the Integrated Microbial Genomes database (IMG). The gene corresponds to β-xylosidase genes in *Caldicellulosiruptor saccharolyticus* (Csac_2411), *Bacteroides cellulosilyticus* (BACCELL_02584 and BACCELL_00858), and *Fibrobacter succinogenes* (FSU_2269/ Fisuc_1769). Clusters containing M1 protein families were also detected around these genes.

with the M1 families in *cip-cel* and *xyl-doc*. This is probably due to the existence of M1 families in the genomes of organisms that do not have cellulosomes, such as *Thermobifida fusca*, which is a model organism for the free enzyme paradigm (see Additional file 2: Section 6). M1 also mapped to a hemicellulolytic gene cluster in *F. succinogenes* [63,64], an organism without cellulosomes that uses an unknown degradation strategy (see Additional file 6: Figure S1). Despite the evidence for a link between M5 and the cellulosome strategy, none of the PDMs proved to be exclusive for a particular degradation paradigm (Figure 2). As described above, the M5 module also contained five Pfam families whose functional descriptions have no known link to lignocellulose degradation (Table 3). These five Pfam families shared co-occurrence patterns with the cohesin and dockerin domains, but they also occurred in organisms using free cellulolytic enzymes, such as some *Caldicellulosiruptor* species (Figure 4). Thus, M5 also covered non-cellulosome-related functionalities (see Additional file 2: Section 7).

Predicting the ability for plant biomass degradation

We predicted the presence of PDMs for the 3,096 remaining genomes and taxonomic metagenome bins if their completeness was greater than or equal to the threshold determined for each PDM (see Methods; see weight thresholds in Table 4). Overall, the presence of one or more PDMs was predicted for 8.4% (28/332) of the taxonomic bins and 24.7% (683/2,764) of the genomes (see

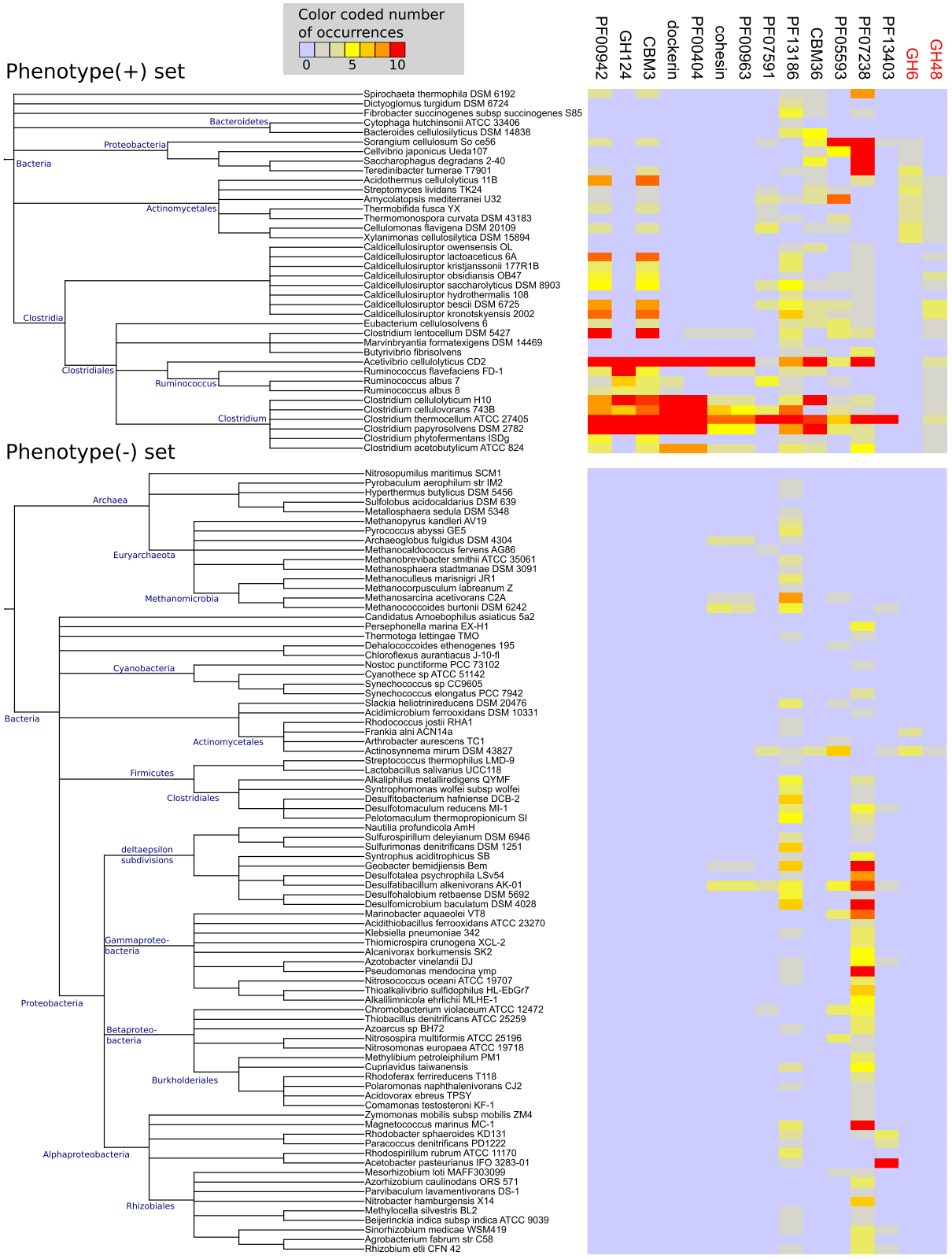


Figure 4 (See legend on next page.)

(See figure on previous page.)
Figure 4 Co-occurrences of the M5 protein families with GH6 and GH48 across known degraders and non-degraders. Two heat maps display the combined co-occurrence profiles for the M5 protein families and two additional cellulases, GH6 and GH48, across the known sets of the phenotype-positive and phenotype-negative genomes, respectively. GH6 and GH48 were not assigned into the plant biomass degradation module M5; in the case of GH48, this was only because of our strict cutoff criteria. However, GH48 was weakly associated with M5, and belonged to the top 50 families of the majority of M5 modules that were used to construct the consensus module. The colors of the heat map cells represent the number of instances of each family in the respective genomes of the organisms (see legends and note that the counted number of instances was limited to a maximum of 10 per genome, as described in Methods). The phylogenetic relationships of the genomes are indicated by dendrograms alongside the rows of the heat maps.

Additional file 10: Tables S1-5). Most genomes and bins to which M1 was assigned also had M2 assigned to them (82% of 132 M1 assignments occurred jointly with M2 assignments). This agreed with the cellulose/hemicellulose-degrading (M1) and hemicellulose-targeting (M2) enzymatic activities we determined for these modules, which are both essential for lignocellulose degradation [44]. The majority (52.5%) of all predictions was exclusive to M4 (see Additional file 11: Venn diagram). As M4 included functionally diverse glycan degradation families, and had the lowest precision (82.1%) of all modules for lignocellulose degraders, these assignments probably reflect a general ability of the respective organisms to degrade carbohydrate substrates of plant origin.

In a previous study [44], Medie *et al.* analyzed the distributions of CAZy families representing cellulases, hemicellulases, and pectinases across approximately 1,500 complete bacterial genomes. The authors classified almost 20% of these organisms as saprophytic bacteria, based on the presence of at least one cellulase and three or more hemicellulases or pectinases. Saprophytes feed on dead organic matter of plant origin, and thus are likely to include lignocellulose-degrading species. Based on the same CAZy families and criteria as described by Medie *et al.* [44], we determined potential saprophytes in our dataset (see Methods). In total, about a quarter (27.2%) of all 3,216 genomes and metagenome bins fulfilled these criteria. The number of predicted saprophytes thus further supports the notion that the ability to degrade plant biomass is a common trait in Bacteria and Archaea species. The genomes and metagenome bins with predicted PDM occurrences were clearly enriched with potential saprophytes (75% of all predictions). This enrichment was particularly large for M1 (99%), M2 (91%), and M3 (100%).

The metagenome bins that were assigned PDMs came from cow rumen, reindeer rumen, manatee gut, Tammar wallaby gut, and termite hindgut samples, and samples of a methylotrophic and a terephthalate-degrading community. Most of these communities, except for the methylotrophic and terephthalate-degrading ones, are known to include lignocellulose-degrading community members; however, their taxonomic affiliations are only partly known [19,65,66]. The coverage and quality of the protein-coding sequences was heterogeneous across the 332 bins that we analyzed: 63 bins could only be annotated with fewer than 10 protein families, while the remaining bins were annotated with 276 different protein families on average. It is well known that the gene content of metagenome bins is often incomplete, particularly for community members of low abundance, which is caused by insufficient sequencing depth or insufficient DNA read lengths [67]. Overall, the PDMs were predicted to be present in 28 bins covering 5 major taxonomic clades (Figure 5). PDMs occurring in metagenome bins of Bacteroides, Prevotella, and Lachnospiraceae (Clostridiales) were in line with the taxonomic affiliations of cellulose degraders found in mammalian gut and rumen microbial communities [68]. Furthermore, the PDMs accurately identified Bacteroidales and Treponema bins that have been shown to be involved in lignocellulose degradation in recent metagenome studies of cow rumen [69], and termite hindgut [60], thus indicating the benefit of our method to guide the discovery of uncultured microbial taxa with lignocellulolytic activities. Our results also indicated two archaeal extremophile species that have plant biomass degradation capabilities (see Additional file 2: Section 8).

Identification of gene clusters and polysaccharide utilization loci in the predicted (meta-)genomes

To identify new candidate clusters of genes encoding the ability to degrade lignocellulosic plant biomass, we searched for gene clusters encoding PDM protein families in the 711 genomes and taxonomic bins with assigned PDMs, using the same criterion as above. We found 379 gene clusters of 4 or more genes for individual PDMs, which mapped to 342 distinct gene clusters in 168 genomes and 6 taxonomic bins. Genome fragmentation caused by incomplete assembly of bacterial draft genomes from IMG and taxonomic bins in our dataset may have decreased the number of detected clusters. The average distance between the genes was 369 bp, which was almost the same as the average intracluster gene distance observed for the detected clusters in the phenotype-positive organisms. Most of the gene clusters occurred in Bacteroidetes (54.3%); 22.4% and

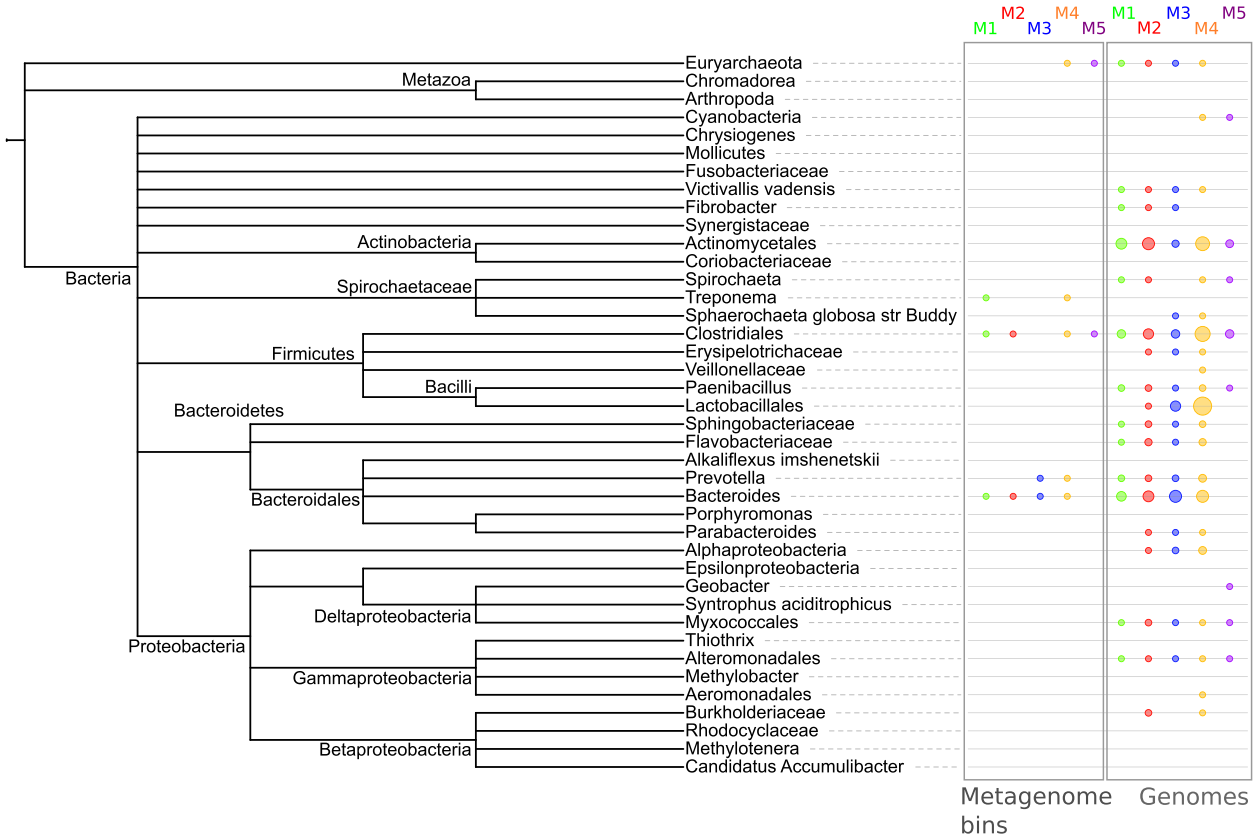


Figure 5 Comparison of plant biomass degradation module (PDM) occurrences in metagenome bins and isolate genomes of corresponding taxa. The colored circles at the leaf nodes of the tree denote the predicted occurrences of the different PDMs in the respective taxa or their subclades. The tree was constructed from the taxonomic assignments of the metagenome bins in our input set (see Methods). We then mapped the predicted occurrences of the PDMs in 28 different metagenome bins to the leaf nodes of the tree. If PDMs were identified in two or more bins, with one being a parental taxon to the other, the parent taxon is displayed. For example, predictions for *Prevotella ruminicola* were summarized with other predictions for the genus *Prevotella*. In addition, PDM occurrences in 403 isolate genomes of corresponding taxa were also mapped to the leaf nodes. The area of the colored circles for the isolate genomes was sized proportionally to the number of genomes for which the respective PDM was identified. The PDM predictions for the metagenome bins covered only five major taxonomic clades. The majority of PDMs in the metagenome bins were assigned to the orders Bacteroidales and Clostridiales. For the genomes, PDMs were identified from a broader range of taxa, including Actinobacteria, Firmicutes, Bacteroidetes, and Proteobacteria, in agreement with the estimated taxonomic range of potential cellulose-degrading species reported in [45]. The differences in the taxonomic distributions of the identified PDMs between genomes and metagenome bins may partly reflect the abundance of Bacteroidales and Clostridiales degraders in plant biomass-degrading communities, but it is also likely that some PDMs were not identified in the metagenome bins because of the partial nature of the genomic information recovered.

12.7% occurred in Firmicutes and Actinobacteria, respectively. The first two phyla are predominant in gut and rumen environmental communities with lignocellulose-degrading abilities [68,70].

Some of the newly identified gene clusters may cover polysaccharide utilization loci (PULs) targeting various kinds of polysaccharides. We found gene clusters in 39 isolate Bacteroides species, which are generally known to possess PULs [65]. As an example, the pectin-related PDM M3 identified gene clusters in *B. thetaiotaomicron* that represent parts of two regions that have been shown to be active in rhamnogalacturonan degradation in a PUL-targeted study [71]. Moreover, LDA inferred a stable functional module related to PULs, which included a suite of outer membrane proteins as well as the two core proteins that are known to define PULs, namely SusD- and SusC(TonB)-like membrane proteins (see Additional file 1: Table S6A). This module was not one of the high-ranking modules, which can be explained by the broad substrate specificity of PULs for various polysaccharides, including starch in particular [15,65]. While analyzing gene clusters of PDM protein families, we found hybrid gene clusters linking the PUL module to the glycoside hydrolases of the PDMs M1 and M2. For example, we identified gene clusters corresponding to previously characterized Sus-like PULs from *Bacteroides ovatus* targeting xyloglucan and xylan [71] (see Additional file 2: Section 9).

Predicting the ability for plant biomass degradation in a cow rumen microbial community

Hess *et al.* [19] reconstructed 15 draft genomes from the metagenome of a switchgrass-degrading microbial community from a cow rumen. In a previous study (Weimann *et al.* [28]), we have cross-linked the data from cellulolytic enzyme screens of the study by Hess *et al.* with protein family annotations of the 15 draft genomes to identify potential plant biomass degraders from the cow rumen metagenome. Strikingly, all 4 families (GH5, GH9, GH10, and GH26), which have been described by Weimann *et al.* to correspond to the (hemi-)cellulolytic enzymes of the cow rumen bins with degradation abilities confirmed by activity screens (see Table 3 in Weimann *et al.* [28]), were grouped together in the PDM M1 in our study.

In the current study, we investigated whether PDM assignments allowed identification of the plant biomass-degrading community members in the cow rumen metagenome (Table 5). The presence of M1 or M2 identified all degraders, in agreement with the enzyme screen results and our previous assignments with a family-centric SVM classifier [28]. M1 was also present in the draft genome APb, for which no lignocellulolytic enzymes were confirmed, but which is closely related to a known plant biomass-degrading species (*Butyrivibrio fibrisolvens*). The PDMs mapped to six gene clusters with four or more genes and several shorter clusters in the draft genomes. We investigated these, and found an interesting cluster in the Bacteroidales-associated draft genome AGa, containing genes annotated with GH5, GH94, and two unannotated gene sequences (see Additional file 2: Section 10; see Additional file 12: Figure S1; see protein sequences in Additional file 13).

Table 5 Module based identification of potential biomass-degrading cow rumen draft genomes

Draft genome	Taxonomic affiliation	M1	M2	M3	M4	M5
AJ[a,b]	Bacteroidales	+	+	+	+	
AGa[a,b]	Bacteroidales	+				
AC2a[a,b]	Bacteroidales		+			
Ala[a,b]	Clostridiales	+			+	
APb[b,c]	Clostridiales	+			+	
AH[b]	Bacteroidales	+			+	
AFa	Spirochaetales				+	
AN	Clostridiales				+	
AWa	Clostridiales				+	
ATa	Clostridiales					+
ADa	Myxococcales					
AMa	Spirochaetales					
AQ	Bacteroidales					
AS1a	Clostridiales					
BOa	Clostridiales					

[a]AJ, AGa, AC2a and Ala are supported by evidence for lignocellulolytic activity according to carbohydrolytic activity tests [19].
[b]The draft genomes AJ, AGa, AC2a, Ala, APb and AH were also predicted by an SVM-based method for predicting lignocellulose degraders (counting only the unambiguous predictions of the SVM classifier) [28].
[c]APb was mapped using 16S rRNA marker genes to the known lignocellulose-degrading organism *B. fibrisolvens* [19].
We used the weights of the consensus plant biomass degradation modules (PDMs) in the 15 draft genomes of the cow rumen metagenome to predict the draft genomes with lignocellulolytic activities (indicated by + signs), using the weight thresholds reported in Table 4.

Conclusions

Degradation of lignocellulosic plant biomass is a complex biological process with a number of mechanisms across different microbial species, which are currently only partially understood. In this paper, we describe functional modules of protein families linked to plant biomass degradation, which we identified based on co-occurrence patterns and partial phenotype information. Using LDA, a state of the art Bayesian inference method, we inferred 400 potential modules from a set of 2,884 genomes and 332 taxonomic bins from 18 metagenomes. Such modules represent sets of functionally coupled protein families, and cover a broad range of biochemical processes, as shown previously [42]. We then determined the presence of modules in genomes of known lignocellulose-degrading species and non-degraders to calculate a ranking of the modules that reflected the strength of their association with the plant biomass degradation phenotype. We analyzed the stability of the top-ranking modules across several executions of the LDA method, and determined five consensus functional modules (PDMs) involved in plant biomass degradation.

For the ranking of the modules, we used the learning set from our previous study [28], in which we had linked individual protein families to plant biomass degradation, and extended it by 20 additional phenotype-positive organisms. Despite of these additions, the number of confirmed degrader species is still small, compared with the estimated abundance of potential plant biomass-degrading species reported in two other studies [44,45]. Based on these estimates, 20 to 25% of bacterial genomes could possess plant biomass degradation capabilities. The unsupervised topic model of LDA allowed us to also include genomes and taxonomic metagenome bins lacking phenotype information in the inference process. The modules could thus be inferred from the known phenotype-positive genomes as well as from currently unknown degraders and cellulolytic communities in the dataset. To our knowledge, this is the first study to globally analyze the available genome sequence and phenotype data to determine the functional modules

of protein families that are linked to plant biomass degradation.

The PDMs included many protein families known to be involved in the degradation of cellulose, xylan, xyloglucan, and pectins, which are the main components of plant cell walls, with families that target the same macromolecules being grouped together. Overall, the PDMs contained 87 CAZy and Pfam families. We discussed 41 of these in detail in the functional contexts of the PDMs in which they were placed. Nineteen CAZy families were also represented by additional isofunctional Pfam families in the PDMs. These two sets account for 60 of the 87 families. Of the 27 remaining families, 7 were carbohydrate-active families, such as GH55, GH88, GH95, or alpha-amylase, which are involved in the degradation of polysaccharides. The remaining 20 families are listed in Table 3. Their functions were less clear, and they represent candidate families with potential roles in plant biomass degradation. Even more potentially interesting families were found in the high-ranking modules, but were not included in the consensus modules because they occurred in less than half of the modules used to construct the consensus. Some of these families might be interesting for further investigation.

The functional coherence of PDM member families was also supported by their localization in gene clusters in lignocellulolytic microbes. These included several known clusters of lignocellulolytic enzymes, such as *cip-cel* and *xyl-doc* from *C. cellulolyticum* H10. Based on the modules, we identified overall more than 400 gene clusters in different organisms of our dataset, which could potentially be linked to the degradation processes. These clusters may include PULs targeting different kinds of polysaccharides. We discussed some examples of identified PULs and Sus-like PULs that have been described as targeting rhamnogalacturonan, xyloglucan, and xylan in previous studies.

In addition, we investigated whether certain modules were specific to different degradation paradigms, as the module M5, for example, contained cellulosome-related families, such as cohesin, dockerin, and CBM3. We found that none of the modules was exclusive to a specific degradation strategy, and the modules instead spanned different paradigms. We believe that the granularity of the modules could be further improved in future if more and better curated phenotype information becomes available, which would allow us to enrich the set of genomes with species having different confirmed paradigms. For instance, the identification of genes from Sus-like cellulose-interacting protein complexes, as reported by Pope and Mackenzie [65], and Naas *et al.* [69], would probably require more accurate profile hidden Markov models (HMMs) for *susD*-like genes. For these, the sequences of relevant genes in more organisms that use the Sus-like paradigm would need to be known. Within our learning set, only *B. cellulosilyticus* uses a Sus-like strategy on hemicellulosic polysaccharides [72].

The PDMs allowed us to predict the ability of lignocellulose degradation with cross-validation accuracies of up to 96.7%, which we used to predict the ability to degrade plant biomass for all genomes and taxonomic bins with unknown degradation status in our dataset. The predicted degraders were clearly enriched with organisms that were likely to have a saprophytic lifestyle. For 15 draft genomes of a microbial community from a cow rumen, we confirmed the predictions by cross-linking to enzymes with demonstrated lignocellulolytic activities. In addition, the PDMs identified metagenome bins with cellulolytic capabilities for several microbial communities.

The PDMs contained many of the protein families that we had previously identified with a family-centric approach in a smaller set of 19 known lignocellulose degraders and 3 metagenomes, including CBM3, CBM4, CBM_4_9, CBM6, GH5, GH10, GH26, GH43, GH55, GH88, and GH95 [28]. Nevertheless, differences in the results existed. For example, in our previous study [28], only a few pectin-related families were identified (PL1, GH88, and GH106), but in the current study we identified an entire PDM (M3) of pectin-degrading families, which included these three families together with PL3, PL9, GH28, GH105, CE12, and additional related ones. Differences were also found for individual families. For example, the PDMs were linked to GH9, GH48, cohesin, and dockerin, as well as to elements of xylan binding, such as the CBM13 and lectin domains, which were not identified with the family-centric approach. However, GH6 and GH44 were not associated with the PDMs.

The families GH6, GH44 and GH48 occurred in less than 5% of the input genomes and metagenome bins, and their co-occurrence patterns with other families were more subtle in our large data collection than in the smaller dataset analyzed previously. These observations, which are in agreement with previous reports about the absence of GH6 and GH48 in the genomes and metagenomes of known lignocellulose-degrading species and microbial communities [17,20,24,60], suggest that GH6, GH44, and GH48 are not universally present in lignocellulose-degrading bacteria. However, we cannot exclude the possibility that remotely related family members that perform the functions of GH6, GH44 or GH48 are encoded in these genomes, which were not detected by the currently available family-specific HMM profile models. This could be further investigated by experimental screening for these enzymatic activities and identification of the respective proteins from the taxa that seem to lack these families.

In addition to differences in dataset sizes and composition, methodological differences between the family-centric and the PDM-based approach are likely to

be responsible for the differences observed and the additional relevant families that were included in the PDMs. Neither approach identified any gene families related to lignin degradation. This may be because lignin-related protein domains, except for the broadly defined peroxidase family PF00141, were largely missing from the Pfam and CAZy/dbCAN databases. Furthermore, reports of lignin decomposition have been dominated by fungi [73], and thus the corresponding mechanisms might have been under-represented in our bacterial and archaeal dataset.

We found evidence for functional links of the protein families in the PDMs with each other and the plant biomass degradation phenotype, which includes the co-occurrences of these families across genomes, co-occurrences with known relevant families, clustering within the genomes of known degraders, and the predictive value of the PDMs for identifying plant biomass degraders. Given this extensive support, an experimental characterization of the protein families with unknown relevance for plant biomass degradation in the PDMs, and their respective gene clusters, is likely to reveal new biochemical functionalities for plant biomass degradation. With the method we have described, other phenotypes, such as, nitrogen fixation or antibiotic resistance, could be studied from existing genome datasets in a similar fashion.

Methods

Latent Dirichlet allocation

LDA is a text-mining method for extracting semantic concepts (that is, topics) from a collection of text documents [41]. The topics reflect groups of semantically related words supported by co-occurrence signals across the document collection. LDA is a generative probabilistic model assuming a well-defined process as the source of the observed documents. With Bayesian inference and MCMC methods such as Gibbs sampling, the generative process can be reversed [74,75], which corresponds to increasing the probability of the model by fitting latent variables to make the outcome of the process match the observed documents as closely as possible. Here, we are interested in inferring the latent variables, not in the outcome of the process itself.

The input for LDA is a collection of N documents, where each document is a collection of words stemming from a controlled vocabulary V. The order of words in the document is not important (termed the "bag of words" assumption). LDA assumes the existence of T latent topics, and each topic is represented as a discrete multinomial distribution over V.

One variable of the model with central meaning is the vector $\vec{z}$, which contains a random variable z for each word of the text collection that models the latent origin of the word with respect to the T topics. According to the model, the probability of observing word w in document d of the collection is given by:

$$P(w|d) = \sum_{t=1}^{T} \underbrace{P(w|z=t)}_{\phi_t(w)} \cdot \underbrace{P(t|d)}_{\theta_d(t)},$$

where $\phi_t(w)$ defines the multinomial distribution representing topic t, and $\theta_d(t)$ corresponds to a multinomial distribution describing the document-specific prior probabilities of the topics. The parameters $\vec{z}$, $\phi_t(w)$, and $\theta_d(t)$ for all documents and topics are latent variables of the hidden generative process, which can be estimated efficiently with MCMC sampling methods.

Genome and metagenome annotation

Protein sequences for bacterial and archaeal species were downloaded from IMG (version 3.4) and metagenomic protein sequences were obtained from IMG with microbiome samples (IMG/M, V3.3). In addition, we collected samples of microbial communities from Svalbard reindeer rumen [20], termite hindgut [60], manatee gut, and Tammar wallaby forestomach [66], as well as draft genomes reconstructed from a metagenome sample of a switchgrass-degrading microbial community in a cow rumen [19]. If no protein-coding sequences were available, genes were predicted by MetaGeneMark [76]. Note that, from the metagenomes, only protein sequences with a predicted taxonomic origin were included in our dataset. For this purpose, taxonomic bins from IMG/M or the original publications, or those generated in-house were used, which were inferred with either *PhyloPythia* [77] or *PhyloPythiaS* [78] using sample-specific training sequences and taxonomic models constructed with taxa that represent the more abundant community populations. Overall, we worked with protein-coding sequences from 2,884 prokaryotic genome sequences and 332 taxonomic bins derived from 18 metagenome samples. Protein sequences were annotated with profile HMMs of protein families from the Pfam database (Pfam-A, V26.0) and CAZy families from the dbCAN database [79] using HMMER 3.0 [80]. Multiple matches of different domains per protein were allowed. All matches were required to satisfy an e-value of 1e-02 and a bit score of 25 or more. For Pfam, the family-specific thresholds from the Pfam database (gathering thresholds) were used if they were stricter than our default threshold. Such family-specific thresholds were not available from dbCAN. For large families (more than 100 amino acids) of the dbCAN database, we used the threshold 1e-04 instead of 1e-02. We then converted the protein family annotations for the genomes and taxonomic metagenome bins into a suitable input collection for LDA (see Additional file 3: Section 1).

Homology-based annotation of protein families can generate some FP or FN annotations [81], which may affect the accuracy of the downstream analysis. Therefore, robust computational methods capable of handling potential annotation errors should be chosen to obtain reliable results. Bayesian probabilistic models such as LDA are well suited for the inference of robust associations from potentially noisy datasets [82,83].

Functional module inference with LDA

We used the protein family collection of the (meta-) genomes as input for the LDA inference procedure to predict potential functional modules, as demonstrated previously [42]. Note that the identifiers of the protein families (e.g. 'GH28') were used to define the words of the vocabulary V in the LDA model. Because of the larger input collection compared with our previous work, we increased the number of topics from 200 to 400. Despite the increased number of documents (3,216 *versus* 575), there was a slight decrease in the vocabulary size (8,413 *versus* 10,431), owing to differences in coverage between the Pfam-A and eggNOG databases. As in [42], we used the parameter value $C = 0.01$ to convert topic probability distributions into discrete sets of protein families, which represented our potential functional modules. Given the vocabulary V, and the multinomial distribution over words of V for topic t, that is, the topic distribution $\phi_t(w)$, we defined module M_t as $M_t := \{w \in V | \phi_t(w) \geq C\}$. The module M_t thus contained the protein family identifiers that were most strongly related to topic t. The families assigned to M_t share common co-occurrence patterns, and were therefore likely to be functionally coupled based on the "guilt by association" principle [39].

Phenotype annotation

We assigned the lignocellulose-degrading phenotype to genomes by manually curating the annotations of "(ligno)cellulose degradation" or "(plant) biomass degradation" that we obtained from the databases of IMG, the Genomes Online Database (GOLD) [84], and the German Collection of Microorganisms and Cell Cultures (DSMZ) [85], based on information from the literature. Removal of ambiguous or inconsistent phenotype annotations resulted in 38 confirmed lignocellulose degraders (phenotype-positive genomes), which degraded some or all components of lignocellulose (see Additional file 14: Table S1). The set of phenotype-positive genomes is a superset of the 19 lignocellulose-degrading microbes (except *Postia placenta*) from our previous work [28]. We adopted the set of 82 phenotype-negative genomes from the same study, which were also manually curated using information from the literature. There was less certainty in phenotype-negative annotations, as it may be that a particular phenotype has not been discussed in the literature; however, the statistical methods we used to determine PDMs from these datasets can tolerate a certain amount of error.

Definition of module weights

The inference of a topic model with LDA from a collection of N input documents results in T potential functional modules. We extracted 400 modules from 3,216 genomes and metagenome bins. We then applied an attribute-ranking approach to sort the modules according to their relevance for lignocellulose degradation. As attributes to be used in the ranking procedure, we defined module weights. A weight, $weight_t(d)$, should reflect how likely the module M_t is to be contained in the genome or metagenome bin encoded as document d of the input collection. Given N genomes or bins as input, and T modules, we can summarize the weights in a weight matrix $W \in \mathbb{R}^{N \times T}$ with entries $w_{dt} := weight_t(d)$.

Two different definitions of weights (probability weights and completeness scores) were tested (see Additional file 3: Section 2). We decided to use completeness scores, as they produced more relevant results, though the rankings obtained with both choices of weights largely agreed (see Additional file 2: Section 11). The completeness score of a module is the percentage of the protein families of a module that occurred in a specific genome or taxonomic bin. More precisely, we defined the weight of module M_t in document d of the (meta-)genome collection based on completeness as:

$$weight_t(d) := \frac{|M_t \cap d|}{|M_t|} \times 100\%,$$

where $|M_t \cap d|$ is the size of the intersection of the protein family sets of module M_t and document d, and $|M_t|$ is the number of protein families contained in M_t.

Identification of phenotype-defining functional modules

To identify phenotype-associated modules, we used the weights of the modules in the input documents that corresponded to the manually curated phenotype-positive and phenotype-negative genomes. We refer to these genomes as the "learning set." The selected weights were used to predict the phenotypes of these genomes, and we scored each of the 400 modules according to its ability to distinguish between the two phenotype classes. More precisely, the classification of the learning set with respect to module M_t was carried out by applying a threshold value γ_t to the weights of the module, that is, genomes were predicted to be phenotype-positive if the respective weights satisfied the threshold, or phenotype–negative if they did not.

The ranking procedure optimized independent thresholds for all modules by finding the threshold that maximized a criterion function. We used the F-measure [86] with the parameter $\beta = 0.5$ for scoring (recall half as important as precision [54]; see Additional file 3: Section 3), which can be computed using the confusion matrix shown in Table 6. Finally, we obtained the ranking of the modules by sorting them in decreasing order, based on their F-measure scores.

Mapping of modules between Gibbs samples and runs

Finding the optimal assignments of protein families to functional modules, such that the observed data can be explained in the best possible way, is a combinatorially challenging task. We used Gibbs sampling to derive statistical estimates for the latent topic distributions of the LDA model, from which we derived the potential functional modules as described. We then searched for similar modules across several LDA runs to identify stable modules, because with the MCMC inference technique used, there is variance in the derived estimates across different runs. We used the Kullback–Leibler divergence [42] and the Jaccard distance [87] to calculate pairwise distances between topics (probability distributions) or modules (discrete protein family sets), respectively. As expected, we observed good agreement between the results with both distance measures. Given the matrix of pairwise distances for the modules of two LDA runs, we used the Hungarian algorithm [88] to find an optimal global mapping between these. The Bron–Kerbosch algorithm [89] was used to find cliques of similar modules efficiently across multiple LDA runs (see Additional file 3: Section 4).

Consensus modules

In theory, Gibbs sampling efficiently estimates the posterior distribution of the model parameters and converges to a global optimum given a sufficient number of iterations [46]. However, in practice, there is variance in the results of individual LDA runs, and a common approach to derive a stable solution is to repeat the inference multiple times and to compare the results from a number of runs [74]. Therefore, we repeated the steps of our analysis several times with the same input data. In comparison with our previous study [42], we doubled the number of LDA runs to 18. In each run, we inferred 400 potential functional modules. As described in the previous section, we tracked the identities of the modules across all runs based on pairwise module distances, and thus characterized the stability of the modules. Next, we applied the described attribute-ranking scheme based on the completeness scores to each of the 18 sets of 400 inferred modules, and determined the top 15 modules for each run. Among these highly ranked modules from different runs, we searched for similar modules that occurred in at least 75% of the 18 runs. From these recurring modules, we derived consensus modules of protein families (see Additional file 1: Tables S1A-5A) as follows. Given a set of similar modules from different LDA runs, which were identified as representing a stable module across 14 runs or more, the corresponding consensus module contained all protein families that occurred in at least 9 modules of this set.

Table 6 Confusion matrix

M_t	Document d is phenotype-positive	Document d is phenotype-negative
$weight_t(d) \geq \gamma_t$	TP	FP
$weight_t(d) < \gamma_t$	FN	TN

FN, false negative; FP, false positive; TN, true negative; TP, true positive.

LOO analysis and 10-fold cross-validation

For the consensus PDMs, we performed LOO and 10-fold cross-validation experiments to assess their predictive accuracy. In a loop, we successively left out each individual genome (or 10% of the genomes) of the learning set, and optimized the weight threshold of a module on the remaining learning set with the F-measure. For the omitted genomes, the PDM was predicted to be present if the genome-specific module weight was greater than or equal to the inferred threshold. In both settings, we obtained exactly one prediction for each genome of the learning set, based on which we calculated performance measures, such as precision and recall, the F-measure, the cross-validation accuracy, and the cross-validation macro-accuracy. For the 10-fold cross-validation experiments, we randomly split the data to create the different folds. The procedure was repeated 10 times, and the results were averaged. For a more accurate estimate of the test error, we also calculated 95% confidence intervals for the cross-validation accuracies of the modules. We used the Clopper–Pearson bound [53], which is an estimate based on the binomial distribution and the observed error rate on the omitted test samples. Note that the number of available test samples (120 in our case) is an important parameter of the binomial, and determines the sizes of the intervals. With a larger set, narrower bounds would be obtained.

Prediction of module occurrences in genomes and metagenome bins

We optimized the cutoff thresholds γ_t for module prediction by maximizing the F-measure using the weights of the consensus modules for all genomes with a known phenotype (that is, the genomes of our learning set). We then considered the module weights in the genomes and

metagenome bins of unknown phenotype to predict occurrences of the modules. We applied the following prediction rule to predict the presence of a module M_t in the genome or metagenome bin that corresponds to document d in the input of LDA:

$$predict\left(d, M_t, \gamma_t\right) = \begin{cases} 1 & if\ \ weight_t(d)\ \ \geq \gamma_t \\ 0 & otherwise \end{cases}$$

Comparison of PDM occurrences in the taxonomic bins of metagenomes and isolate genomes of the corresponding clades

We constructed a tree based on the NCBI taxonomy tree with the tool iTOL [90] for the taxa represented by the metagenome bins in the dataset. Metagenome bins with fewer than 10 protein families were excluded from consideration. We used the taxonomic assignments inferred by the binning methods *PhyloPythia* [77] and *PhyloPythiaS* [78], except for high-ranking bins, such as bacteria. To visualize the common occurrences of the PDMs at the leaf nodes of the tree, we collapsed some of the original leaf nodes to new leaf nodes of higher ranks. This was carried out if two or more of the PDMs were predicted to occur in taxa of the same clade, but with different ranks. In these cases, the PDMs involved were displayed for the highest common observed rank. The PDMs were predicted to occur in the bins of only five major taxa across the different metagenomes (Figure 5). In addition, we also mapped isolate genomes of the corresponding taxa with predicted occurrences of the PDMs to the leaf nodes of the tree.

Identification of saprophytic genomes and taxonomic bins

As described by Medie *et al.* [44], we classified a genome or metagenome bin as belonging to a cellulase- and hemicellulase-containing saprophyte if the corresponding annotation set contained at least one cellulase and three or more hemicellulases or pectinases from the following families.

- Cellulase families: GH5, GH6, GH8, GH9, GH12, GH44, GH45, GH48, GH74, and GH124.
- Hemicellulase and pectinase families: GH10, GH11, GH26, GH28, GH30, GH43, GH53, GH67, GH78, PL1, PL2, PL9, PL10, PL11, and PL22.

Implementation and parameter settings

We used the LDA implementation from the topic modeling toolbox [91]. The LDA model depends on two hyperparameters, α and β, which control the Dirichlet priors of the multinomial distributions. We used the default values of the topic modeling toolbox, that is, $\alpha = 50/T$ and $\beta = 0.01$. We specified 2,000 iterations as the burn-in phase of a run. After burn-in, we collected 50 Gibbs samples and derived the topic distributions by averaging over the samples (see Additional file 3: Section 5).

Additional files

Additional file 1: Protein families of the consensus plant biomass degradation modules (PDMs). The Tables S1A-5A show each consensus module as a list of Pfam/CAZy terms. The consensus modules summarize highly similar modules from the 18 LDA runs and contain all elements that occurred in nine runs or more. The tables S1B-5B contain information about all additional Pfam/CAZy families that occurred in the similar modules in less than nine runs. Tables S6A and S6B list the families of the additional PUL module (see Results section).

Additional file 2: Supplementary Note. Additional details about the results of the main manuscript.

Additional file 3: Supplementary Methods. Additional details about the methods and preparation of the input data.

Additional file 4: Plant biomass degradation module (PDM) assignments to the genomes of the learning set. PDM assignments to the genomes of known plant biomass degraders and non-degraders are visualized in Figure S1, and were obtained by leave-one-out classification.

Additional file 5: Single predictions of the consensus modules on the learning set of genomes. Each sheet of the Excel file lists the predictions of one of the consensus modules with respect to the learning set of genomes (Tables S1-5). Table S6 contains the predictions of the additional PUL module. We used different colors to mark true-positive (TP), true-negative (TN), false-positive (FP) and false-negative (FN) predictions (a description of the color coding is contained in the first sheet of the file). For each classified sample, we have provided several details (for example, name and phylum), as well as the genome-specific module weight (completeness score) of the respective consensus module.

Additional file 6: Hemicellulolytic gene cluster in *Fibrobacter succinogenes* S85. The gene cluster in Figure S1 encodes more than 10 hemicellulose-targeting enzymes in the genome of *F. succinogenes* S85. The protein domain architecture of the cluster genes has been described by Yoshida *et al.* [63,64]. *F. succinogenes* does not use a cellulosome-based degradation strategy, but rather a degradation paradigm that is still uncharacterized [92,93].

Additional file 7: Co-occurrence profiles of the M1 protein families and GH6/GH48 across the learning set. Two heat maps display the combined co-occurrence profiles of the M1 protein families and two additional cellulases, GH6 and GH48, across the sets of the known phenotype-positive (Figure S1) and phenotype-negative (Figure S2) genomes, respectively. GH6 and GH48 were not assigned to module M1. The colors of the heat map cells represent the number of instances of each family in the respective genomes of the organisms (see legends and note that the counted number of instances was limited to a maximum of 10 per genome, as described in Methods). The phylogenetic relationships of the genomes are indicated by dendrograms alongside the rows of the heat maps.

Additional file 8: Co-occurrence profiles of protein families that were weakly associated with M5 (across the learning set). Two heat maps displaying protein family co-occurrence profiles across the known phenotype-positive (Figure S1) and phenotype-negative (Figure S2) genomes. The columns of the heat maps represent the families that were weakly associated with the M5 module. These families did not satisfy the required threshold $C = 0.01$, but they belonged to the 50 protein families with the highest probabilities in the 16 topics that were used to create the M5 consensus module. Because the families failed to match the threshold, they were not counted for the consensus of M5. One example for such a family is GH48 (discussed in the main text). Families are ordered from left to right according to the number of topics in which they occurred. Families that occurred in less than 9 of the 16 topics are not displayed. We also added the cohesin and dockerin domains

of M5 for a comparison. The colors of the heat map cells encode the number of instances of each family in the respective genomes of the organisms.

Additional file 9: Leave-one-out results for the consensus plant biomass degradation modules (PDMs) obtained with the threshold $C = 0.005$. The threshold $C = 0.01$ was used to convert the discrete topic probability distributions of the LDA model into potential functional modules (see Methods). In additional tests, we used the threshold $C = 0.005$ instead. This cutoff level is less strict and allowed families with smaller probabilities to be included in the potential functional modules. This also resulted in enlarged consensus modules for the PDMs. Table S1 summarizes the results obtained in leave-one-out validation for the PDMs M1 to M5 based on the threshold $C = 0.005$.

Additional file 10: Single predictions of the consensus modules on the remaining set of genomes and metagenome bins. Each sheet of the Excel file lists the predictions of one of the consensus modules with respect to all genomes and metagenome bins, except for the 120 known (non-)degraders used for learning (Tables S1-5). Table S6 contains the predictions of the additional PUL module. For each classified sample, we provide several details (for example, name and phylum), as well as the genome-specific module weight (completeness score) of the respective consensus module.

Additional file 11: Venn diagram of the predicted occurrences of the modules M1 to M4. The diagram displays the overlap between the genomes and metagenome bins with predicted occurrences of the modules M1, M2, M3, and M4. Genomes from the learning set were excluded.

Additional file 12: Gene cluster in the cow rumen draft genome AGa. The red box in Figure S1 marks a gene cluster (NODE_457020_ORF_01660 to NODE_457020_ORF_01710), which was identified based on the families assigned to highest scoring module (M1). The cluster is located on a 97,191-bp contig of the draft genome AGa (Bacteroidales) from the cow rumen metagenome [19]. The cluster includes three cellulases, based on assignments of the GH5 family, and a cellobiose phosphorylase (GH94; EC 2.4.1.20) with an attached putative carbohydrate binding domain (PF06204). The GH94 family was not assigned to the consensus module of M1, but it was contained in the M1 modules in 7 of 18 LDA runs. Depending on the presence or absence of GH94 in the M1 modules of different runs, the gene cluster was identified either partly or completely. The cluster includes two genes (genes 01680 and 01690; green rectangle) without any annotated functional domains; these are uncharacterized genes that may be relevant for the degradation of lignocellulose. The presence of two Pfam families related to the major facilitator superfamily in gene 01640 (marked by the yellow box) indicates a link between the (hemi)cellulases of the GH5 and GH94 families, and sugar-binding or transport proteins located in the outer membrane (see Additional file 2: Section 10).

Additional file 13: Protein sequences of the identified gene cluster in the cow rumen draft genome AGa. Protein sequences (NODE_457020_ORF_01620 to NODE_457020_ORF_01740) from the cow rumen metagenome representing a gene cluster in the draft genome AGa (discussed in the main manuscript), as well as its surrounding genes.

Additional file 14: Microbial isolate strains (lignocellulose degraders and non-degraders) that were used as the learning set. Table S1 represents a manually curated list of 120 phenotype-positive or phenotype-negative prokaryotic genomes, including the respective literature references.

Abbreviations

CAZyme: Carbohydrate-active enzyme; CBM: Carbohydrate-binding module; CE: Carbohydrate esterase; DUF: Domain of unknown function; EC: Enzyme classification; GH: Glycoside hydrolase; LDA: Latent Dirichlet allocation; PDM: Plant biomass degradation module; PL: Polysaccharide lyase; PUL: Polysaccharide utilization locus; Sus: Starch utilization system; SVM: Support vector machine.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SGAK and ACM designed the study, interpreted the results, and wrote the manuscript; SGAK conducted the experiments under the supervision of ACM; SGAK and AW computed the CAZy and Pfam protein annotations, and curated the learning set. PBP was involved in the interpretation of the results and revised the final manuscript. All authors read and approved the manuscript.

Acknowledgements

SGAK, AW, and ACM were supported by the Max-Planck society and Heinrich Heine University Düsseldorf. PBP gratefully acknowledges support from the Research Council of Norway (Project number 214042).

Author details

[1]Max-Planck Research Group for Computational Genomics and Epidemiology, Max-Planck Institute for Informatics, University Campus E1 4, Saarbrücken 66123, Germany. [2]Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Post Office Box 5003, 1432 Ås, Norway. [3]Department of Algorithmic Bioinformatics, Heinrich Heine University Düsseldorf, Düsseldorf 40225, Germany.

Received: 19 April 2014 Accepted: 5 August 2014
Published online: 09 September 2014

References

1. Kumar R, Singh S, Singh OV: **Bioconversion of lignocellulosic biomass: biochemical and molecular perspectives.** *J Ind Microbiol Biotechnol* 2008, **35:**377–391.
2. Kohse-Hoinghaus K, Osswald P, Cool TA, Kasper T, Hansen N, Qi F, Westbrook CK, Westmoreland PR: **Biofuel combustion chemistry: from ethanol to biodiesel.** *Angew Chem Int Ed Engl* 2010, **49:**3572–3597.
3. Himmel ME, Ding SY, Johnson DK, Adney WS, Nimlos MR, Brady JW, Foust TD: **Biomass recalcitrance: engineering plants and enzymes for biofuels production.** *Science* 2007, **315:**804–807.
4. Gowen CM, Fong SS: **Exploring biodiversity for cellulosic biofuel production.** *Chem Biodivers* 2010, **7:**1086–1097.
5. Xing MN, Zhang XZ, Huang H: **Application of metagenomic techniques in mining enzymes from microbial communities for biofuel synthesis.** *Biotechnol Adv* 2012, **30:**920–929.
6. Minic Z, Jouanin L: **Plant glycoside hydrolases involved in cell wall polysaccharide degradation.** *Plant Physiol Biochem* 2006, **44:**435–449.
7. Burton RA, Gidley MJ, Fincher GB: **Heterogeneity in the chemistry, structure and function of plant cell walls.** *Nat Chem Biol* 2010, **6:**724–732.
8. Sweeney MD, Xu F: **Biomass converting enzymes as industrial biocatalysts for fuels and chemicals: Recent developments.** *Catalysts* 2012, **2:**244–263.
9. Gilbert HJ, Stalbrand H, Brumer H: **How the walls come crumbling down: recent structural biochemistry of plant polysaccharide degradation.** *Curr Opin Plant Biol* 2008, **11:**338–348.
10. Jayani RS, Saxena S, Gupta R: **Microbial pectinolytic enzymes: a review.** *Process Biochem* 2005, **40:**2931–2944.
11. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B: **The carbohydrate-active enzymes database (CAZy): an expert resource for glycogenomics.** *Nucleic Acids Res* 2009, **37:**D233–D238.
12. Morais S, Barak Y, Lamed R, Wilson DB, Xu Q, Himmel ME, Bayer EA: **Paradigmatic status of an endo- and exoglucanase and its effect on crystalline cellulose degradation.** *Biotechnol Biofuels* 2012, **5:**78.
13. Wilson DB: **Microbial diversity of cellulose hydrolysis.** *Curr Opin Microbiol* 2011, **14:**259–263.
14. Fontes CM, Gilbert HJ: **Cellulosomes: highly efficient nanomachines designed to deconstruct plant cell wall complex carbohydrates.** *Annu Rev Biochem* 2010, **79:**655–681.
15. Martens EC, Koropatkin NM, Smith TJ, Gordon JI: **Complex glycan catabolism by the human gut microbiota: the Bacteroidetes Sus-like paradigm.** *J Biol Chem* 2009, **284:**24673–24677.
16. Bolam DN, Koropatkin NM: **Glycan recognition by the Bacteroidetes Sus-like systems.** *Curr Opin Struct Biol* 2012, **22:**563–569.
17. Wilson D: **Evidence for a novel mechanism of microbial cellulose degradation.** *Cellulose* 2009, **16:**723–727.
18. Horn SJ, Vaaje-Kolstad G, Westereng B, Eijsink VG: **Novel enzymes for the degradation of cellulose.** *Biotechnol Biofuels* 2012, **5:**45.

19. Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T, Mackie RI, Pennacchio LA, Tringe SG, Visel A, Woyke T, Wang Z, Rubin EM: **Metagenomic discovery of biomass-degrading genes and genomes from cow rumen.** *Science* 2011, **331:**463–467.
20. Pope PB, Mackenzie AK, Gregor I, Smith W, Sundset MA, McHardy AC, Morrison M, Eijsink VG: **Metagenomics of the Svalbard reindeer rumen microbiome reveals abundance of polysaccharide utilization loci.** *PLoS ONE* 2012, **7:**e38571.
21. Graham JE, Clark ME, Nadler DC, Huffer S, Chokhawala HA, Rowland SE, Blanch HW, Clark DS, Robb FT: **Identification and characterization of a multidomain hyperthermophilic cellulase from an archaeal enrichment.** *Nat Commun* 2011, **2:**375.
22. Kim SJ, Lee CM, Han BR, Kim MY, Yeo YS, Yoon SH, Koo BS, Jun HK: **Characterization of a gene encoding cellulase from uncultured soil bacteria.** *FEMS Microbiol Lett* 2008, **282:**44–51.
23. Wang F, Li F, Chen G, Liu W: **Isolation and characterization of novel cellulase genes from uncultured microorganisms in different environmental niches.** *Microbiol Res* 2009, **164:**650–657.
24. Duan C-J, Feng J-X: **Mining metagenomes for novel cellulase genes.** *Biotechnol Lett* 2010, **32:**1765–1775.
25. Rubin EM: **Genomics of cellulosic biofuels.** *Nature* 2008, **454:**841–845.
26. Park BH, Karpinets TV, Syed MH, Leuze MR, Uberbacher EC: **CAZymes Analysis Toolkit (CAT): web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database.** *Glycobiology* 2010, **20:**1574–1584.
27. Wang PI, Marcotte EM: **It's the machine that matters: predicting gene function and phenotype from protein networks.** *J Proteomics* 2010, **73:**2277–2289.
28. Weimann A, Trukhina Y, Pope PB, Konietzny SG, McHardy AC: ***De novo*** **prediction of the genomic components and capabilities for microbial plant biomass degradation from (meta-)genomes.** *Biotechnol Biofuels* 2013, **6:**24.
29. Kastenmüller G, Schenk ME, Gasteiger J, Mewes HW: **Uncovering metabolic pathways relevant to phenotypic traits of microbial genomes.** *Genome Biol* 2009, **10:**R28.
30. Yosef N, Gramm J, Wang Q-F, Noble WS, Karp RM, Sharan R: **Prediction of phenotype information from genotype data.** *Commun Inf Syst* 2010, **10:**99–114.
31. Vey G, Moreno-Hagelsieb G: **Metagenomic annotation networks: construction and applications.** *PLoS ONE* 2012, **7:**e41283.
32. Padmanabhan K, Wilson K, Rocha AM, Wang K, Mihelcic JR, Samatova NF: **In-silico identification of phenotype-biased functional modules.** *Proteome Sci* 2012, **10**(Suppl 1)**:**S2.
33. Slonim N, Elemento O, Tavazoie S: ***Ab initio*** **genotype-phenotype association reveals intrinsic modularity in genetic networks.** *Mol Syst Biol* 2006, **2:**1–14.
34. Lingner T, Muhlhausen S, Gabaldon T, Notredame C, Meinicke P: **Predicting phenotypic traits of prokaryotes from protein domain frequencies.** *BMC Bioinformatics* 2010, **11:**481.
35. Jeffery C: **Moonlighting proteins: implications and complications for proteomics.** *Protein Sci* 2004, **13:**124–124.
36. Liu B, Pop M: **MetaPath: identifying differentially abundant metabolic pathways in metagenomic datasets.** *BMC Proc* 2011, **5**(Suppl 2)**:**S9.
37. Schmidt MC, Rocha AM, Padmanabhan K, Shpanskaya Y, Banfield J, Scott K, Mihelcic JR, Samatova NF: **NIBBS-search for fast and accurate prediction of phenotype-biased metabolic systems.** *PLoS Comput Biol* 2012, **8:**e1002490.
38. De Filippo C, Ramazzotti M, Fontana P, Cavalieri D: **Bioinformatic approaches for functional annotation and pathway inference in metagenomics data.** *Brief Bioinform* 2012, **13:**696–710.
39. Aravind L: **Guilt by association: contextual information in genome analysis.** *Genome Res* 2000, **10:**1074–1077.
40. Kensche PR, van Noort V, Dutilh BE, Huynen MA: **Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution.** *J R Soc Interface* 2008, **5:**151–170.
41. Blei DM, Ng AY, Jordan MI: **Latent dirichlet allocation.** *J Mach Learn Res* 2003, **3:**993–1022.
42. Konietzny SG, Dietz L, McHardy AC: **Inferring functional modules of protein families with probabilistic topic models.** *BMC Bioinformatics* 2011, **12:**141.
43. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P: **STRING: known and predicted protein-protein associations, integrated and transferred across organisms.** *Nucleic Acids Res* 2005, **33:**D433–D437.
44. Medie FM, Davies GJ, Drancourt M, Henrissat B: **Genome analyses highlight the different biological roles of cellulases.** *Nat Rev Microbiol* 2012, **10:**227–234.
45. Berlemont R, Martiny AC: **Phylogenetic distribution of potential cellulases in bacteria.** *Appl Environ Microbiol* 2013, **79:**1545–1554.
46. Gilks WR, Richardson S, Spiegelhalter DJ: *Markov Chain Monte Carlo in Practice.* Boca Raton, Florida, USA: Chapman and Hall/CRC; 1999.
47. Himmel ME, Xu Q, Luo Y, Ding S-Y, Lamed R, Bayer EA: **Microbial enzyme systems for biomass conversion: emerging paradigms.** *Biofuels* 2010, **1:**323–341.
48. Boraston AB, Bolam DN, Gilbert HJ, Davies GJ: **Carbohydrate-binding modules: fine-tuning polysaccharide recognition.** *Biochem J* 2004, **382:**769–781.
49. McCartney L, Blake AW, Flint J, Bolam DN, Boraston AB, Gilbert HJ, Knox JP: **Differential recognition of plant cell walls by microbial xylan-specific carbohydrate-binding modules.** *Proc Natl Acad Sci U S A* 2006, **103:**4765–4770.
50. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci U S A* 1999, **96:**2896–2901.
51. Ballouz S, Francis AR, Lan R, Tanaka MM: **Conditions for the evolution of gene clusters in bacterial genomes.** *PLoS Comput Biol* 2010, **6:**e1000672.
52. Duda RO, Hart PE, Stork DG: *Pattern Classification.605 Third Avenue.* New York, USA: John Wiley & Sons; 2012.
53. Anguita D, Ghelardoni L, Ghio A, Ridella S: **Test Error Bounds for Classifiers: A Survey of Old and New Results.** In *Proceedings of the IEEE Symposium on Foundations of Computational Intelligence (FOCI) 2011.* Paris, France; 2011:80–87.
54. Lewis DD: **Evaluating and optimizing autonomous text classification systems.** In *Proceedings of the 18th annual international ACM-SIGIR conference on Research and Development in Information Retrieval.* Seattle, WA: ACM; 1995:246–254.
55. Anderson I, Abt B, Lykidis A, Klenk HP, Kyrpides N, Ivanova N: **Genomics of aerobic cellulose utilization systems in actinobacteria.** *PLoS ONE* 2012, **7:**e39331.
56. Boraston AB, Tomme P, Amandoron EA, Kilburn DG: **A novel mechanism of xylan binding by a lectin-like module from *Streptomyces lividans* xylanase 10A.** *Biochem J* 2000, **350**(Pt 3)**:**933–941.
57. Blouzard J-C, Coutinho PM, Fierobe H-P, Henrissat B, Lignon S, Tardif C, Pagès S, de Philip P: **Modulation of cellulosome composition in *Clostridium cellulolyticum*: adaptation to the polysaccharide environment revealed by proteomic and carbohydrate-active enzyme analyses.** *Proteomics* 2010, **10:**541–554.
58. Kotake T, Dina S, Konishi T, Kaneko S, Igarashi K, Samejima M, Watanabe Y, Kimura K, Tsumuraya Y: **Molecular cloning of a b-galactosidase from radish that specifically hydrolyzes b-(1- > 3)- and b-(1- > 6)-galactosyl residues of arabinogalactan protein.** *Plant Physiol* 2005, **138:**1563–1576.
59. Olson DG, Giannone RJ, Hettich RL, Lynd LR: **Role of the *CipA* scaffoldin protein in cellulose solubilization, as determined by targeted gene deletion and complementation in *Clostridium thermocellum*.** *J Bacteriol* 2013, **195:**733–739.
60. Warnecke F, Luginbuhl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, Cayouette M, McHardy AC, Djordjevic G, Aboushadi N, Sorek R, Tringe SG, Podar M, Martin HG, Kunin V, Dalevi D, Madejska J, Kirton E, Platt D, Szeto E, Salamov A, Barry K, Mikhailova N, Kyrpides NC, Matson EG, Ottesen EA, Zhang X, Hernandez M, Murillo C, Acosta LG, *et al*: **Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite.** *Nature* 2007, **450:**560–565.
61. Schwarz WH: **The cellulosome and cellulose degradation by anaerobic bacteria.** *Appl Microbiol Biotechnol* 2001, **56:**634–649.
62. Kitago Y, Karita S, Watanabe N, Kamiya M, Aizawa T, Sakka K, Tanaka I: **Crystal structure of Cel44A, a glycoside hydrolase family 44 endoglucanase from *Clostridium thermocellum*.** *J Biol Chem* 2007, **282:**35703–35711.
63. Yoshida S, Hespen CW, Beverly RL, Mackie RI, Cann IK: **Domain analysis of a modular a-L-arabinofuranosidase with a unique carbohydrate binding strategy from the fiber-degrading bacterium *Fibrobacter succinogenes S85*.** *J Bacteriol* 2010, **192:**5424–5436.
64. Yoshida S, Mackie RI, Cann IK: **Biochemical and domain analyses of *FSUAxe6B*, a modular acetyl xylan esterase, identify a unique**

carbohydrate binding module in *Fibrobacter succinogenes S85*. *J Bacteriol* 2010, **192:**483–493.
65. Mackenzie AK, Pope PB, Pedersen HL, Gupta R, Morrison M, Willats WG, Eijsink VG: **Two SusD-like proteins encoded within a polysaccharide utilization locus of an uncultured ruminant bacteroidetes phylotype bind strongly to cellulose.** *Appl Environ Microbiol* 2012, **78:**5935–5937.
66. Pope PB, Denman SE, Jones M, Tringe SG, Barry K, Malfatti SA, McHardy AC, Cheng JF, Hugenholtz P, McSweeney CS, Morrison M: **Adaptation to herbivory by the Tammar wallaby includes bacterial and glycoside hydrolase profiles different from other herbivores.** *Proc Natl Acad Sci U S A* 2010, **107:**14793–14798.
67. Dröge J, McHardy AC: **Taxonomic binning of metagenome samples generated by next-generation sequencing technologies.** *Brief Bioinform* 2012, **13:**646–655.
68. Flint HJ, Bayer EA, Rincon MT, Lamed R, White BA: **Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis.** *Nat Rev Microbiol* 2008, **6:**121–131.
69. Naas AE, Mackenzie AK JM, Schückel J, Willats WGT, Eijsink VGH, Pope PB: **Do rumen Bacteroidetes utilize an alternative mechanism for cellulose degradation?** *mBio* 2014, **5:**e01401–e01414.
70. Morrison M, Pope PB, Denman SE, McSweeney CS: **Plant biomass degradation by gut microbiomes: more of the same or something new?** *Curr Opin Biotechnol* 2009, **20:**358–363.
71. Martens EC, Lowe EC, Chiang H, Pudlo NA, Wu M, McNulty NP, Abbott DW, Henrissat B, Gilbert HJ, Bolam DN, Gordon JI: **Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts.** *PLoS Biol* 2011, **9:**e1001221.
72. McNulty NP, Wu M, Erickson AR, Pan C, Erickson BK, Martens EC, Pudlo NA, Muegge BD, Henrissat B, Hettich RL, Gordon JI: **Effects of diet on resource utilization by a model human gut microbiota containing Bacteroides cellulosilyticus WH2, a symbiont with an extensive glycobiome.** *PLoS Biol* 2013, **11:**e1001637.
73. Floudas D, Binder M, Riley R, Barry K, Blanchette RA, Henrissat B, Martinez AT, Otillar R, Spatafora JW, Yadav JS, Aerts A, Benoit I, Boyd A, Carlson A, Copeland A, Coutinho PM, de Vries RP, Ferreira P, Findley K, Foster B, Gaskell J, Glotzer D, Górecki P, Heitman J, Hesse C, Hori C, Igarashi K, Jurgens JA, Kallen N, Kersten P, *et al*: **The paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes.** *Science* 2012, **336:**1715–1719.
74. Steyvers M, Griffiths T: **Probabilistic Topic Models.** In *Handbook of Latent Semantic Analysis. Volume 427.* Edited by Landauer T, McNamara D, Dennis S, Kintsch W. Colorado, USA: Laurence Erlbaum; 2007:427–440.
75. Griffiths TL, Steyvers M: **Finding scientific topics.** *Proc Natl Acad Sci U S A* 2004, **101**(Suppl 1):5228–5235.
76. Zhu W, Lomsadze A, Borodovsky M: **Ab initio gene identification in metagenomic sequences.** *Nucleic Acids Res* 2010, **38:**e132.
77. McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I: **Accurate phylogenetic classification of variable-length DNA fragments.** *Nat Methods* 2007, **4:**63–72.
78. Patil KR, Haider P, Pope PB, Turnbaugh PJ, Morrison M, Scheffer T, McHardy AC: **Taxonomic metagenome sequence assignment with structured output models.** *Nat Methods* 2011, **8:**191–192.
79. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y: **dbCAN: a web resource for automated carbohydrate-active enzyme annotation.** *Nucleic Acids Res* 2012, **40:**W445–W451.
80. Eddy SR: **Accelerated profile HMM searches.** *PLoS Comput Biol* 2011, **7:**e1002195.
81. Friedberg I: **Automated protein function prediction–the genomic challenge.** *Brief Bioinform* 2006, **7:**225–242.
82. Friedman N: **Inferring cellular networks using probabilistic graphical models.** *Science* 2004, **303:**799–805.
83. Wilkinson DJ: **Bayesian methods in bioinformatics and computational systems biology.** *Brief Bioinform* 2007, **8:**109–116.
84. Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC: **The genomes online database (GOLD) v. 4: status of genomic and metagenomic projects and their associated metadata.** *Nucleic Acids Res* 2012, **40:**D571–D579.
85. **Deutsche Sammlung von Mikroorganismen und Zellkulturen.** [http://www.dsmz.de/]
86. Van Rijsbergen CJ: *Information Retrieval.* 2nd edition. London, Boston: Butterworths; 1979.
87. Levandowsky M, Winter D: **Distance between sets.** *Nature* 1971, **234:**34–35.
88. Kuhn HW: **The Hungarian method for the assignment problem.** *Nav Res Log* 1955, **2:**83–97.
89. Bron C, Kerbosch J: **Algorithm 457: finding all cliques of an undirected graph.** *Commun ACM* 1973, **16:**575–577.
90. Letunic I, Bork P: **Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy.** *Nucleic Acids Res* 2011, **39:**W475–W478.
91. **Matlab Topic Modeling Toolbox.** [http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm]
92. Wilson DB: **Three microbial strategies for plant cell wall degradation.** *Ann N Y Acad Sci* 2008, **1125:**289–297.
93. Suen G, Weimer PJ, Stevenson DM, Aylward FO, Boyum J, Deneke J, Drinkwater C, Ivanova NN, Mikhailova N, Chertkov O, Goodwin LA, Currie CR, Mead D, Brumm PJ: **The complete genome sequence of *Fibrobacter succinogenes S85* reveals a cellulolytic and metabolic specialist.** *PLoS ONE* 2011, **6:**e18814.

doi:10.1186/s13068-014-0124-8
Cite this article as: Konietzny *et al.*: Inference of phenotype-defining functional modules of protein families for microbial plant biomass degraders. *Biotechnology for Biofuels* 2014 **7**:124.

CHAPTER 9

Related projects

We based our methods for functional module inference on latent Dirichlet allocation (LDA); however, the underlying concepts can of course be generalized to other topic models. In the following, LDA is compared with supervised topic models.

9.1 Supervised topic models

We have also tested supervised probabilistic topic models such as the *author-topic* model (Rosen-Zvi *et al.*, 2004), *sLDA* (McAuliffe and Blei, 2008), and *MedLDA* (Zhu *et al.*, 2012). These models allow an incorporation of document labels into the inference process. In text mining settings, such models were used, for example, to analyze the relationship between the usage of words (e.g. 'joy', 'entertained', 'tiring') and the final numerical rating (between 1 to 5) in review comments about movies (see, for example, McAuliffe and Blei (2008)).

Supervised models appear to be promising solutions to the problem of finding phenotype-related functional modules. We expected that the explicit use of *a priori* knowledge (in the form of phenotypic labels for genomes) could improve the discovery of phenotype-related functional modules. But in our tests with supervised topic models, we did not observe this effect.

9.1.1 Author topic model

The *author topic* model (Rosen-Zvi *et al.*, 2004) is distinct from the baseline LDA model because it introduces an additional layer in the hierarchical graphical model (Figure 9.1). The effect of this modification can be described as follows. The original LDA model proposed by Blei *et al.* (2003) implicitly assumes that the complete text corpus was written by one and the same author. In contrast to this, the design of the *author topic* model assumes multiple authors (see details below). Apart from that, the two models are almost identical. In particular, both models assume a fixed set of K latent topic distributions.

The benefit of the *author topic* model stems from the fact that one can incorporate meta knowledge about the relationships between topics. We provide the model a list of tagged authors along with every document. In other words, the author tags are made explicit, they are not treated as latent variables of the model like topic assignments. When we infer the *author topic* model from a tagged input dataset, we derive estimates for two different kinds of relationships. On the one hand, we see the associations between words and topics, just as in the case with the original LDA method. On the other hand, we also see associations of topics and authors, and we can interpret these associations as the favorite topics of each author.

The model can be adapted to the inference of phenotype-related functional modules under the following assumptions:

- The phenotypic traits of organisms are mainly caused by the activities of functional modules that are encoded in the genomes.

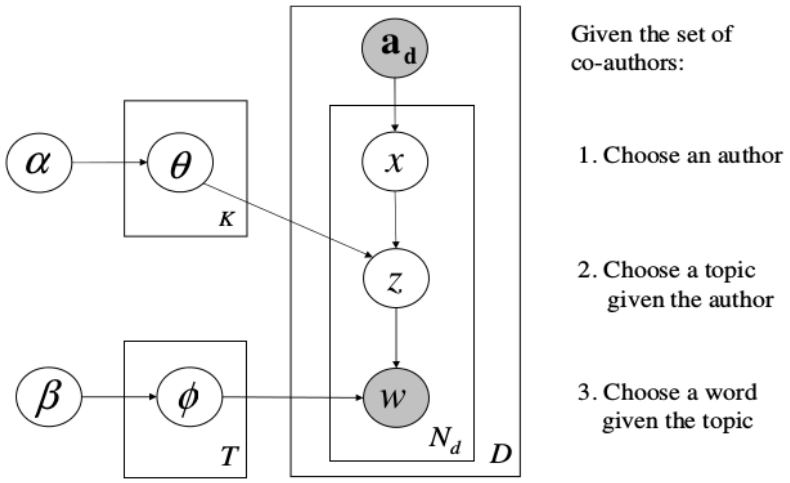


Figure 9.1: The author topic model in plate notation. Different to the baseline LDA model by Blei *et al.*, the generative process of the *author topic* model by Rosen-Zvi *et al.* does not assume a single author as the creator of all the documents. Instead of this, multiple authors could have contributed. This is a supervised approach where the information about the authors of documents is given to the model as *a priori* information in the form of documents labels. The variables have the same meanings as in the basic LDA model (Section B.2), except x and a_d. The displayed generative process describes how, at each iteration of the algorithm, a new word instance w gets created depending on the topic preferences of an author x, who is chosen at random from a given group of authors, a_d. (Image source: Steyvers *et al.* (2004))

- A single phenotypic trait may correspond to the activities of one or more functional modules.

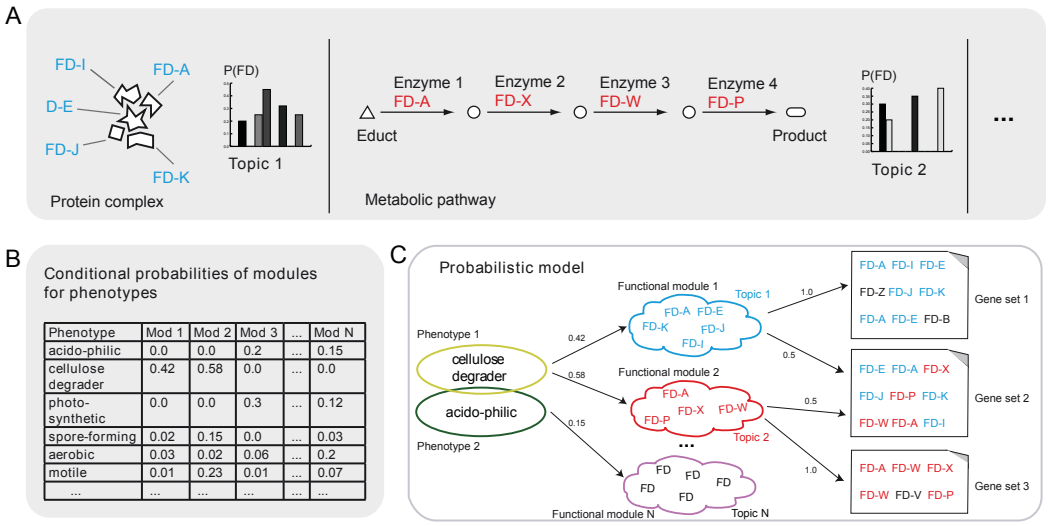
Based on these assumptions, we modeled the relationships between phenotypes and functional modules on the basis of the *author topic* model (Figure 9.2). Since some phenotypic traits of organisms can be observed by measurements, knowledge of their presence or absence for individual input genomes can be passed as *prior* information to the model. The basic idea is to use the phenotypic labels of organisms to define the author tags of the model. In this way, we can use phenotypic labels to perform a supervised inference of phenotype-related functional modules. It basically means that similar 'combinations' (distributions) of functional modules – observed in different genomes – likely correspond to similar phenotypes of the organisms. This closely follows

the analogy to the aforementioned relationship between a specific author and his favorite topics.

The proposed approach can handle multiple phenotypes at once. For testing the use of the *author topic* model, we acquired phenotypic information for genomes from IMG and the GOLD database (Liolios *et al.*, 2010). Unfortunately, the available datasets turned out to be sparse in phenotypic labels because only a few microbial phenotypes have been annotated consistently as of today. Among the phenotypes we could test was 'cell motility' which, along with 'oxygen requirement', is one of the most frequently annotated binary phenotypes. Strinkingly, the *author topic* model correctly related the motility phenotype with the flagellar module that we already discussed in Konietzny *et al.* (2011), a link that was easy to confirm. For almost all other phenotypes, it was not clear whether the inferred phenotype-module associations were correct or not. An in-depth evaluation of this question would have required well curated sets of different microbial phenotypes and appropriate gold standards to confirm the relevance of the identified potential functional modules for each individual phenotype. Such information is difficult to access because it is mostly contained in unstructured scientific articles.

9.1.2 Simulations with an artificial pathway

Prior to our choice of a specific topic model for the inference of phenotype-related functional modules, we had tested the capabilities of different models to detect a target pathway that is known to be responsible for the labeling of the samples in a controlled set of genomes. For this purpose, we modeled the protein family set of an artificial pathway which consisted of 11 elements and could occur in two slightly different variants. Each pathway element was represented by a hypothetical protein family identifier that we additionally introduced into the controlled vocabulary of identifiers. We distributed the set of pathway elements across a set of genomes, and labeled the pathway-containing genomes as phenotype-positives and the others as phenotype-negatives, respectively. To make the scenario realistic, we altered the pathway configuration by randomly leaving out some of its elements in certain genomes, or by using alternative elements which



Phenotype	Mod 1	Mod 2	Mod 3	...	Mod N
acido-philic	0.0	0.0	0.2	...	0.15
cellulose degrader	0.42	0.58	0.0	...	0.0
photo-synthetic	0.0	0.0	0.3	...	0.12
spore-forming	0.02	0.15	0.0	...	0.03
aerobic	0.03	0.02	0.06	...	0.2
motile	0.01	0.23	0.01	...	0.07
...	...	...	...	...	...

Figure 9.2: A proposed model for relationships between genes, functional modules, and microbial phenotypes. Panel A: Examples of functional modules, represented by a probability distribution over functional descriptors (FD-terms, e.g. COGs, Pfam-domains) for genes. **Panel B:** Inferred associations between phenotypic traits and functional modules, expressed as conditional probabilities. Each trait relates to a specific set of functional modules. **Panel C:** Ovals on the left represent phenotypic traits, each associated with several modules. Clouds in the middle represent the FD-terms with the highest associated probabilities for each module. Gene sets on the right represent (meta-)genomic contents. The direction of arrows represents the process of generating the observed data. This model could be expressed in form of a text mining model – the *author topic* model by Rosen-Zvi *et al.*.

should reflect the two different pathway variants. We then tested the different methods with the labeled set of genomes.

It turned out that a hybrid approach consisting of the LDA baseline method combined with a suitable attribute ranking scheme – published in Konietzny *et al.* (2014) – performed best in identifying the correct elements of the pathway. It was not possible to show a significant improvement with the other methods, and some supervised models like *MedLDA*[1] performed substantially worse in this task.

[1] *MedLDA* couples the concepts of probabilistic topic models and support vector machines in a single hybrid model.

9.2 Seeding of LDA topics

We were also trying to incorporate additional sources of *prior* information about the targeted cellular processes into our method. In many practical cases, there exist two different types of knowledge about phenotype-related processes: *i)* a range of organisms known (or assumed) to possess the phenotype, and *ii)* pre-existing characterizations of some of the key proteins that are essential for the functional processes of interest.

We planned to use a combination of two different strategies to exploit both types of *prior* knowledge. The formerly described ranking scheme for functional module weights already accounts for the first aspect. It guides the discovery of phenotype-related functional modules by considering phenotypic labels of the genomes. The second aspect should be covered by a new concept that we call 'seeding of topics'. The function of seeding is to account for *prior* knowledge which may be available about the key elements of the functional modules that are the targets of the analysis.

The key idea of seeding is to use already known protein families as 'seeds' for functional modules. Seeds serve to attract other elements which often co-occur together with them. Similar concepts of 'seeding' have been used in information retrieval (Andrzejewski and Zhu, 2009). Basically, we try to optimize the initial state of the LDA model before we actually begin to infer potential functional modules.

The obvious way of doing this would be to explicitly inform the model about some of the relationships between protein families and functional modules. In this case, seeding would have to be done manually, guided by expert knowledge. A biologist would need to define certain key protein families as the seeds of topics. However, this would be a tedious and time-consuming process, and we intended to automate the procedure in order to reduce manual efforts. We developed a two-step strategy that automatically computes a seeded topic model optimized for the identification of phenotype-related functional modules (Figure 9.3; Appendix C provides additional details):

1. The first step involves the computation of protein families that are significantly correlated with the binary phenotype across the input genomes, and running LDA

with this limited vocabulary only. In theory, this results in candidate functional modules that subdivide the set of phenotype-correlated families into groups of functionally-coupled items.

2. The resulting modules of step 1 define the 'seeds' for a new run of LDA, which uses the complete set of protein family identifiers to define the vocabulary. This step involves a modification of the initial topic assignments in the Gibbs sampling procedure of LDA, and a slight modification of the inference procedure to keep some of the assignments fixed (Appendix C provides the details).

The first step of the procedure should find clusters of protein families that represent drafts of phenotype-related functional modules. These are likely incomplete because LDA was only run on a subset of all protein families. The second step then adds potentially missing elements.

Step 2 mainly targets protein families with activities in both phenotype-related and non-related cellular processes (Figure 9.4). Examples for such families are multifunctional auxiliary enzymes with activities in several cellular processes. Multifunctional proteins with activities in processes of phenotype-positive and phenotype-negative organisms have weak correlation with the phenotypic labels, and would thus be missing in the input vocabulary of the first LDA run. However, they might play an active role in phenotype-related cellular processes, which means they need to be somehow functionally-coupled to the families selected in the first step. This suggests that there exist co-occurrence signals in the genomes which link the families. If this is the case, the missing ambiguous families are likely to be included into the correct seeded modules in the second run of LDA. The second step thus augments the seeded modules with the missing elements.

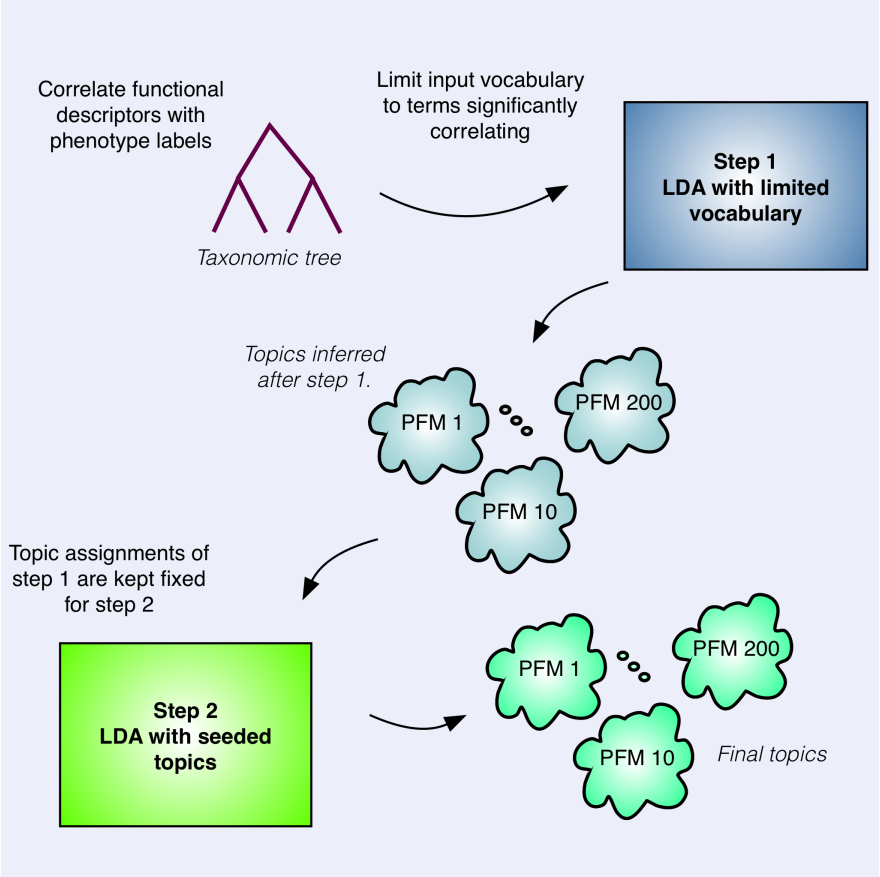


Figure 9.3: Fully-automated seeding of phenotype-specific topic models. The proposed approach consists of two consecutive LDA runs. The input vocabulary for the first LDA run gets limited to protein family identifiers that significantly correlate with the binary phenotype labels (for example, across a taxonomic tree). The objective of the first LDA run is to find a substructure of functional modules in the list of correlated protein families. However, the resulting modules will likely be incomplete, due to the fact that the majority of protein family identifiers were filtered out because they were not correlated with the phenotype. A second run of LDA with the full vocabulary and topic distributions that are seeded with information from the first run serves to complete the initially found modules with co-occurring elements.

9.3 Preliminary results

The effects of seeding depend on the stability of the results after the first step of the approach. The proposed two-step approach will only work, if LDA succeeds in the first step to infer stable modules from the reduced set of vocabulary terms. The main

goal of the second step is then to fine-tune the pre-determined module drafts that were generated in the first phase.

We conducted a large series of experiments to address this question in the context of microbial lignocellulose degradation:

1. We generated ten different random input sets, and included a common core of phenotype-positive labeled genomes into each one of it.

2. We then determined a subset of the protein families that were highly correlated with the phenotype status of the genomes in the input set (p-values ≤ 0.05, based on Pearson's pairwise correlation of the phylogenetic profiles and the phenotypic labels). Except from the known phenotype-positives, all other genomes were assumed to be phenotype-negative.

3. LDA was executed three times on each set with the phenotype-correlated vocabulary only, resulting in three independent topic models that were used for seeding in the successive step. We ranked the resulting modules with the attribute ranking scheme based on module weights to determine their associations with the phenotype.

4. Each seed initialized a corresponding consecutive run of LDA, performed on the same random set, but this time using the full vocabulary for creating the input collection. Once again, a ranking of the modules was performed afterwards.

5. We compared the top-ranked modules in the final runs with the top-ranked modules in the initial runs that were used for seeding.

Altogether, this amounts to 30 initial and 30 successive seeded runs on ten different input sets, meaning that we had a broad basis to assess the robustness of the approach. In an additional test, we repeated identically seeded runs to assess the stability of the second step of the seeding approach. Figure 9.5 provides an overview of the evaluation setting, showing the number of runs performed, and which runs served as sources or targets for seedings.

It should be noted that the unmodified version of LDA is an unsupervised method. Similar to unsupervised clustering algorithms, LDA does not assign consistent labels to the probabilistic word clusters that it infers. More precisely, the internal indexing $i \in (1 \dots K)$ of the topic distributions is not consistent across runs. If we search for stable topics across repeated executions of LDA, these cannot be identified by their index i, which is randomly initialized in each execution. In seeding, however, we transfer the topic index assignments of a subset of words (in a subset of documents) between two successive runs. Therefore, we expect to stabilize the identity of topics. Note that the number K of latent topics for seeded runs needs to be equal or exceed the number of topics in the initial run.

Overall, the effect of seeding could clearly be observed. For each pair of a seeded run and its corresponding initial run, we compared the topic at rank one, that is, the single top-ranked module. We found both an accordance of the topic identity (i.e. the internal index $i \in (1 \dots K)$ used by LDA) in 27/30 cases, as well as an agreement of the modules' protein family contents. Moreover, we observed that, in general, several of the top-5 modules of the rankings kept their identities. In contrast to that – for the unmodified version of LDA[2] – we would see topics with the same index i at first rank only by chance, with a probability of $1/K$.

The results clearly showed that seeding had an effect by stabilizing the identity of topics and, more importantly, that specific knowledge about groupings could be effectively transferred between the two phases of the seeding approach. The same could be observed for repeated runs with identical seedings.

At the time of these early analyses, our evaluation setup suffered from the insufficient knowledge about the true functional modules in lignocellulose degradation that was available. High-quality gold standards are needed to compare the performances of LDA and seeded topic models because qualitative differences depend on details of the processes, such as, for example, the precise protein family configurations of the true functional modules. Lacking a true gold standard dataset, we could not prove that

[2] In general, two successive LDA runs on the same input will infer similar results, but the indexing of topics will be different.

seeding significantly improved the quality of the results. We therefore concentrated our research efforts on the improvement and in-depth evaluation of the attribute ranking scheme that had proven to deliver promising results in our tests with the artificial pathway.

9.4 Conclusions

It was not possible to show that the more advanced supervised approaches improved upon the baseline LDA model. The main reasons for this were the good performance of LDA, together with a lack of suitable gold standards that would be needed to compare the differences of the models in sufficient detail.

While it is possible to compare the predictions of different methods with respect to known cellular processes (such as, for example, KEGG pathways), comparing their performance in unravelling unknown aspects of phenotype-related processes is very difficult. Therefore, we used our STRING-based criterion for the functional coherence of protein family modules (Konietzny *et al.*, 2011) to compare the results of different topic models. Compared to LDA, we could thus see an improvement with the unsupervised *collocation* LDA model (Section 6) but less improvement with the *author topic* model.

We considered complex phenotypes like cell motility, methane degradation, and lignocellulose degradation in more detail. Overall, the lack of fundamental knowledge about the true underlying functional modules for phenotypes like lignocellulose degradation (Rubin, 2008; Wilson, 2011) made it extremely difficult to compare the relative improvements between different models. Moreover, it was generally a challenge to identify phenotype-positive and phenotype-negative genomes for training and test sets. For example, with the available information from the literature, we could only compile a small list of relevant organisms with respect to methane degradation. This observation illustrates an important general problem: only a few microbial phenotypes, such as cell motility, are accurately annotated for large sets of organisms.

Using prior knowledge about the key protein families of target modules, such as in

the case of topic seeding, means additional challenges for the evaluation. The way how we incorporated knowledge is to keep the assignments of certain protein family members to functional modules fixed during the inference process, or at least at the beginning of the process (Figure C.1). This is a form of semi-supervision because not all members of a protein family need to be affected in the same way. Note that it would not suffice to globally fix the assignments of whole protein families to modules because the members of a protein family might or might not be involved in variants of a biological process across different genomes. As a suitable gold standard for semi-supervised module inference, we need species-specific annotations of functional modules that allow us to identify the corresponding genes in individual genomes for training and testing. However, the existing information sources about functional modules in different species, such as MetaCyc, contain relatively few data (or very short pathways – with only a couple of reaction steps), and they are biased towards only a few model organisms. If these small datasets need to be split into training and test sets, we can only use a sparse fraction of the proteins of the overall input collection to evaluate semi-supervised approaches on the level of individual protein family members.

For future work, it would likely be a useful approach to consider the final results from this thesis as a baseline for the comparison with more advanced probabilistic graphical models.

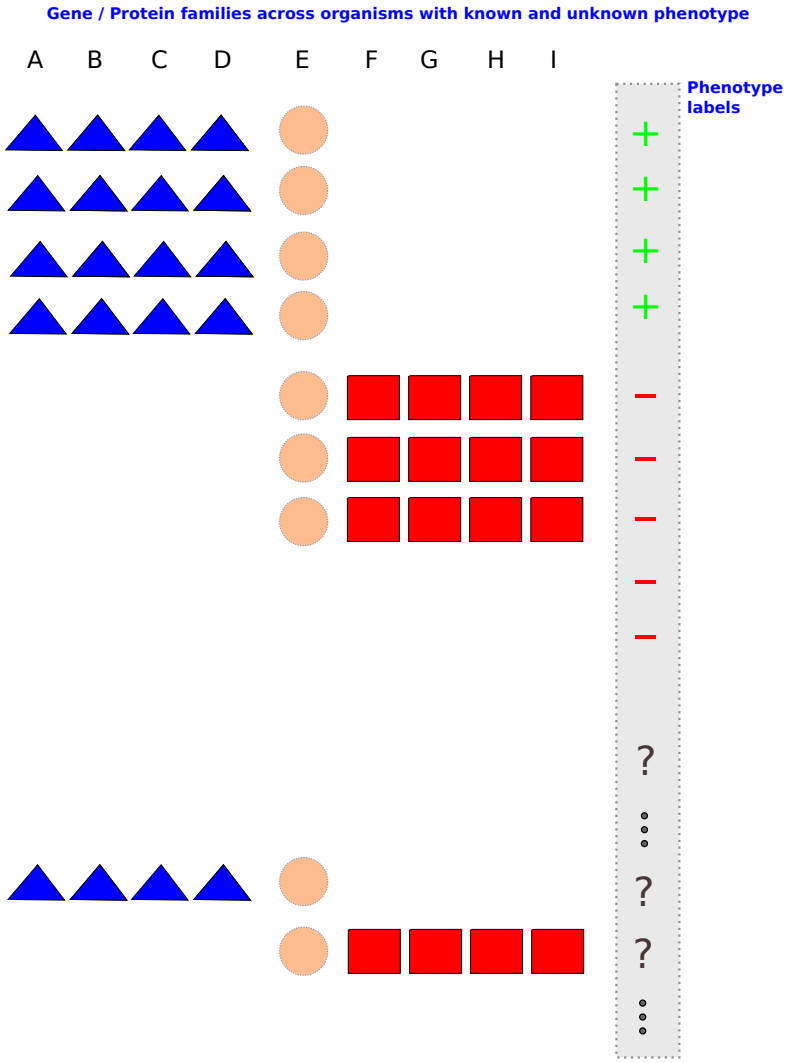


Figure 9.4: Gene co-occurrence patterns of moonlighting proteins. Moonlighting protein families (shown here in column E) can be active in different cellular processes (indicated by the blue and red functional modules) – possibly related with different phenotypic traits of the organism. In contrast to proteins with a single function, the phylogenetic profile of moonlighting proteins is generally more difficult to relate with the phenotypic labels of organisms, simply because the respective genes can occur in both phenotype-positive and phenotype-negative genomes. However, detecting their joint co-occurrence with genes from a phenotype-specific functional module would clearly help to associate them with the phenotype. This is the aim of the proposed seeding strategy.

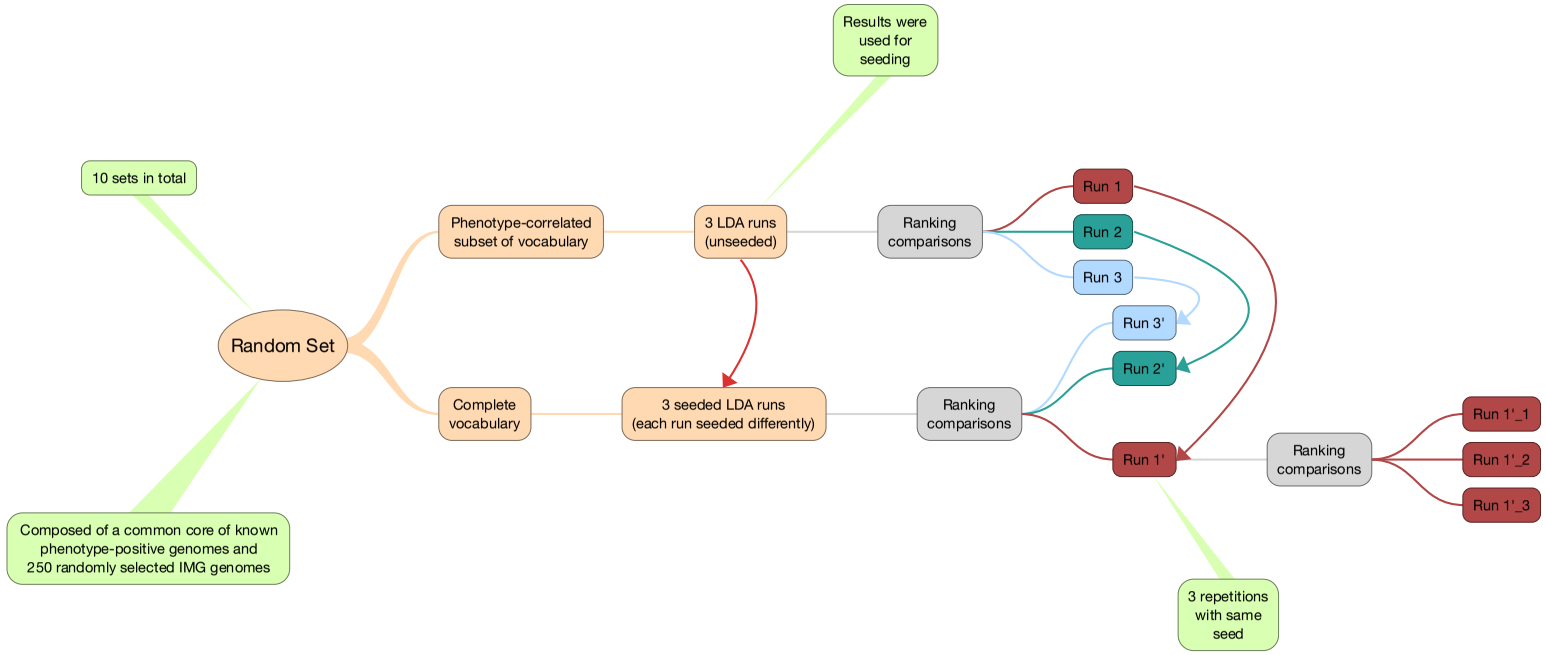


Figure 9.5: Tests of topic seeding on randomized input sets. The following procedure was repeated for ten different input sets with a variable fraction of genomes and a common core of phenotype-positive genomes. LDA was run three times on each set with the phenotype correlated vocabulary only, resulting in three independent topic models that were used for seeding in a successive phase. Each seed then initialized a corresponding consecutive run of LDA, performed on the same random set, but this time using the full vocabulary for creating the input collection. Moreover, identically seeded runs were also repeated, to further assess the stability of the results. Arrows in the picture describe how runs were seeded by other runs. The resulting modules of each LDA run were ranked based on the module weights in phenotype-positive and phenotype-negative genomes, respectively.

CHAPTER 10

Summary of Part III

As described, there is a need for computational methods to target the protein families and functional modules that are underlying processes associated with a certain phenotype. The types of methods can be subdivided into family-centric and pathway-centric methods. Family-centric methods consider protein families as basic attributes, whereas pathway-centric methods target functional modules and treat them as atomic functional units. In both cases, prioritizing the most decisive elements, that is, doing an attribute ranking is a straightforward solution to the problem.

SVM-based family-centric approach

In Weimann *et al.* (2013), we have proposed a family-centric method that determines the most relevant families from the entries of the weight vector of a support-vector-machine-based classifier for genomes and metagenomes. The method was shown to

successfully identify protein families that are distinctive for lignocellulose degrading species.

Support vector machines (SVMs) are powerful supervised classification models that achieve high levels of accuracy in classification tasks (Cortes and Vapnik, 1995). Given a set of labeled input vectors as the so-called training set, a SVM model can be optimized to distinguish between the different label classes. One can thus train SVMs to distinguish between phenotype-positive and phenotype-negative genomes from a suitable training set. A particular benefit of SVMs is the ability to generalize their predictive performance to new and unknown input samples, even if the training set is small.

SVMs are usually treated as 'black box models', that is, one is interested in the classification performance of the model but not in the delineation of the specific input attributes that were decisive for the classification outcomes. In Weimann *et al.* (2013), we chose the type and parameters of the SVM model in such a way that it became possible to identify the most relevant input attributes that determine the classification decision. We used this model to distinguish lignocellulose-degrading from non-degrading microbial organisms by processing their repertoires of protein families as input attributes. The resulting method achieved high values of classification accuracy, and correctly identified known protein families of lignocellulose degradation processes.

LDA-based pathway-centric approach

In Konietzny *et al.* (2014), we have extended the unsupervised framework for functional module inference proposed in Konietzny *et al.* (2011) by an attribute ranking scheme to enable the targeted discovery of phenotype-related modules. Thus, we have developed a hybrid framework that combines a LDA-based approach for unsupervised module inference with a supervised post-processing step based on attribute ranking. As attributes, we defined 'genome-specific weights' for the modules. The weights are used in a threshold-based classifier to distinguish phenotype-positive from phenotype-negative genomes, and the ranking procedure determines the most discriminative functional

modules from the model – based on the obtained classification results.

We used a mixed input collection of genomes and metagenome bins to infer the potential functional modules. Therefore, metagenomic information was explicitly taken into account, which is important because several of these microbial communities were known or assumed to possess lignocellulose-degrading abilities. In contrast to support vector machines, the described hybrid method provides the benefit of being able to process partially labeled datasets. The reason is that its first part – which is based on LDA – works in a completely unsupervised fashion, meaning that the labels of genomes are only used for the subsequent ranking of the modules. As a consequence, phenotype-related functional modules can be learned from both known and unknown phenotype-positive input samples. In Konietzny *et al.* (2014), we used 25 times more genomes and metagenome bins for the process of functional module inference than for the subsequent module ranking step. Learning from partially labeled datasets is especially useful to process metagenomes because, typically, we observe the phenotype of a microbial community as a whole. However, we are unable to assign phenotypic labels to individual metagenome bins, and thus cannot use them for creating a labeled training set of metagenomic data.

Again, we used microbial lignocellulose degradation as a candidate phenotype to test our method. Lignocellulose is a complex molecular structure, and thus many different subprocesses need to be involved in its decomposition. In line with that, we identified five large functional modules whose respective weight distributions within known degrading and non-degrading microbial species were significantly different. The modules were specific to lignocellulose-degraders, that is, they clearly represented attributes of the phenotype-positive class. We therefore named the five potential functional modules 'plant biomass degradation modules' (PDMs). A detailed study of the protein family contents of the PDMs eventually proved their individual functional coherence, and their involvement in different subprocesses of lignocellulose degradation, such as cellulose, hemicellulose, and pectin degradation.

We made sure that the PDMs represented a stable result across repeated LDA runs.

This was particularly important because, in the phenotype-targeting setting, we were exclusively interested in a marginal aspect of the overall topic model which in this case consisted of 400 inferred potential functional modules. Therefore, the stability of the model itself could not secure the stability of the few modules we were interested in. This aspect needed to be checked carefully.

Importantly, the new approach allows the discovery of unknown aspects of phenotype-related biological processes. For example, we identified interesting protein families (e.g. 'domains of unknown functions') that were implied to be functionally linked with already known phenotype-related families.

The PDMs mapped to several gene clusters in phenotype-positive genomes which represents additional evidence for functional relationships – based on the 'guilt by association principle' of gene neighborhood. In total, we identified more than 400 gene clusters distributed across the whole input set of genomes and metagenome bins with help of the modules, only some of which were described before.

Notably, mapping the PDMs to gene clusters is also useful for finding additional gene sequences that are linked to known lignocellulose families by genomic proximity. From the perspective of a genome researcher, this feature demonstrates a particular benefit of the method compared to a gene-centric method. Note that, in general, complex phenotypes of organisms can be decomposed into several functional sub-contexts, as e.g. the different processes of cellulose and pectin degradation in the case of lignocellulose decomposition. A gene-centric method would not distinguish between these different subaspects. It typically outputs an unstructured list of protein families which are correlated with the phenotype labels of the organisms. The resulting list can be quite long, and it might contain a substantial fraction of falsely correlated protein families. Therefore, using the whole list in a search for matching gene clusters is inefficient, and may lead to the detection of many irrelevant gene clusters. In contrast to the gene-centric paradigm, a module-centric approach infers groups of functionally coupled protein families that are jointly related to the phenotype of interest. Therefore, matching these small phenotype-related groups against (meta-)genomes for detecting potentially

interesting gene clusters is a far better strategy. It efficiently limits the search space of gene clusters, for example, with respect to more than 3,000 genomes and metagenome bins that we analyzed.

Part IV

Synopsis and Outlook

CHAPTER 11

Synopsis

This work was motivated by the need for new computational methods that enable large-scale analyses of the sets of protein families that are underlying biological processes in microbes (Section 1.3). Such methods would help to explain the relationships between the gene repertoires of cells and their observable phenotypic traits, which could finally lead to the discovery of novel proteins that are potentially useful for biotechnological or medical applications.

Moreover, since the majority of microbes cannot be cultivated with the currently known laboratory protocols, their protein sequences need to be accessed with metagenomic techniques. New methods for the analysis of functional modules of protein families should therefore be applicable to the gene repertoires of genomes of cultivated isolate species as well as of metagenomes of microbial communities.

This thesis comprises three related research projects that have addressed the above mentioned requirements. The previous chapters presented the individual results of the

projects and the newly developed methods in the form of the original publications. In the following, the main achievements and some of the results are summarized and discussed.

11.1 Main achievements

LDA for functional module inference

The main focus of the research was on the *de novo* discovery of functional modules of protein families. In Konietzny *et al.* (2011), we have pioneered the use of probabilistic topic models for the purpose of functional module inference (Chapter 5). We have demonstrated how potential functional modules can be inferred from the co-occurrence patterns of protein family identifiers with the latent Dirichlet allocation (LDA) model. Thus, we have adapted the LDA model from the field of text mining applications to the domain of phylogenetic profiling, that is, the inference of functional interaction partners based on genomic context.

We inferred a large set (200) of functional modules with diverse biochemical activities. The modules are useful resources contributing to our general knowledge of biological processes in microbes, and their back-mapping to genomes can be used to improve the process-level annotations of the analyzed genomes. Moreover, the modules were shown to represent individual functional contexts for the protein families that they contain. Based on the 'guilt by association principle', this can guide the assignments of functional roles to uncharacterized protein families, due to the assumed functional dependencies between the families of a functional module. LDA is an unsupervised scheme whose inference procedure does not require *a priori* knowledge about the composition of the inferred modules. Therefore, our approach is well suited for an exploratory analysis of the existing microbial processes, and enables the discovery of both (partially) known as well as novel functional modules. Due to the assumption of a global topic distribution in the LDA model, the modules likely correspond to common biological processes of the analyzed microbial genomes.

Validating the inferred functional modules by experiments in the laboratory would be time- and labor-intensive. This approach was not feasible for a large-scale evaluation of our inference method. We therefore sought a suitable gold standard to validate the predictions. However, the available datasets suffered from various insufficiencies which reflect the multitude and diversity of microbial processes in nature and the narrow boundaries of our knowledge (Section 3.6.2). In particular, it is very difficult to validate potential functional modules that do not map to known processes. A sound evaluation of the methods was thus a particular challenge in the research projects, and we proposed a mapping to high confidence edges in the STRING functional interaction network as a solution to this problem.

Identification of phenotype-related protein families and functional modules

The development of methods that can identify phenotype-related genomic elements is tightly related to the aforementioned work. With respect to biotechnological and medical applications, scientists are most often interested in elucidating the underlying mechanisms of specific cellular processes that correspond to a particular phenotype of microbial cells. We here consider phenotypes with a binary state that indicates the presence or absence of an organism's ability to perform certain biochemical activities. The task in this setting is to identify genomic elements whose occurrence patterns across genomes correlate with the phenotypic features of the organisms. In Weimann *et al.* (2013), and Konietzny *et al.* (2014), we have developed two new methods for this purpose, which both use supervised attribute ranking schemes (Section 8.3 and Section 8.4). The basic idea behind attribute ranking schemes is to weight each individual genomic element by its effectiveness to distinguish phenotype-positive from phenotype-negative genomes in a set of labeled training examples. The main difference between the two new methods lies in the genomic elements which are targeted as attributes, that is, either individual protein families or entire functional modules.

Once phenotype-related genomic elements have been identified, they can be used

to distinguish phenotype-positive from phenotype-negative genomes. We have demonstrated that our methods are capable to classify genomes with high accuracy, which highlights the relevance of the identified genomic elements. While it is difficult and labor-intensive to determine phenotypic traits of cells in wet lab experiments, our computational methods allow a rapid screening of large organismal libraries to identify reasonable candidates for further testing.

Phylogenetic profiling for metagenomic datasets

The new LDA-based approach for functional module inference builds on similar assumptions as phylogenetic profiling. However, previous phylogenetic profiling methods were only designed for the analysis of functional relationships between protein families in microbial genomes.

Whereas phylogenetic profiling works well on genomic datasets, its effectiveness for metagenomics greatly suffers from the uncertainties in the phylogenetic profiles that can be constructed on metagenomic datasets. Prior to the assembly step, metagenomic DNA (shotgun) sequences are extremely fragmented, and consist of tiny DNA pieces with just a few DNA nucleotides. Although post-processing of these datasets with specialized bioinformatics methods may recover large parts of the original gene inventories, the outcome will typically be incomplete and associated with uncertainty. In particular, uncertainties in the presence and absence profiles of genes across metagenomic datasets will directly affect the performance of phylogenetic profiling methods. These methods are highly sensitive to disturbed co-occurrence signals. With the new methods, we introduced a Bayesian framework for phylogenetic profiling. It is known that this type of methods is well suited for the analysis of noisy and less certain input data.

Moreover, the application of phylogenetic profiling to metagenomes is difficult because metagenomic datasets are huge in size and lack internal structure. The reason is that a metagenome by definition represents the pooled genomes of various individuals in a microbial community. This drawback of metagenomic datasets potentially also affects the effectiveness of LDA on input collections of metagenomes. We therefore proposed to

use predicted taxonomic binnings of metagenomes to define the input documents for LDA. This corresponds to splitting the metagenome of a microbial community into small bins of protein sequences which roughly approximate the genomes of the unknown community members, at least in terms of gene repertoires at the level of higher-order taxonomic groups for the most abundant species. We can then substitute the metagenome by its corresponding taxonomic bins in the input of LDA, which means substituting a single huge document by a collection of small documents. This preprocessing step increases the obtainable information from the co-occurrence patterns of protein families for the metagenome of a microbial community. It should be noted, however, that (even with the best currently available binning methods) only small fractions of the proteins of a metagenome can be assigned to taxonomic bins. We therefore cannot process the majority of protein sequences in a typical metagenome in this fashion. Nevertheless, with the described approach, we attempt to reconstruct the latent community structure, which is a necessary step for relating the inferred functional modules to individual members of the community. Indeed, we have successfully demonstrated that functional modules can be ascribed to individual metagenome bins. This is important because typically merely some unknown members of a microbial community possess the desired biochemical activities, whereas the measurable indicators of such activities can only be studied for the community as a whole.

11.2 Conclusions

We have demonstrated that functional modules can be inferred from (meta-)genomes, whose gene sequences were annotated with protein family identifiers, with a probabilistic topic model such as latent Dirichlet allocation. We discovered large quantities of potential modules with diverse functionalities and demonstrated that most of them captured biologically relevant functional interactions between protein families. The modules represent known and novel biological processes, and we have discussed some examples of modules representing protein complexes, signal transduction cascades and metabolic

processes. Moreover, we have developed two new methods for the targeted discovery of single protein families and entire functional modules that are key players in processes related to specific microbial phenotypes of particular relevance for biotechnological or medical applications. The discovered modules can be predicted for genomes and metagenomes to improve process-level annotations. In some cases, the modules may also guide protein-level functional annotations based on the guilt by association principle because they have put many weakly characterized families into a new functional context.

From another – slightly philosophical – perspective, this success of topic models on genomic input data seems to imply similar mechanisms for the human act of text writing and the way how evolution forms the gene repertoires of bacterial genomes. It is therefore appealing to think of evolution as 'an author that expresses the *topics of life*'. The analogy is obvious: while texts become shaped by the semantic concepts that an author bears in mind, functional modules have left characteristic patterns in the gene repertoires of organisms.

However, topic models were clearly not designed to model evolutionary processes, and the evolutionary forces that shape bacterial genomes are extremely complex. They include mechanisms like gene duplication, gene loss, and horizontal gene transfer, as well as more subtle processes like mutation events that slowly alter the functions of gene products. Moreover, we might have an incomplete picture of the true gene repertoires due to false positive or false negative predictions during gene detection and function prediction. All of these aspects greatly influence the joint conservation across genomes of the protein families that make up functional modules, which reduces the effectiveness of phylogenetic profiling approaches that assume evolutionary cohesiveness (Section 3.3).

Nevertheless, many of the described evolutionary processes finally result in discrete changes of gene frequencies in genomes, and this phenomenon matches the assumptions of topic models about word count frequencies in text documents. Therefore, topic models can, at least, capture certain effects of evolution, even though, of course, the depiction of evolution as an author should be seen as a metaphor.

The LDA-based approach relies on similar assumptions as phylogenetic profiling,

however, its group-oriented mode of operation has clear advantages when it comes to the detection of complex transitive relationships between the members of biological processes. As probabilistic frameworks, topic models are very flexible in the way how they model groups of functionally related protein families. Therefore, functional modules can also be inferred from partial (incomplete) observations, which again reflects their ability to detect transitive relationships between elements. This flexibility of the models is very useful with respect to the heterogeneous levels of evolutionary cohesiveness that is found for functional modules (Section 3.3).

Overall, our results support the assumption of the evolutionary cohesiveness of functional modules. In future analysis, one could analyze the impact of horizontal gene transfer on the distribution of modules across genomes.

CHAPTER 12

Outlook

The two main projects of this thesis have established and evaluated a new framework for inferring functional modules by means of probabilistic topic models (Section 5.2 and Section 8.4). As the projects were pioneering this novel use of probabilistic topic models, the main focus was on testing the assumption that topic models qualify for phylogenetic profiling. Therefore, it was crucial to prove that the discovered modules do indeed capture cellular processes and protein complexes. This task was especially difficult for potential modules that could not be mapped to known cellular functions. We focused on addressing this problem because – from an industrial perspective – modules containing weakly characterized protein families could be of special importance, possibly representing novel cellular processes or parts thereof.

This chapter outlines some ideas for future work which - based on the results of this thesis - may guide further studies and improvements of the methodologies.

Delineation of potential functional modules from the topic distributions

In both LDA-based projects, we used $C = 0.01$ as the default threshold for transforming the inferred topic distributions into potential functional modules. This specific choice was based on previous tests with different values, and visual inspections of the plotted topic distributions. In Konietzny *et al.* (2014), we also presented results obtained with an alternative threshold value ($C = 0.005$). This served as an example of a permissive threshold which resulted in significantly enlarged potential functional modules. Overall, many of our tests indicated that permissive thresholds also yield good results. But it should be noted that it becomes difficult to manually assess the functional cohesiveness of very large groups of protein families. The default threshold thus represents a compromise which resulted in functional modules of a reasonable average size, capturing both small and large functional modules.

In future work, one could build on the results of this thesis, and try to optimize the choice of the threshold. For example, one could rely on the STRING-based measure of coverage – introduced in Konietzny *et al.* (2011) – to systematically investigate how different choices of the C parameter would impact the average results of the modules. Likewise, monitoring the modules' capabilities to distinguish phenotype-positive from phenotype-negative genomes with the F-measure could be useful to compare different choices of C, as demonstrated in Konietzny *et al.* (2014). As a long-term goal it would be desirable to include the process of determining the optimal choice of C directly into the inference process.

Scaling to massive datasets

Topic models and Bayesian inference methods are efficient means to perform large-scale machine learning studies of the evolutionary patterns of functional modules across genomes. This performance feature of the models is important with respect to the available amounts of genomic and metagenomic datasets.

The field of (meta-)genomics has already been transformed into big data analytics. The main reason for this are high-throughput techniques like, for instance, massively parallel DNA sequencing. According to Moore's law, the capacity of data processing doubles every two years because the density of transistors on hardware chips can be doubled in this time period (Moore *et al.*, 1998). However, since 2008, the velocity of DNA sequence generation by high-throughput techniques even outpaced this rate by a factor of 4 (O'Driscoll *et al.*, 2013). Therefore, we can expect huge amounts of data in the near future that will allow large-scale analyses to reveal hidden patterns of cellular processes. For example, the National Cancer Institute (USA) announced to sequence a million human genomes over the next years in order to understand biological pathways and genomic variations (O'Driscoll *et al.*, 2013).

These expected amounts of data require scalable analysis methods and distributed compute infrastructures such as Apache Hadoop to be handled. The running time of LDA scales linearly with the input size as well as with the number of topics; however, the running time for our analyses (comprising about 3,000 genomes and metagenomic bins) was already in the range of days. Thus, with even bigger dataset sizes the time required might become very long, which is why there have been many attempts to speed up the algorithm by parallelized, distributed implementations (Newman *et al.*, 2009; Xiao and Stibor, 2010; Yan *et al.*, 2009; Zhai *et al.*, 2012). Moreover, it would be worth to try related deep neural network models such as *word2vec* (Mikolov *et al.*, 2013). Notably, deep learning methods have become very powerful tools in the field of big data analytics (Schmidhuber, 2015).

Advanced models

Besides LDA, we had tested several other probabilistic topic models (Chapter 6), and – in particular – supervised models (Chapter 9). Many of these have promising features that could, in theory, improve functional module inference. However, apart from *collocation* LDA, we had difficulties to show their superior performance over the baseline LDA

model. The scope and level of detail of available gold standards were not sufficient to prove a significant improvement of the results with the more advanced models. For instance, the exact protein families that are involved in lignocellulose degradation and their interplay are still unclear (Wilson, 2011), which complicates the comparison of model predictions that differ on a very fine-grained level (in just a few families). However, these intermediate results do not imply that the tested alternative models are not worth to be considered again in the future. The opposite is the case, as our knowledge of the principles for applying topic models to genomic input data has just begun to mature. The results from this thesis, combined with improved gold standards should guide future work. For example, one could use the plant biomass degradation modules discussed in Konietzny *et al.* (2014) as a reference set to test the performances of different models – especially, supervised ones. One could also extend our tests with artificial pathways (Subsection 9.1.2). We only tested a single and isolated pathway. More realistic tests could use a couple of pathways to simulate complex phenotypes such as lignocellulose degradation which consist of several subprocesses. Supervised models usually base the classification of documents on the vectors of topic weights. If there is only a single target pathway, its relevance for the topic weight vector could be too small to be noticed by the supervised model. Combinations of pathways would then correspond to multiple features in the vectors, and thus classification could become more performant and robust.

Resolution of the protein family space

With a few exceptions, the specificities of current HMM protein family models of databases like Pfam or CAZy are suboptimal. One example is the CAZy family GH5 which comprises a large and heterogeneous set of proteins, among which several should be subcategorized instead (Aspeborg *et al.*, 2012). Insufficient levels of details also become evident by looking at the protein domain architectures of the cellulolytic *cip-cel* and hemicellulolytic *xyl-doc* gene clusters described by Blouzard *et al.* (2010) (Konietzny

et al., 2014). The clusters are composed of biochemically diverse carbohydrate-active enzymes but the protein domain architectures of the genes capture this functional heterogeneity only on a very coarse level of detail. The rather coarse definitions of protein families thus represent a fundamental bottleneck for methods that operate on this kind of input data. A subcategorization of CAZy and Pfam families would therefore help to increase the resolution of the protein family space and likely result in more accurate, fine-grained functional modules when LDA is applied. We can certainly expect to see appropriate (iterative) updates of the protein family databases in the near future – for example, a subcategorization of carbohydrate-active families (Aspeborg *et al.*, 2012).

Inclusion of additional data sources

Bayesian graphical models are modular and flexible tools that can describe almost all aspects of a real world problem domain. Thus, topic models are extensible, and we could include random variables that model additional features of (meta-)genomes such as, for example, gene activity measurements from gene expression experiments. Such extensions of the model would allow to incorporate additional relevant data, and the formulation of conditional constraints. For instance, the protein families assigned into a common functional module could be restricted to possess similar activity patterns in controlled experiments. The main difficulty would be to adjust the update rules of Gibbs sampling for the extended models which is not trivial, and requires a profound expertise concerning statistical modeling.

Coupling with metabolic models

Trying to look ahead into the future, we can expect a continuous improvement of both the coverage and the quality of the contents in protein family databases. This will lead to more detailed genome annotations, and thus a better performance of phylogenetic profiling methods and topic models with respect to functional module inference. It will then become interesting to couple functional module inference with metabolic

simulations.

For example, we could simulate a flux balance model for a model organism with the objective to identify missing links in the metabolic pathways. Missing links suggest that some unknown elements could potentially lead to more efficient fluxes in the metabolic network. But the question is why these elements are missing? In many cases, the lack of elements turns out to be an artefact of genome annotation, meaning that existing proteins were not annotated correctly in the process of genome annotation. If we cross-link an observed gap in a metabolic pathway with the results of functional module inference, we might be able to identify proteins that are functionally coupled with the protein set of the pathway. These would represent candidates for filling the gap – even if they have missing or false annotations.

Guilt-by-association

So far, we assumed that functional module inference provides a mean to guide biologists in identifying gene or protein candidates that are involved in cellular processes. This assumption is motivated by the 'guilt-by-association principle' (Section 3.7). This thesis concentrated on the inference of functional modules. For future work, it would be interesting to use the functional context of a functional module to automatically propose biochemical functions for insufficiently characterized protein families that are contained in the module. As a starting point, I would propose the excellent review article by Janga *et al.* (2011), which describes several strategies for the propagation of functional annotations in functional interaction networks that could be adapted for this purpose.

Future perspectives

Functional module inference remains a difficult challenge because the variety of cellular processes, the complexity of their molecular mechanisms, and their evolutionary traces in genomes are very difficult to capture in statistical models. Many of the underlying mechanisms are still unknown, and history has shown that the delineation of model

pathways such as *glycolysis* required years of intensive biochemical studies. Nevertheless, with respect to the huge amounts of genetic measurements accumulating in databases, there seems to be enough information available for data-driven approaches. For the future, we can certainly expect to see an immense progress in the computational study of metabolic networks. Key factors for that development will be a steady improvement of the quality of protein- and process-level annotation resources, as well as the adoption of techniques from big data analytics.

Part V

Appendix

APPENDIX A

Bayesian inference

A Bayesian model is defined by a joint probability distribution over a set of random variables. In general, three different kinds of variables can be distinguished. *Observable* variables represent features of the model for which data is obtainable through measurements in experiments or other forms of data acquisition. *Latent* (or hidden) variables, by contrast, are used for the design of the model, but their status cannot be observed directly. Their function is to explain the values of the observable variables by modeling the assumed dependencies between the variables. Latent variables might closely correspond to real world entities, such as a disease state that explains the observable symptoms of a patient, or they might represent abstract, hypothetical concepts without an exact correspondence in the real world domain. The third type of variables are *hyper parameters* of the model. Their values can be specified at the time of model application to control the characteristics of the model. Thus, we could write the density function of the joint probability distribution of the model as:

$$p(\vec{o}, \vec{l}|\vec{h}) = p(o_1, \ldots, o_s, l_1, \ldots, l_r|h_1, \ldots, h_q), \tag{A.1}$$

where $\vec{o} = (o_1, \ldots, o_s)$, $\vec{l} = (l_1, \ldots, l_r)$ and $\vec{h} = (h_1, \ldots, h_q)$ are vectors summarizing the sets of observable variables, latent variables and hyper parameters, respectively.

Bayesian models can be defined for large sets of random variables, and be used to process high-dimensional input vectors containing values for the observable aspects of the model. Given the hyper parameters of the model, and a set of data for the observable variables, the task of Bayesian inference is to infer the posterior distribution for the latent variables of interest. According to Baye's rule, the density of the posterior distribution for a set of continuous variables, for example, the latent variables l_1 and l_2 can be calculated as follows:

$$p(l_1, l_2|\vec{o}, l_3, \ldots, l_r, \vec{h}) = \frac{p(\vec{o}, l_3, \ldots, l_r|l_1, l_2, \vec{h}) \cdot p(l_1, l_2)}{\int\limits_{l_1, l_2} p(\vec{o}, l_3, \ldots, l_r|l_1, l_2, \vec{h}) \cdot p(l_1, l_2)\, dl_1\, dl_2} \tag{A.2}$$

In high-dimensional vector spaces, solving this formula is difficult to achieve because calculating the high-dimensional integral for determining the normalization constant requires too much computational efforts. With Markov Chain Monte Carlo (MCMC) methods, the solutions to the integrals can be estimated efficiently.

Latent variables of the model, which are not of direct interest, can be marginalized out from the joint distribution. This requires the specification of a prior and an integration over the whole parameter space of the variables. For example, the latent variable l_3 can be marginalized out from the model as follows:

$$\begin{aligned} p(\vec{o}, l_1, l_2, l_4, \ldots, l_r|\vec{h}) &= \int\limits_{l_3} p(\vec{o}, \vec{l}|\vec{h})\, dl_3 \\ &= \int\limits_{l_3} p(\vec{o}, l_1, l_2, l_4, \ldots, l_r|l_3, \vec{h}) \cdot p(l_3)\, dl_3 \end{aligned} \tag{A.3}$$

In some cases, the integrals can be solved analytically; however, often this is not

feasible. Fortunately, the integrals that are needed for the marginalization process can be handled by the same Markov Chain Monte Carlo methods used for the approximation of the posterior probability.

APPENDIX B

The LDA model

This section provides a more detailed description of the LDA topic model, its generative process, and a procedure for model inference based on Gibbs sampling. For additional details on these topics, the reader is referred to an excellent technical report by Gregor Heinrich (Heinrich, 2009).

B.1 Model description

The LDA model describes a collection of M text documents that were generated from an underlying set of K semantic concepts, called the topics, which are modeled as probabilistic clusters over a vocabulary of words. Each word w in a document d of the collection represents an instance of a term v from the vocabulary V. A word can be generated from a specific topic according to a latent multinomial probability distribution. There only exists a single, global set of topics that shape the contents

of all documents, and the relative influence of a topic for an individual document is described by a document-specific prior probability for every topic.

Based on these definitions, the probability of observing a specific word can be modeled in terms of a mixture model, that is, a convex combination of a set of component distributions, where the mixture proportions sum to one. The LDA mixture model describes the relative influence of the topics on the word content of the document:

$$p(w = v) = \sum_{k=1}^{K} \underbrace{p(w = v|z = k)}_{\varphi_k^v} \cdot \underbrace{p(z = k)}_{\vartheta_d^k}, \tag{B.1}$$

where z is a random variable that is associated with the word w and describes the choice of topic. The formula expresses the probability of observing an instance of the vocabulary term $v \in V$ in a particular document. It represents the weighted sum over the likelihoods of the word for each of the K underlying topics, where the weight factors correspond to the prior probabilities of the topics for the document under consideration. As can be seen from the formula, a word may originate from several topics with different probabilities. Both types of probability distributions, the distribution over V for every topic, and the distribution describing the prior probabilities of the topics are modelled as multinomial distributions. For a given document d, and a topic k, these distributions are commonly referred to as $\vec{\vartheta}_d$ and $\vec{\varphi}_k$, respectively[1].

B.2 The generative process of the LDA model

The generative process for the construction of a text corpus can be described as an iterative procedure. The basic operation of this process is sampling from a Dirichlet-multinomial model.

[1] More precisely, the terms $\vec{\vartheta}_d$ and $\vec{\varphi}_k$ refer to parameter vectors that specify the respective multinomial distributions in the LDA model.

The Dirichlet-multinomial model

Dirichlet distribution

The Dirichlet distribution is a distribution over the parameter space of a multinomial distribution. This means that parameter vectors $\vec{p} = (p_1, \ldots, p_n)$ drawn from a Dirichlet can be used to define a multinomial distribution because they fulfill the condition: $\sum_j p_j = 1$, which moreover implies that the values p_j are embedded in a $(n-1)$-dimensional simplex in $\mathbb{R}^n$ (Figure B.1).

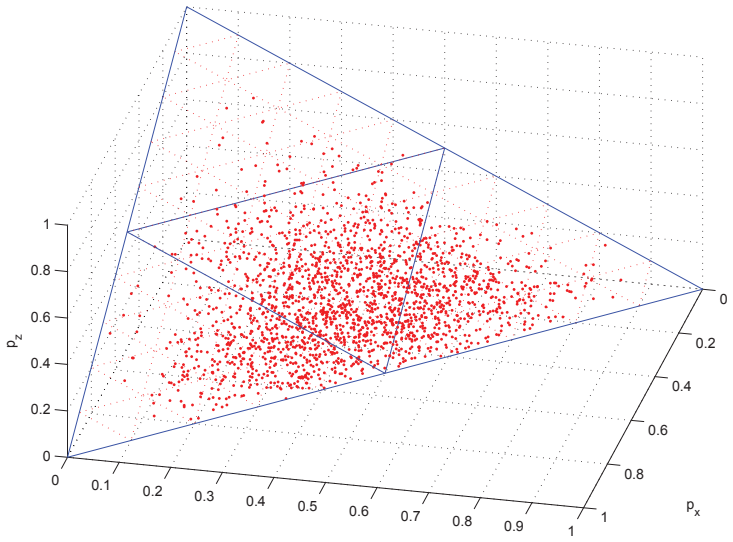


Figure B.1: 2000 samples from a Dirichlet distribution. The samples were drawn from the Dirichlet distribution $Dir(4, 4, 2)$. All samples are on a simplex embedded in the three-dimensional space, due to the constraint: $\sum_j p_j = 1$. (Image source: Heinrich (2009))

In the Dirichlet-multinomial model, the generative process for creating a word looks as follows:

$$(i) \quad \vec{p} \sim Dir(\vec{p}|\vec{\beta}) \tag{B.2}$$

$$(ii) \quad w \sim Mult(w|\vec{p}) \tag{B.3}$$

A parameter vector $\vec{p}$ is sampled from a Dirichlet distribution (i), and serves to

define a multinomial distribution which is used to sample the word w (ii). Note that the characteristics of the Dirichlet distribution can be controlled by the hyper parameter $\vec{\beta}$.

The task of Bayesian inference is to invert the process, that is, to generate the latent parameter $\vec{p}$ from a sequence of observed words. The Dirichlet-multinomial model uses the fact that the Dirichlet distribution is a conjugate prior for the multinomial distribution.

The generative model of LDA

The generative process assumed by the model of LDA is described by the iterative procedure in Figure B.2.

⊡ "topic plate"
for all topics $k \in [1, K]$ **do**
  sample mixture components $\vec{\varphi}_k \sim \text{Dir}(\vec{\beta})$
end for
⊡ "document plate":
for all documents $m \in [1, M]$ **do**
  sample mixture proportion $\vec{\vartheta}_m \sim \text{Dir}(\vec{\alpha})$
  sample document length $N_m \sim \text{Poiss}(\xi)$
  ⊡ "word plate":
  for all words $n \in [1, N_m]$ in document m **do**
    sample topic index $z_{m,n} \sim \text{Mult}(\vec{\vartheta}_m)$
    sample term for word $w_{m,n} \sim \text{Mult}(\vec{\varphi}_{z_{m,n}})$
  end for
end for

Figure B.2: The iterative process of the LDA model. See description in the main text. (Image source: Heinrich (2009))

A run through the generative process looks as follows. First, the parameters $\vec{\varphi}_k$ for K multinomial distributions, which correspond to the topics of the model, are sampled from a Dirichlet distribution that depends on the hyper parameter $\vec{\beta}$. Next, the collection of M documents is created in a loop. For every document m, the prior probabilities $\vec{\vartheta}_m$ of the topics are sampled from a Dirichlet distribution that depends on

the hyper parameter $\vec{\alpha}$. The size N_m of the current document is a parameter that can be sampled from, for example, a Poisson distribution. The size parameter then determines the number of iterations of the inner loop that creates the sequence of words according to a Dirichlet-multinomial model. In detail, for every word $w_{m,n}$, a random number between 1 and K is drawn from the multinomial distribution $Mult(\vec{\vartheta_m})$ that describes the prior probabilities of the topics. This number is assigned to the topic-indicator variable $z_{m,n}$. Depending on the chosen topic, the word then gets generated by sampling from the topic's corresponding multinomial distribution $Mult(\varphi_{\vec{z_{m,n}}})$.

The generative process can be visualized in a so-called plate notation diagram. Figure B.3 shows the graphical model of LDA in plate notation. Arrows represent dependencies between the variables and rectangles represent the repetition of the enclosed sampling steps. The numbers at the bottom of the rectangles specify the numbers of repetitions. The double circle around the variable for the words should indicate that these values are actually observed. The plate notation for LDA can be separated into the 'topic plate' at the left hand side, and the 'document plate' at the right hand side.

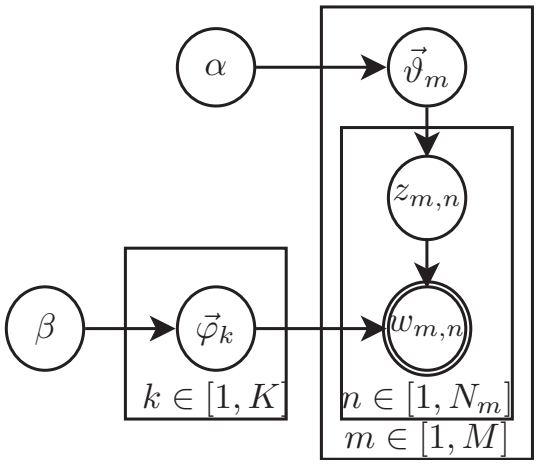


Figure B.3: Graphical model of LDA in plate notation. See description in the main text. (Image source: Heinrich (2009))

B.2.1 Model inference

The task of model inference is to estimate the latent variables of the LDA model from a given collection of M texts. In the beginning, we do not know the relationships between words and topics.

As described, every word w of the collection is associated with a random variable z, the topic-indicator. The indicator holds an index between 1 and K and describes the current assignment of the word to one of the K latent topics. In the beginning of the inference procedure, the correct assignments of words to topics are unknown and chosen arbitrarily. In principle, we would have to test all possible assignments for the whole document collection, which is infeasible due to the combinatorial complexity. Let $W = (w_{11}, w_{12}, \ldots, w_{1N_1}, \ldots, w_{M1}, w_{M2}, \ldots, w_{MN_M})$ be the concatenated vector of all words of the collection with total length $|W|$. The number of possible assignments of topics to words is then $K^{|W|}$, where K is again the number of topics. This factor is exponential in the input length and therefore we cannot compute it in a naïve fashion. However, with a Markov Chain Monte Carlo method such as Gibbs sampling, the optimal assignments can be approximated.

Collapsed Gibbs sampling

The conjugacy relationship between Dirichlet and multinomial distributions allows for a simplification of the LDA model. To this end, it is possible to marginalize out the latent parameter vectors $\vec{\varphi_k}$ for $k \in (1 \ldots K)$ and $\vec{\vartheta_m}$ for $m \in (1 \ldots M)$, which define the multinomial distributions, from the joint probability of the model. We then speak of a collapsed Gibbs sampler. The simplified model can be formulated in terms of count variables that express the frequencies of the assignments of words to topics across the whole document collection. Notably, these counts are sufficient statistics to capture the essential information content of the model. They can be summarized in two matrices, C_{VK} and C_{MK}, where an entry of the matrices describes the number of times that word $v \in V$ was assigned to topic k across the whole collection of M documents, or the number of times that words of document m were assigned to topic k, respectively.

With Gibbs sampling, we can simulate a Markov Chain that efficiently samples from the posterior distribution of the latent topic assignments. As a result, the values of the topic indicators get updated at every iteration of the Gibbs sampling procedure. Once the procedure has stabilized at later iterations, that is, the Markov chain has reached its stationary distribution, we can take 'snapshots' of the topic assignments to collect a Gibbs sample – for example, at every 10-*th* iteration. Finally, the parameters that determine the multinomial distributions of the LDA model can be approximated as statistics from the collected Gibbs sample. This step compensates for the fact that some of the latent parameters had been marginalized out in the construction of the model. Therefore, Gibbs sampling allows for the recovery of the unknown multinomial distributions from the words of a given text collection. The asymptotic complexity of the collapsed Gibbs sampler is $\mathcal{O}(K\,|W|\,S)$ (Porteous *et al.*, 2008; Xiao and Stibor, 2010). It is linear in the size of the input (the number $|W|$ of words in the collection), but it scales with the number K of topics, and the number S of Gibbs samples that are simulated for the Markov chain.

APPENDIX C

Automated seeding of topic models

The key idea of seeding is to semi-supervise the topics in the unsupervised LDA topic model, that is, we want to keep the associations of certain protein family instances to modules (topics) fixed. These fixations - or anchors - will serve to attract other proteins which often co-occur together with them. More specifically, the topic distributions of the LDA model become seeded based on the results of a former LDA run with the objective to stabilize the model, as well as to incorporate prior knowledge about intended groupings of protein families (Figure 9.3).

We developed a fully-automated two-step approach for deriving a seeded topic model that is optimized to predict phenotype-defining functional modules. The procedure minimizes the amount of manual curation needed to define suitable protein families as seeds of the modules.

C.1 Preprocessing step

Given the protein family repertoires of a labeled set of phenotype-positive and phenotype-negative genomes, we can use Pearson's correlation to assess the strengths of the individual associations of the families with the binary phenotypic labels. Protein families are treated in a binary fashion (existent or not) and coded as either 0 or 1 for each genome. Finally, we select all families with a significant correlation (p-value $\leq$ 0.05).

Another possibility would be to compute the correlation between protein families and the binary phenotypic labels across a taxonomic tree.

C.2 Two-step approach for the seeding of the topic model

The following procedure results in a seeded topic model that is set up to infer phenotype-related functional modules (Figure 9.3):

Step 1 Running LDA on a specifically chosen subset of the vocabulary

1. Limit the genomic input collection to the chosen subset of vocabulary terms.
2. Execute LDA on the collection.
3. Retrieve the final topic assignments for the term instances in the collection (that is, the triplets $(FD, d, z = t)$, where FD represents an annotated functional descriptor in document d, which has currently been assigned to topic t by the Gibbs sampling procedure.

Regarding *step 1.3*, some more explanations and remarks about the LDA implementation are necessary. We refer to an implementation based on Gibbs sampling, where each instance of a functional descriptor gets assigned to one of K topics (Figure 5.2). Let FD be a protein family identifier and (FD, d) be a corresponding annotation of a gene that is a family member of FD residing in the genome that is encoded as document d of the collection. Gibbs sampling iteratively assigns topic assignments to the term

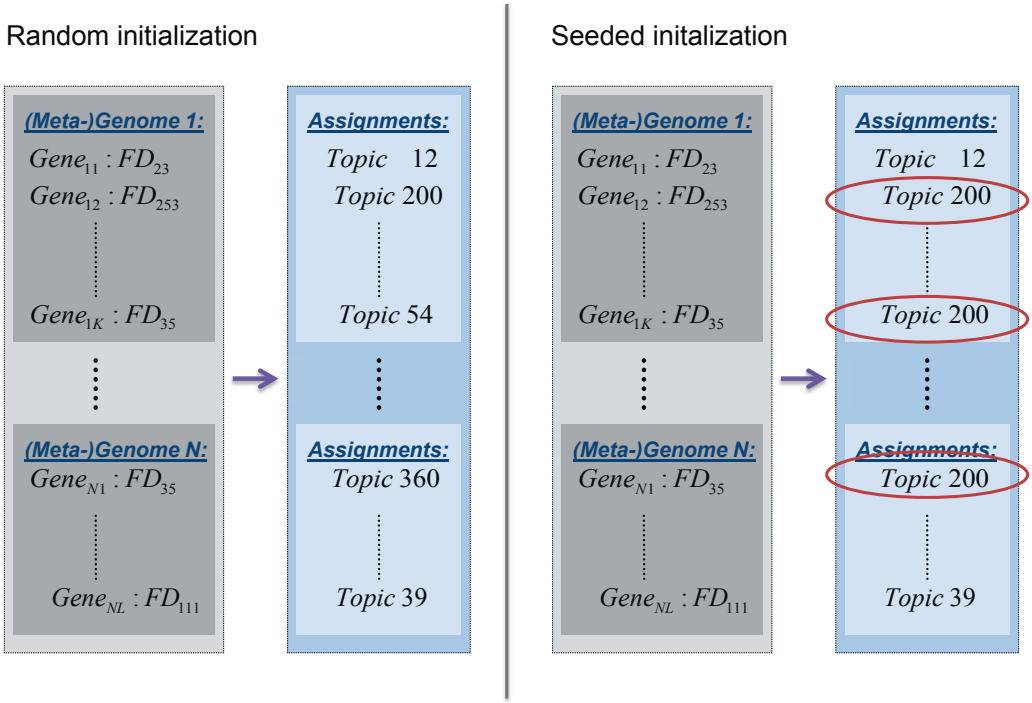


Figure C.1: Picture illustrating how seeding is performed. In Gibbs sampling, the instances of annotation terms are iteratively assigned to topic indices, while the method generally converges to an optimal assignment. At the beginning, all assignments are random. However, when seeding the model, we modify some of the random assignments with assignments from another run of LDA on the same or similar input set.

instances (Appendix B). Thus, we can query the information $z = t$ (where t is the topic assignment for the annotation instance (FD, d)) from a given Gibbs sample. We write this information as the triplet $(FD, d, z = t)$.

Step 2 Seeding a second run of LDA on the complete vocabulary

1. Choose a collection of documents that includes the exact term instances (FD, d) (i.e. identity of both annotation term and genome or bin) from the first step of the procedure. For example, use the same set of input documents as in step 1, but with the full vocabulary of functional descriptors. However,

one could also use the former set of documents and add additional ones.

2. Perform a single iteration of LDA to get a random topic assignment for all instances of annotated functional descriptors. Note that the number K of latent topics needs to be equal or exceed the number of topics in the initial run.
3. Adopt the topic assignments of step 1 of the procedure for all tuples (FD, d) that have been processed in the first step (note that in the second step both the vocabulary as well as the amount of documents may differ). That is, substitute $(FD, d, z = t')$ by $(FD, d, z = t'')$, where t' is the current random topic assignment and t'' is the final assignment for the combination (FD, d) from step 1 of this procedure.
4. Re-start the Gibbs sampling with the constrained model.

The modified topic assignments are used in the internal Gibbs sampling procedure and control the inferred associations of terms to topics. Importantly, we have three options of handling the modified assignments in the final Gibbs sampling process:

1. We keep them fixed over all iterations.
2. We use them only as the initial state of the model. From some chosen iteration on, the model will be left free to alter the assignments in future iterations.
3. We keep part of the assignments fixed – for example, the most relevant ones – giving the model freedom to change the assignments that we think are less certain with respect to our problem at hand.

Depending on the choice of these three options, we need to modify the Gibbs sampling procedure in order to skip the updates of topic assignments for modified triplets $(FD, d, z = t'')$ that should be kept fixed.

Seeding in such a way corresponds to semi-supervised learning. Importantly, we do not globally fix the assignments of functional descriptors to functional modules. Instead of this, we only do this for individual instances of the functional descriptors.

There might still be other instances of the same functional descriptors that are not constrained by these assignments. This feature allows instances of a protein family to be ambiguously assigned to functional modules. Regarding the multitude of variants of real biological modules and their possible overlaps, it would be a bad idea to force global associations between protein families and functional modules, independent of the genome in which an instance of the family occurs.

APPENDIX D

Supplementary files

This chapter includes the additional files of the published articles (text files and some figures), as well as additional information resources for topics discussed in this work. For the sake of compactness of this document, comprehensive supplements containing many pages or large datasets are not printed here (except supplementary notes and methods), however, they are accessible from the enclosed CD-ROM medium.

All of the files are also available online (see journal web links in the published articles).

D.1 Publication - Konietzny *et al.* 2011

D.1.1 Supplementary Note

The following section includes the Supplemenatary Note accompanying Konietzny *et al.* (2011).

– SUPPLEMENTARY NOTE –

Inferring functional modules of protein families with probabilistic topic models

Sebastian GA Konietzny, Laura Dietz and Alice C McHardy

Analysis of results with alternative settings for the number of topics inferred with our method

We matched the stable modules from each experiment against the KEGG reference pathway database to compute profiles of pathways that posses at least 6 hits to any one of the modules. For *k*=200 and *k*=400 this yielded most diverse profiles with 20 different KEGG pathways. However, a single KEGG pathway may be mapped by several of the identified stable modules, and we observed higher numbers of occurrences of the matched KEGG pathways for *k*=400 than for *k*=200. Moreover, the KEGG profile over the complete set of 400 inferred modules is more diverse than for *k*=200.

The stable modules which are discussed in more detail, i.e. 'ChemoTax', 'Flagell' and 'VitB12', could be identified in all runs for *k* = 200, 300, 400 and 500. For *k* = 100, especially the 'VitB12' module was unstable and could not be identified in two of three runs. The 'Ribosome' related module wasn't found for *k* = 100 at all.

In total, 256 OGs belong to the sets of identified stable topics across all tested settings of *k*. The most abundant COG functional categories are general function prediction only (R), amino acid transport and metabolism (E) as well as signal transduction mechanisms (T) and carbohydrate transport and metabolism (G). This set of 256 re-occurring OGs makes up 72.8% (k=100), 26.9% (k=200), 19.8% (k=300), 15.2% (k=400) und 15.1% (k=500) of the OGs associated with the corresponding sets of stable modules.

We observed a few modules to be more stable across runs for higher values of *k*. For example, the presented 'Ribosome' related module could be tracked across nine runs for *k*=400, but it was less stable across runs with k=*200*. In general, the set of stable modules for *k*=400 and *k*=500 is larger than for runs with smaller *k*. Moreover, the averaged coverage values are surprisingly high for larger *k,* though more and larger modules are evaluated, which also comprise a larger set of distinct OGs. Thus, the modules capture a larger interaction network for higher numbers of inferred topics.

In summary, we suggest to use a large number for *k,* i.e. $k \geq 200$. However, the more modules are being inferred, the more difficult it becomes to extract the relevant modules from the set, especially if one is interested in finding new groupings of OGs that do not map to known pathways, such as KEGG pathways. In these cases one can use the coverage values as guidance to finding interesting, functionally coherent groupings, as shown for the 'ChemoTax' module in the presented study. Also, enrichments of COG functional categories are useful. An alternative is to query the set of modules for a list of OGs of interest.

Distribution of the probability weights of the modules across the analyzed genomes

Every genome defines a probability distribution, *P(PF-Module|Genome)*, that describes for each of the *k* inferred modules the probability to be encoded in the respective genome (Methods). The figure in Additional file **9** visualizes the distributions of probability weights for every genome. The majority of genomes features only a small subset of modules with a high probability, instead of giving equal probability to all modules. We also observed that some PF-modules receive high weights for almost all of the genomes. Analysis of the corresponding topics revealed that these correspond to empty modules, meaning that OGs with highest probability in the topic distribution do not exceed the threshold criterion of *C=0.01*. This means that none of the OGs of the vocabulary are specifically associated with these modules, and thus it is likely that they serve somehow as 'default topics' which associate with OGs without significant co-occurrence signals.

For the 'Flagell' module we used the module's probability weights to select a subset of genomes where this module seems to be encoded. As cutoff threshold for selecting potentially 'Flagell' encoding genomes we chose the mean of the weights of the 'Flagell' module across all genomes. Thus, the set of genomes was divided into 228 potentially 'motile' and 347 'nonmotile' genomes. We then acquired phenotype annotations for the genomes from the GOLD-Genomes Online Database [2]. In total, 302 genomes were thus labeled as 'truly motile', whereas 215 genomes were flagged as 'truly nonmotile'. With this setting, our method achieved a recall of 65.2% and a very high precision of 94.3%. As most of the genomes with a predicted 'motility' are indeed motile, this provides further evidence for the relation of the inferred module to 'motility' of the organism. The weaker value of recall could be explained by the fact that within some of the genomes only a smaller subset of the OGs related to the module were annotated, resulting in a decreased probability weight of the module for these genomes.

Choosing hyperparameters α and β

We tested performance of the Latent Dirichlet Allocation model for different settings of the hyperparameters α and β. In many applications of LDA, settings

$$\alpha = 1/k \qquad \text{and} \qquad \beta = \frac{1}{size(vocabulary)}$$

were used. However, we obtained significantly worse results with this setting. Parameter values used in our study, which yielded best results, are the default settings of the 'topic modeling toolbox' [1], which is an alternative implementation of the LDA model.

References:

1. Steyvers M, Griffiths T (Authors): **Topic Modeling Toolbox 1.3.2** [http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm]

2. Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC: **The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata**. *Nucleic acids research (*2008) 36, no. Database issue (January): D475-9.

D.1.2 Overview of additional files

Additional file 1: A list of 198 potential functional modules. The Supplementary Tables S1-198 show 198 potential functional modules that were identified in a randomly chosen, exemplary run of the presented method ($k = 200$). Tables S1-70 represent the subset of particularly stable modules that could be tracked consistently across nine independent runs of the method. [Included on CD-ROM, not printed]

Additional file 2: Comparison of histograms over COG functional categories. Comparison of two histograms over COG functional categories for (A) 70 stable modules and (B) modules that could not be tracked across all nine runs. (Figure D.1 and Figure D.2, respectively)

Additional file 3: Profile of KEGG pathways with at least six matches to one of the 198 modules. List of KEGG pathways with at least six matches of their KO terms to one of the 198 potential functional modules inferred in the exemplary run with $k = 200$. (Figure D.3)

Additional file 4: Visualized matches to the KEGG pathway 'Porphyrin and chlorophyll metabolism'. KO terms that are matched by the OGs of the respective potential functional module are highlighted in pink. (Figure D.4)

Additional file 5: 'Ribosome'-related functional module. OGs of the 'Ribosome'-related functional module that was identified in nine runs with $k = 400$. (Figure D.5 and Figure D.6, respectively)

Additional file 6: Visualized matches to the KEGG pathway 'Ribosome'. KO terms that are matched by the OGs of the respective potential

functional module are highlighted in pink. (Figure D.7)

Additional file 7: Supplementary Note. This document includes additional details of the comparison of results for different settings of k, and a discussion on the distribution of the probability weights of the modules across the analyzed genomes. (Subsection D.1.1)

Additional file 8: Visualization of the functional network spanned by the OG pairs of the reference set. Pairwise functional interactions are defined by the reference set as edges between OGs in a network graph. The subset of verified pairwise predictions from the modules is shown in green, whereas the subset of verified predictions by pairwise co-occurrence profiling is shown in blue. Functional interactions that are predicted by both methods are colored red, and those not detected by any of the methods are shown in gray. (Figure D.8)

Additional file 9: Visualization of the distribution of probability weights of the modules across the analyzed genomes. The unclustered heatmap indicates the strengths of probability weights of the 198 modules across the genomes. Rows represent the genomes, whereas columns represent the weights of the modules. The brighter the color of a cell, the larger is the probability weight for the respective PF-module. We re-scaled the values of each row, using minimum and maximum values, to fit values to the interval [0,1]. A discussion of this heatmap is part of the Supplementary Note in Subsection D.1.1. (Figure D.9)

Additional file 10: Reference set of high confidence pairwise OG interactions. List of high confidence interactions with evidence support values from the individual STRING channels, and modified combined

scores. [Included on CD-ROM, not printed]

Additional file contents

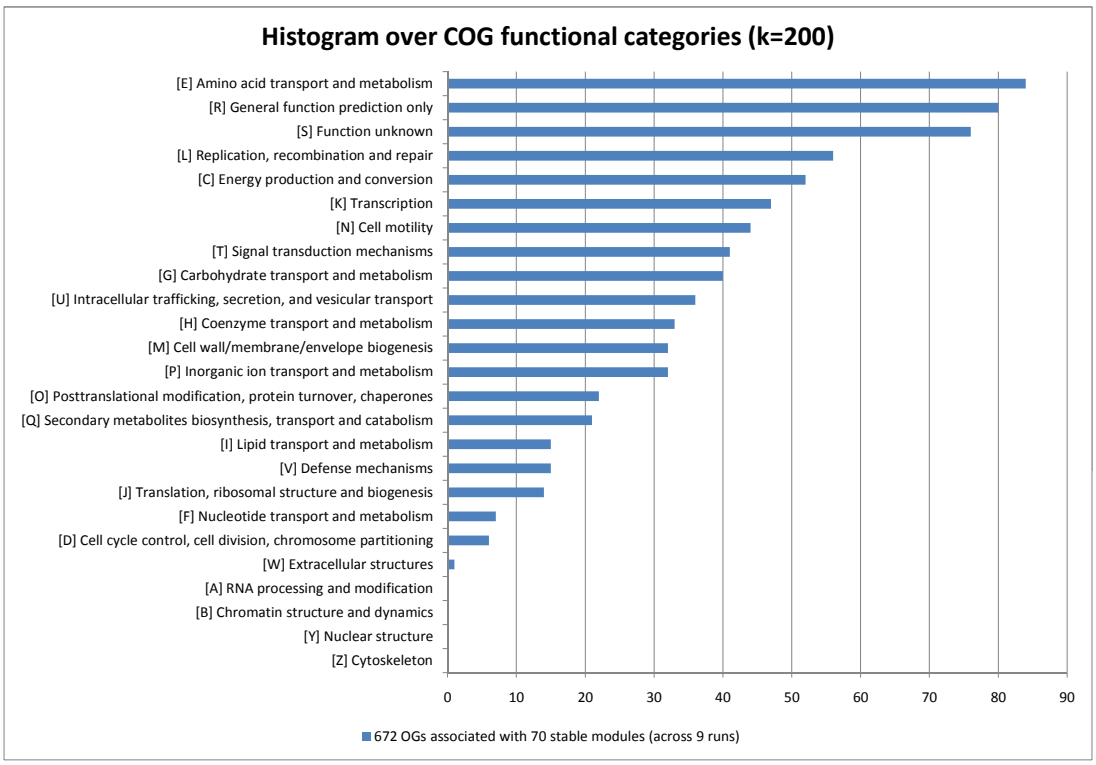


Figure D.1: Comparison of histograms over COG functional categories. (A) Over 70 stable modules

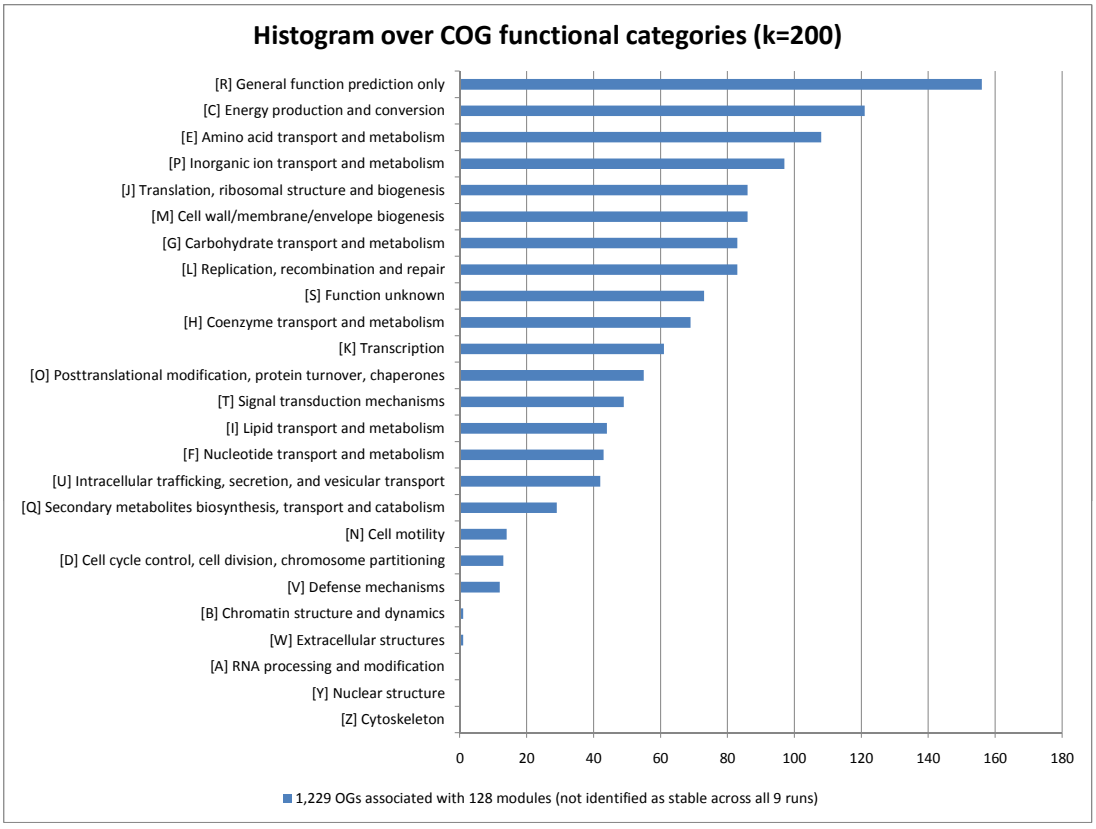


Figure D.2: Comparison of histograms over COG functional categories. (B) Over modules that could not be tracked across all nine runs.

KEGG-Name	Modules
ABC transporters	12
Two-component system	10
Porphyrin and chlorophyll metabolism	3
Oxidative phosphorylation	2
Flagellar assembly	1
Bacterial secretion system	3
Purine metabolism	3
Amino sugar and nucleotide sugar metabolism	3
Ribosome	1
Phosphotransferase system (PTS)	2
Phenylalanine, tyrosine and tryptophan biosynthesis	1
Aminoacyl-tRNA biosynthesis	1
Arginine and proline metabolism	1
Phenylalanine metabolism	1
Fructose and mannose metabolism	1
Riboflavin metabolism	1
Propanoate metabolism	1
Peptidoglycan biosynthesis	1
Benzoate degradation via CoA ligation	1
Starch and sucrose metabolism	1
Galactose metabolism	1
Pyruvate metabolism	1
Butanoate metabolism	1

Figure D.3: Profile of KEGG pathways with at least six matches to one of the 198 modules. List of KEGG pathways with at least six matches of their KO terms to one of the 198 potential functional modules inferred in the exemplary run with $k = 200$.

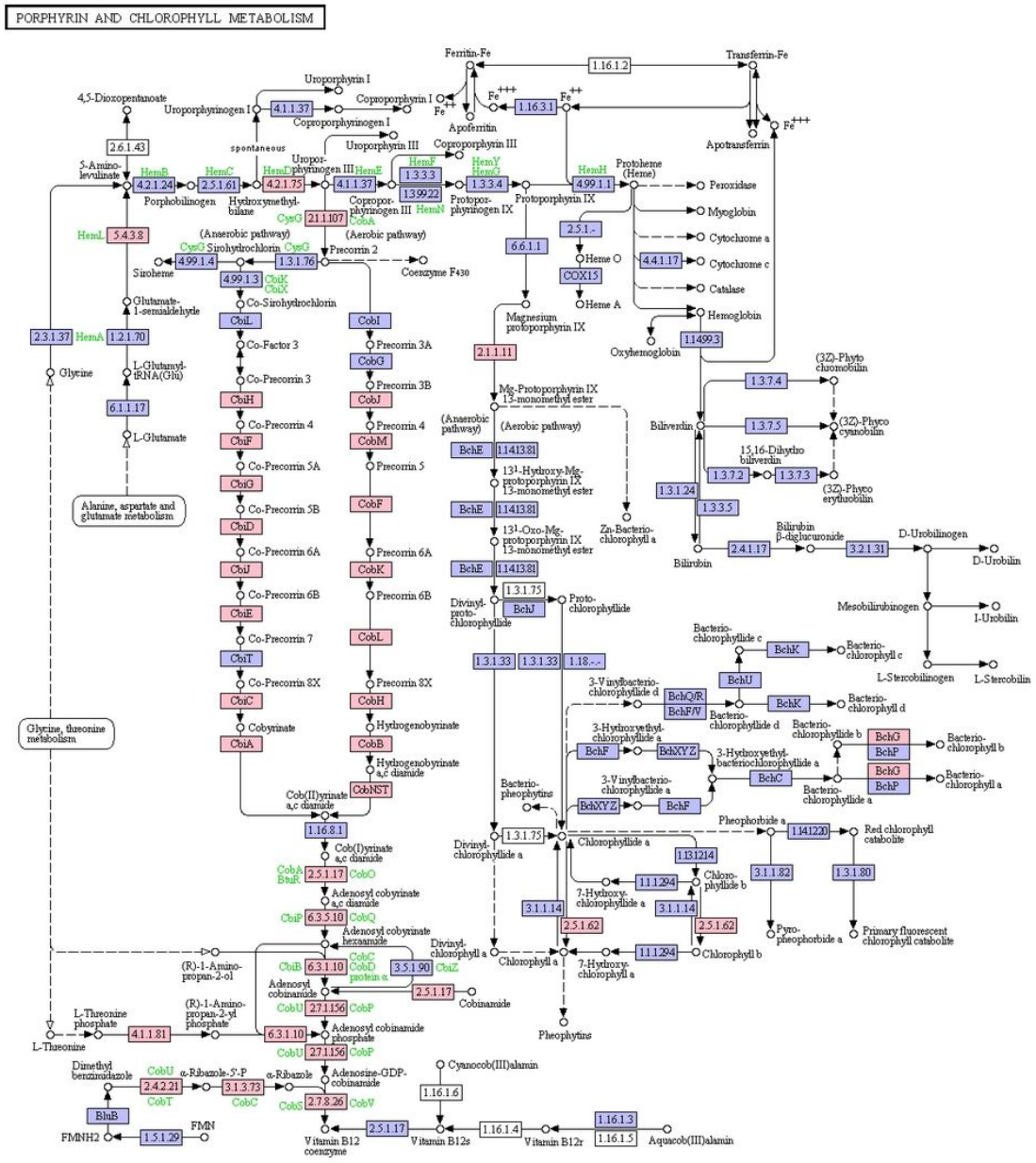


Figure D.4: Visualized matches to the KEGG pathway 'Porphyrin and chlorophyll metabolism'. KO terms that are matched by the OGs of the respective potential functional module are highlighted in pink.

Table S 1: Module 'Ribosome' (stable across many runs with k=200; stable across all nine runs with k=400. Displayed table is taken from runs with k=400.).

Rank	Probability	OG	Description
1*	0.018	COG0051 [J]	Ribosomal protein S10
2*	0.018	COG0100 [J]	Ribosomal protein S11
3*	0.017	COG0013 [J]	Alanyl-tRNA synthetase
4*	0.017	COG0124 [J]	Histidyl-tRNA synthetase
5*	0.017	COG0103 [J]	Ribosomal protein S9
6*	0.017	COG0541 [U]	Signal recognition particle GTPase
7*	0.017	COG0244 [J]	Ribosomal protein L10
8*	0.017	COG0250 [K]	Transcription antiterminator
9*	0.017	COG0522 [J]	Ribosomal protein S4 and related proteins
10*	0.016	COG0048 [J]	Ribosomal protein S12
11*	0.016	COG0162 [J]	Tyrosyl-tRNA synthetase
12*	0.016	COG0087 [J]	Ribosomal protein L3
13*	0.016	COG0199 [J]	Ribosomal protein S14
14*	0.016	COG0088 [J]	Ribosomal protein L4
15*	0.015	COG0143 [J]	Methionyl-tRNA synthetase
16*	0.015	COG0089 [J]	Ribosomal protein L23
17*	0.015	COG0256 [J]	Ribosomal protein L18
18*	0.015	COG0528 [F]	Uridylate kinase
19*	0.015	COG0060 [J]	Isoleucyl-tRNA synthetase
20*	0.015	COG0092 [J]	Ribosomal protein S3
21*	0.015	COG0201 [U]	Preprotein translocase subunit SecY
22*	0.015	COG0231 [J]	Translation elongation factor P (EF-P)/translation initiation factor 5A (eIF-5A)
23*	0.014	COG0049 [J]	Ribosomal protein S7
24*	0.014	COG0202 [K]	DNA-directed RNA polymerase, alpha subunit/40 kD subunit
25*	0.014	COG0358 [L]	DNA primase (bacterial type)
26*	0.014	COG0441 [J]	Threonyl-tRNA synthetase
27*	0.014	COG0495 [J]	Leucyl-tRNA synthetase
28*	0.014	COG0099 [J]	Ribosomal protein S13
29*	0.014	COG0016 [J]	Phenylalanyl-tRNA synthetase alpha subunit
30*	0.014	COG0185 [J]	Ribosomal protein S19
31*	0.014	COG0084 [L]	Mg-dependent DNase
32*	0.014	COG0080 [J]	Ribosomal protein L11
33*	0.014	COG0126 [G]	3-phosphoglycerate kinase
34*	0.014	COG0361 [J]	Translation initiation factor 1 (IF-1)
35*	0.014	COG0090 [J]	Ribosomal protein L2

Figure D.5: 'Ribosome'-related functional module. (Part 1 of 2) OGs of the 'Ribosome'-related functional module that was identified in nine runs with $k = 400$.

Rank	Probability	OG	Description
36*	0.013	COG0091 [J]	Ribosomal protein L22
37*	0.013	COG0180 [J]	Tryptophanyl-tRNA synthetase
38*	0.013	COG0552 [U]	Signal recognition particle GTPase
39	0.013	COG0221 [C]	Inorganic pyrophosphatase
40*	0.013	COG0081 [J]	Ribosomal protein L1
41*	0.013	COG0258 [L]	5-3 exonuclease (including N-terminal domain of PolI)
42*	0.013	COG0468 [L]	RecA/RadA recombinase
43	0.013	COG0533 [O]	Metal-dependent proteases with possible chaperone activity
44*	0.013	COG0575 [I]	CDP-diglyceride synthetase
45*	0.013	COG0030 [J]	Dimethyladenosine transferase (rRNA methylation)
46*	0.012	COG0525 [J]	Valyl-tRNA synthetase
47*	0.012	COG0096 [J]	Ribosomal protein S8
48*	0.012	COG0072 [J]	Phenylalanyl-tRNA synthetase beta subunit
49*	0.012	COG0172 [J]	Seryl-tRNA synthetase
50*	0.012	COG0052 [J]	Ribosomal protein S2
51*	0.012	COG0024 [J]	Methionine aminopeptidase
52*	0.012	COG0442 [J]	Prolyl-tRNA synthetase
53*	0.012	COG0198 [J]	Ribosomal protein L24
54*	0.011	COG0186 [J]	Ribosomal protein S17
55*	0.011	COG0550 [L]	Topoisomerase IA
56*	0.011	COG0149 [G]	Triosephosphate isomerase
57*	0.011	COG0592 [L]	DNA polymerase sliding clamp subunit (PCNA homolog)
58*	0.011	COG0094 [J]	Ribosomal protein L5
59*	0.011	COG0184 [J]	Ribosomal protein S15P/S13E
60*	0.011	COG0112 [E]	Glycine/serine hydroxymethyltransferase
61*	0.011	COG0097 [J]	Ribosomal protein L6P/L9E
62*	0.011	COG0102 [J]	Ribosomal protein L13
63*	0.011	COG0480 [J]	Translation elongation factors (GTPases)
64*	0.010	COG0200 [J]	Ribosomal protein L15
65*	0.010	COG0093 [J]	Ribosomal protein L14
66*	0.010	COG0532 [J]	Translation initiation factor 2 (IF-2; GTPase)

Figure D.6: 'Ribosome'-related functional module. (Part 2 of 2) OGs of the 'Ribosome'-related functional module that was identified in nine runs with $k = 400$.

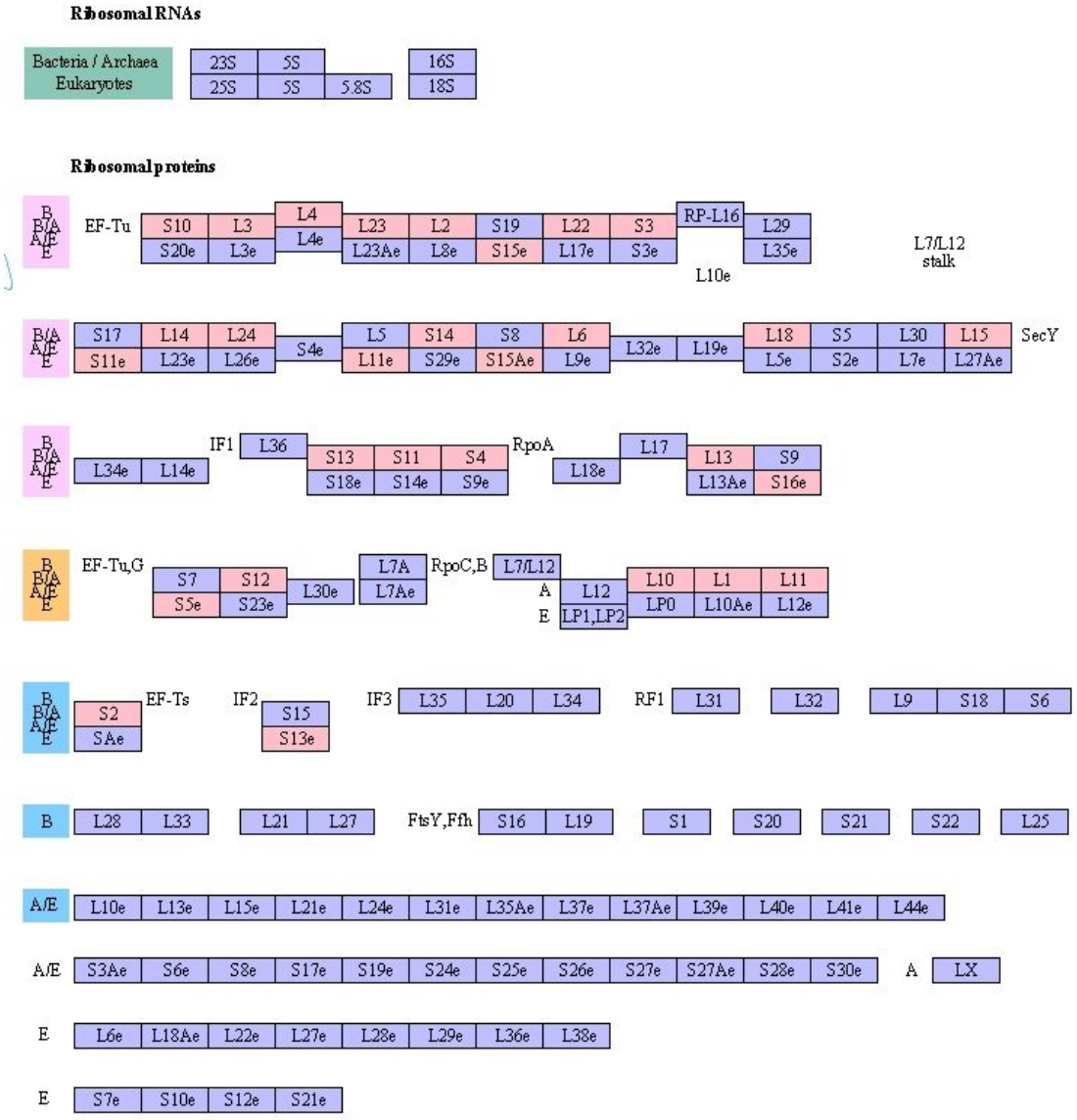


Figure D.7: Visualized matches to the KEGG pathway 'Ribosome'. KO terms that are matched by the OGs of the respective potential functional module are highlighted in pink.

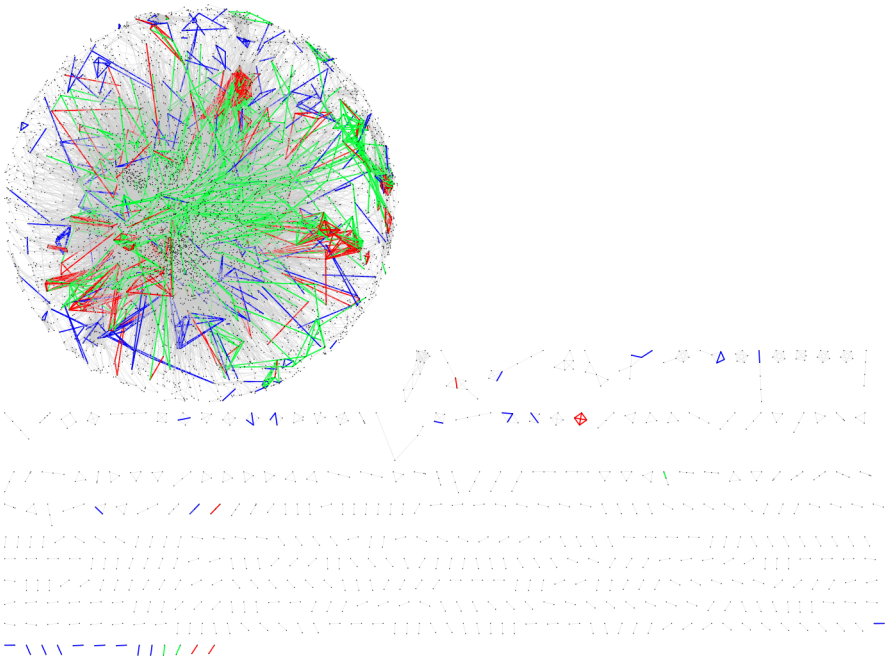


Figure D.8: Visualization of the functional network spanned by the OG pairs of the reference set. Pairwise functional interactions are defined by the reference set as edges between OGs in a network graph. The subset of verified pairwise predictions from the modules is shown in green, whereas the subset of verified predictions by pairwise co-occurrence profiling is shown in blue. Functional interactions that are predicted by both methods are colored red, and those not detected by any of the methods are shown in gray.

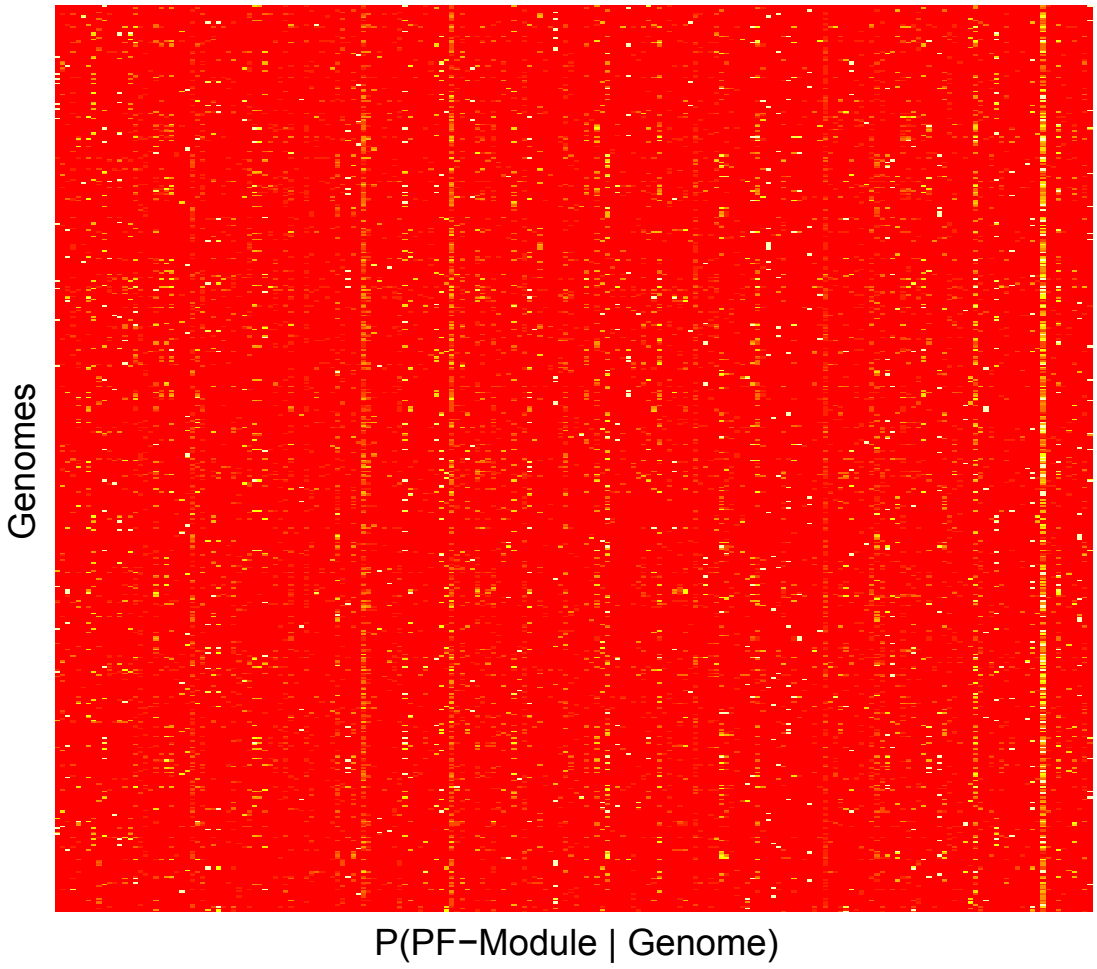


Figure D.9: Visualization of the distribution of probability weights of the modules across the analyzed genomes. The unclustered heatmap indicates the strengths of probability weights of the 198 modules across the genomes. Rows represent the genomes, whereas columns represent the weights of the modules. The brighter the color of a cell, the larger is the probability weight for the respective PF-module. We re-scaled the values of each row, using minimum and maximum values, to fit values to the interval [0,1]. A discussion of this heatmap is part of the Supplementary note in Subsection D.1.1.

D.2 Publication - Weimann *et al.* 2013

D.2.1 Overview of additional files

Additional file 1: Table S1 Isolate strains and metagenome samples used in this study. The signs '+' and '-' indicate availability of CAZy or Pfam annotation data. The symbol * marks strains for which we provide another reference than the genome publication characterizing the metabolic capacities of the respective strain. [Included on CD-ROM, not printed]

Additional file 2: Table S2 Evaluation and meta-parameter settings of the ensembles of classifiers. The ensembles were used for feature selection and phenotype classification of the (draft) genomes and metagenomes. The macro-accuracy for each model for a discrete set of values for the parameter C was calculated in cross-validation experiments. The five best models were selected based on macro-accuracy. The mean of the exponentially transformed parameter C and the mean macro- accuracy for these five models are shown for all trained classifiers. For details on the different ensemble classifiers, see the Results section in the manuscript. (Figure D.10)

Additional file contents

	Mean parameter *C*	Mean macro-Accuracy
$eSVM_{bPFAM}$	$10^{-1.7}$	0.93
$eSVM_{fPFAM}$	$10^{-1.5}$	0.87
$eSVM_{CAZY\_A}$	$10^{-1.0}$	0.95
$eSVM_{CAZY\_B}$	$10^{-1.9}$	0.95
$eSVM_{CAZY\_C}$	$10^{-1.9}$	0.94
$eSVM_{CAZY\_a}$	$10^{-1.1}$	0.93
$eSVM_{CAZY\_b}$	$10^{-1.6}$	0.94
$eSVM_{CAZY\_c}$	$10^{-1.8}$	0.92

Figure D.10: Table S2 Evaluation and meta-parameter settings of the ensembles of classifiers. The ensembles were used for feature selection and phenotype classification of the (draft) genomes and metagenomes. The macro-accuracy for each model for a discrete set of values for the parameter C was calculated in cross-validation experiments. The five best models were selected based on macro-accuracy. The mean of the exponentially transformed parameter C and the mean macro-accuracy for these five models are shown for all trained classifiers. For details on the different ensemble classifiers, see the Results section in the manuscript.

D.3 Publication - Konietzny *et al.* 2014

D.3.1 Supplementary Note

The following section includes the Supplemenatary Note accompanying Konietzny *et al.* (2014).

Inference of phenotype-defining functional modules of protein families for microbial plant biomass degraders

Sebastian G.A. Konietzny, Phillip B. Pope, Aaron Weimann, Alice C. McHardy

– Supplementary Note –

1. Pfam families with potential relevance for plant biomass degradation

Excluding most of the protein families with references in the CAZy database, the consensus modules M1 to M5 contained at least 20 Pfam families with rather unclear relationships to plant biomass degradation (see Table 3 of the main text). Examples of these are two 'domains of unknown function', DUF303 and DUF4008. Notably, DUF303 was a part of the hemicellulose-targeting gene cluster of *Fibrobacter succinogenes* (see section 6 below), and DUF4008 of module M5 seemed related to the CipA cellulosome scaffoldin protein of *Clostridium thermocellum*, as indicated by close genomic neighborhood (see discussion of PDM M5 in the main text). Further examples are 'fibronectin domains' (PF00041, PF14310), which are often observed as linker elements in modular cellulolytic enzymes [1], and a 'GDSL-like lipase/acylhydrolase' family (PF00657). The latter is an example of a family which is hard to assess at first sight, but is described as being involved in lignin degradation processes of white rot fungi [2]. Moreover, we discussed the 'ricin-type β-trefoil lectin' domains of module M2 in the main text, which could have functions for xylan-binding.

Many additional interesting Pfam and CAZy terms were associated with the PDMs, because they occurred in up to 8 corresponding modules of the 18 LDA runs (see the

numbers of their occurrences in the tables S1B-5B of Additional file 1). However, they were excluded from consideration due to the rather conservative construction of the consensus modules, which required a minimum of 9 occurrences. Despite the fact that these families occurred less frequently, they were clearly associated with the modules and represent further resources for the discovery of as yet unknown relationships among families. Note that in addition to the DUFs already contained in the PDMs, we identified 50 more DUFs that co-occur with module families within the identified gene clusters (18% of these were found in the 81 gene clusters of the phenotype-positive genomes).

2. Misclassified genomes of the learning set

Among the few genomes that were jointly misclassified by the modules M1 and M2 in our cross-validation experiments were *Bryantella formatexigens* (FN), *Xylanimonas cellulosilytica* (FN), *Thermonospora curvata* 43183 (FN) and *Actinosynnema mirum* (FP).

One reason for the misclassification of the first two species might be that both genomes were lacking annotations for GH9 (PF00759) (among other elements of M1). This cellulase family was described as an important component of the M1-associated *cip-cel* operon (see Figure 3A of the main text) and has essential cellulolytic activities in several organisms of our phenotype-positive set (see section 3 below). Based on this observation, it is likely that the reported activity of β-glucosidase in *B. formatexigens* [3] is mediated by families other than GH9, e.g. GH1, GH3, GH4 or GH5. Similarly, important elements of M2 – i.e. GH16, CBM13, CBM35, CBM61 and CBM47 – were missing in the annotations of the FN genomes.

In contrast to this, ‘misclassification’ of the other two species may be seen as a proof of concept. *A. mirum* and *T. curvata* were presumably correctly classified by the modules, whereas the phenotype characterization of the species in the original studies seems to have been wrong. According to a recent review on Actinobacteria [4], evidence of high

cellulase activity in *A. mirum* has been found, whereas *T. curvata* 43183 shows no cellulolytic activity and thus was probably characterized incorrectly in previous literature.

3. A discussion of the GH9 family and its role in lignocellulose-degrading species

The GH9 family is a large family of endoglucanases, some of which have been shown to act as processive cellulases on crystalline cellulose [5]. Despite the family's distribution across aerobic and anaerobic bacteria from a broad range of taxa, we found GH9 family annotations in only 313 isolate genomes and 22 metagenome bins, i.e. ~10% of our input collection. Remarkably, the GH9 family was annotated for 32 of the 38 phenotype-positive genomes, whereas it was almost absent in the phenotype-negative set, where it occurred in only 4 out of 82 genomes (see Additional file 7: Figures S1 and S2 (heat maps)). Thus GH9 was very specific for the known degraders.

The GH9 family has a key role in many cellulosomes [6], in which it is assumed to act in concert with other cellulosomal enzymes [7]. However, it has also been demonstrated to have cellulolytic activity in organisms of our phenotype-positive set that do not possess cellulosomes, such as *Caldicellulosiruptor bescii* [8] and *Thermobifida fusca* [9]. Furthermore, the cellulolytic activity of GH9 has been demonstrated for two more strains of species in our phenotype-positive set, *Clostridium phytofermentans* [10] and *Ruminoccus albus* 8 [11].

As described in the main text, GH9 occurred together with GH5 and CBM4 in the *cip-cel* operon of *Clostridium cellulolyticum* H10, for which it is known that the contained genes are involved in cellulose degradation [12]. Even more direct interactions of GH9 and GH5 families have been described. GH9 has been observed to be functionally linked with GH5 in a gene (Cel9B/Man5A) of *C. bescii* that encodes a highly thermophilic cellulase with potential application in biofuel production [8].

Concerning the reasons for the assignment of GH9 to module M1, we measured strong pairwise Pearson correlation of GH9 with the other elements across the learning set,

and identified gene clusters of GH9 and other M1 elements, such as GH10, GH43, CBM35, PF00756, PF02927, PF02018 and PF13472. Notably, in accordance with the known role of GH9 in cellulosomes, the GH9 family was also found to be associated with the cellulosome-related module M5 (the family was present in 7 out of 16 modules).

4. Characterization of the PDM M4

Many of the M4 protein families, such as GH2, GH3, GH5 and GH43, represent large protein families with a variety of possible functions. Members of GH3, GH5 and GH43 possess hemicellulose degradation activities (see Table 2 of the main text). Interestingly, the M4 module contained two different groups of structurally related protein families. These were β-galactosidase families (EC 3.2.1.23), like GH2, GH35 and GH42, and the three known members of the GH-D clan (a superfamily of α-galactosidases), i.e. GH27 (PF02065), GH31 and GH36. The GH31 family represents the most recent addition to this clan, based on structural and mechanistic similarities to GH27 and GH36 [13]. GH-D families have been studied mostly in the context of the degradation of galactoglucomannans (hemicelluloses) [13]. However, GH31 is also characterized as an α-xylosidase (EC 3.2.1.177) that catalyzes the hydrolysis of terminal xyloside residues at the extreme reducing end of xyloglucan-oligosaccharides.

To gain further insights into the possible functions of the M4 families, we investigated the gene clusters that were predicted by M4 in known lignocellulose degraders. Evidence for a link of M4 to xyloglucan degradation was found in the form of a long gene cluster of 8 genes (BACCELL_02066, BACCELL_02068–69 and BACCELL_02071–75) in the lignocellulose degrader *Bacteroides cellulosilyticus* DSM 14838, which was identified based on the protein family content of module M4. According to the gene annotations in IMG, the cluster comprises genes for 1 β-galactosidase, 2 α-glucosidases, 1 melibiase and 4 β-xylosidases. Three genes of the cluster were linked to the MetaCyc pathway of xyloglucan degradation based on their assigned EC numbers. In particular, EC number 3.2.1.23 of the GH2 family, which was annotated for the gene BACCELL_02066, mapped to a ‘xyloglucan oligosaccharide β-

galactosidase' enzyme of the MetaCyc pathway. The M4 families that mapped to the gene cluster are: GH2, GH5, GH31, GH32, GH36, GH43 and PF02065 (melibiase/GH27).

We then searched for additional evidence of a common link between β-galactosidases and members of the GH-D clan. Larsbrink *et al*. recently proposed a model for xyloglucan utilization by *Cellvibrio japonicus* based on functional evidence [14], which comprises β-galactosidases, a member of the GH31 family and an endo-xyloglucanase enzyme (EC 3.2.1.151). In accordance with this, the GH5 family, which was also included in M4, contains members with endo-xyloglucanase activities [15]. Thus there is evidence for the participation of M4 families in xyloglucan degradation. It is also interesting to see the clustering of two groups of structurally related protein families in M4 (i.e. β-galactosidases and the GH-D clan), both of which have members with known activities in xyloglucan degradation.

5. Low abundance of GH6 and GH48 in our dataset

The families GH6 and GH48 play important roles in cellulose hydrolysis, but they are not universally present in known lignocellulose degraders. Their absence is known for *Fibrobacter succinogenes*, *Cytophaga hutchinsonii* and various gut/rumen metagenomes with lignocellulose-degrading capabilities [16-19]. As further examples, an absence of GH48 has been described for *Cellvibrio japonicus* [20] and *Saccharophagus degradans* [21]. According to our in-house CAZy/Pfam annotation sets, only 11/38 (29%) and 24/38 (63%) of the phenotype-positive genomes in our learning set contained GH6 or GH48, respectively.

We observed sparse co-occurrence signals for these two families. Both families were annotated for less than 5% of the 3,216 input documents. Only one protein-coding sequence of the analyzed metagenome bins was predicted to contain GH6 by our HMM-based annotation pipeline (none for GH48), i.e. both families were largely absent from the annotations of the metagenome bins. It is of course possible that some members of these families remained undetected if they had only remote sequence

homology to the family models of the dbCAN and Pfam databases that we used. Taupp *et al.* have described how GH6 could not be detected in the Tammar wallaby foregut metagenome by sequence analysis; however, it was later found in functional screens of fosmid libraries from the same source material [22]. Interestingly, our own annotation sets revealed a GH6-annotated gene sequence in a Clostridia-related bin of the Tammar wallaby forestomach microbiome.

Given these sparse and weak signals of GH6 and GH48, the LDA model performed well in placing the two GH families into a functional context, at least in the case of the GH48 family. Both families exhibited considerable associations (topic probabilities ≥0.005) to some of the inferred functional modules. Although the strength of these associations was less than that required by our rather strict threshold value (C = 0.01) for converting the topic probabilities to potential functional modules, further analyses showed that GH6 and GH48 were always placed among the top 50 protein families of their respective modules (typically at positions 30–35 in the list of families sorted by decreasing probability). We concluded that the sparseness of the co-occurrence signals was the main reason why both families had low probabilities in our topic model.

The modules with associations to GH6 did not match our stability criteria and we therefore did not report a consensus module for them. Apparently, LDA could not learn a stable clustering for GH6. In contrast to this, GH48 was weakly associated with the modules used for creating the M5 consensus module. M5 grouped cohesin and dockerin elements together and seemed to be related to the structural components of the cellulosome complex. A weak association to this particular module can be explained by the fact that many of the known bacterial cellulosomes contain proteins of the GH48 family [23]. However, members of the family GH48 also occur in organisms with multifunctional (*Caldicellulosiruptor saccharolyticus*) and free cellulolytic enzymes [24], e.g. in *Thermobifida fusca*. Therefore, GH48 is by no means exclusive to cellulosome complexes and so one should not expect a much stronger association with M5.

Concerning the role of GH48 in cellulosomes, the family was shown to be non-essential for the cellulose degradation abilities of *Clostridium thermocellum*. Olson *et al*. have created a Cel48S/Cel48Y double knockout-mutant of a *C. thermocellum* strain that

maintained its ability for solubilization of crystalline cellulose, albeit with a reduced rate compared to the wild type [25].

GH6 and GH48 did not correlate well with the elements of module M1 based on pairwise Pearson correlations, so it makes sense that they were not assigned into this module (see Additional file 7: Figures S1 and S2 (heat maps)).

6. Predicted occurrences of M1 in *Fibrobacter succinogenes* S85 and *Thermobifida fusca* YX

The protein families of M1 were organized in a large gene cluster in the genome of *Fibrobacter succinogenes* S85, which is centered on the gene FSU_2269 (note that in the ATCC 19169 strain that we analyzed, this corresponds to the gene Fisuc_1769). The cluster mainly contained xylanases composed of CBM6, CBM35, GH43 and DUF303 protein families, and has previously been described by Yoshida *et al*. [26, 27] (see Additional file 6: Figure S1). *F. succinogenes* is assumed to use an as yet uncharacterized degradation paradigm [28]. In contrast to this, *Thermobifida fusca* is one of the model organisms for the free enzyme strategy [29]. Almost 80% of the protein families in M1 were annotated in *T. fusca* YX, such that the organism was predicted to encode module M1 by our method; CBM4, CBM6, PF02018, GH5, GH9, GH10 and GH43 were among these. Although we found only very short gene clusters in this genome, this observation demonstrates how the module M1 spans organisms with different degradation paradigms. It is therefore likely that in organisms using the free enzyme strategy, such as *T. fusca,* the elements of M1 act alongside elements of other modules. As an example, we identified functional modules that were rich in 'type-II secretion system'-related Pfam families in many LDA runs, which, in principle, might be responsible for the secretion of cellulolytic enzymes. Type-II secretion systems for cellulases have been described previously, e.g. in the plant pathogen *Erwinia chrysanthemi* [30]. However, for *T. fusca* in particular*,* doubts about a type-II secretion process for secreting cellulolytic enzymes have been raised, based on missing homologs [29].

7. Predictions of the cellulosome-related PDM M5

In some species, as shown for the cellulosome-related gene clusters *cip-cel* and *xyl-doc* in *Clostridium cellulolyticum* H10, the protein families of M1 and M5 mixed together. It is a good result that LDA split the protein families involved into different modules, despite the co-occurrence patterns induced by such mixed gene clusters. However, the M5 module only covered part of the structural components of cellulosomes (cohesin and dockerin) and thus was a weak predictor for the cellulosome paradigm. For example, M5 was assigned to 4 out of 8 *Caldicellulosiruptor* species, though these species are assumed to employ the free enzyme strategy [31]. It has been described recently that cohesins and dockerins appear in many bacteria that have no cellulosomes, where they seem to mediate diverse functions [32]. Although this might explain some of the observations where M5 was predicted for non-cellulosomal organisms, it is not an explanation of why 4 of the 8 *Caldicellulosiruptor* species fulfilled the weight threshold condition for M5, as these organisms had neither cohesin nor dockerin annotations. Instead, they possessed annotations for other elements of M5, such as CBM3 (PF00942), CBM36, PF07591 (Pretoxin HINT domain), PF13186 (DUF4008), PF05593 (RHS repeat) and PF07238 (PilZ domain) (see Figure 4 of the main text; see Additional file 5: Table S5). These elements were grouped into M5 because they often co-occurred with the cohesin/dockerin families across the input collection (though not in gene clusters in general). Some of the predicted occurrences of M5 in non-cellulosomal organisms were due to the module's low completeness threshold (38.46%), which was the lowest for all PDMs, meaning that just a fraction of the member families of M5 needed to be present in a genome for a positive prediction. A similar case could be the M5 prediction for *Sorangium cellulosum*, because there were no cohesin or dockerin domains present in the genome. Finally, we would like to mention that unravelling the lignocellulolytic capabilities of *Caldicellulosiruptor* species is a topic of recent research, as these are thermophilic bacteria that have the potential to improve industrial biofuel production [8, 33].

8. Predicted occurrences of PDMs in metagenome bins

PDMs were mainly identified for the taxonomic bins of the orders Clostridiales and Bacteroidales (see Figure 5 of the main text). Species of these clades, together with species of the order Fibrobacter, have cellulolytic activities in microbial communities in the rumen and large intestine of mammalian species [34]. Naas *et al.* have described cellulose degradation enzymes in Bacteroidales-related genome assemblages reconstructed from the rumen microbiome of cows [35], which agrees well with the predictions of M1 to M4 for the Bacteroidales bins of this metagenome (see Table 5 of the main text). Predictions of the PDMs for Bacteroidales-affiliated bins of the foregut microbial community from the Tammar wallaby [36] and the reindeer rumen [19] could thus indicate similar capabilities in species from this order. Overall, only two metagenome bins in our dataset, from the termite hindgut metagenome and related fosmid sequences, represented the order Fibrobacter. These bins were annotated with only a few CAZy families, including GH9 and GH2, but not GH3, GH5, GH6, GH10, GH26, GH30, GH43, GH44 or GH48, which may be why there were no PDMs identified for these bins.

Interestingly, M1 and M4 were predicted to occur in the Treponema bins of the termite hindgut metagenome and the corresponding fosmid sequences. Species of this genus are involved in the degradation of cellulose and hemicellulose in the termite hindgut community [18]. Treponema species seem to be quite diverse in this respect, as no PDMs were identified for the six isolated *Treponema* species of our dataset. *Treponema bryantii*, which was not included in our dataset, is known to promote the degradation activities of cellulolytic species like *Fibrobacter succinogenes* but is itself likely non-cellulolytic [37, 38].

M4 and M5 were predicted for two Euryarchaeota (Archaea) bins of terephtalate-degrading microbial communities from a bioreactor. So far, only very few archaeal species that are capable of degrading lignocellulosic biomass are known; however, the extremophiles in particular have great potential to improve industrial biofuel production [39]. One of the predicted Euryarchaeota bins had an annotation for the cohesin family, while the other bin encoded 84% of the M4 families, including GH2, GH3 and GH5. For

this phylum, three isolate genomes were predicted by the PDMs. The genomes of *Halorhabdus utahensis* (M1, M2) and *Haloterrigena turkmenica* (M3, M4) both have large repertoires of glycoside hydrolases, and, in case of *H. turkmenica,* for pectin degradation as well [40]. *H. utahensis* grows on xylan and it was recently discovered that it possesses an active GH5 cellulase gene, with possible application for biofuel production due to its tolerance of extreme conditions during the pretreatment of biomass [40].

9. Predicted polysaccharide utilization loci (PULs) and Sus-like PUL systems

We identified two gene clusters (the genes BT4145–50 and BT4152–58) predicted by the pectin module M3 for *Bacteroides thetaiotaomicron*, which closely correspond to two genomic regions (BT4145–51 and BT4152–55, BT4158) that were characterized as being active in rhamnogalacturonan degradation in a PUL-targeted study [41]. Moreover, it is well known that some PULs contain SusD- and TonB-like membrane proteins. The Sus gene cluster was originally characterized as a membrane-bound degradation system for starch [42], but this hypothesis has since been generalized to include PULs targeting other polysaccharides and cellulose (Sus-like systems [35, 43, 44]). Our method did not assign the SusD or TonB protein families to any one of the highly ranked modules. However, in most of the 18 LDA runs, it consistently grouped these elements into potential functional modules of lower ranks which also incorporated some protein families that are involved in starch binding or known to be located at the outer membrane. The modules were stable and we included their consensus module (PUL module) in Additional file 1 (Table S6A). PULs are known to be involved in the degradation of various polysaccharides [44], particularly starch, which explains why the PUL module was not identified as being highly specific for lignocellulose degraders and thus obtained lower ranks in our module rankings. Because the PUL module contained elements of Sus-like systems but no glycoside hydrolases, we combined the protein families of this module with those of modules M1 and M2, and used this pooled family set to search for associated gene clusters in the combined prediction sets of the three modules. We found almost 300 gene clusters of four or more genes including *susD*

(PF07980), a TonB-dependent receptor domain (PF00593) and one or more genes annotated with the GH5, GH9, GH10 or GH43 family. Among these clusters, we identified a few Sus-like PULs from *Bacteroides ovatus* that have previously been characterized [41]. The PUL cluster BACOVA_02649–56 contained SusD, TonB, GH5, GH9 and GH43 annotations, and corresponds to a genomic region (BACOVA_02644–56) which has been characterized as targeting xyloglucan [41]. Similarly, SusD, TonB, GH10, GH30 and GH43 families were annotated in another predicted cluster (BACOVA_03424–33) that is embedded within a slightly longer genomic region, BACOVA_3421–36, which has been characterized as targeting oat spelt xylan and wheat arabinoxylan [41]. These well characterized clusters represent only a few examples from the overall ~300 gene clusters.

10. Gene clusters identified in the cow rumen bin AGa

The 15 draft genomes from the cow rumen metagenome were partially fragmented and not fully assembled [45]. The PDMs mapped to six gene clusters with four or more genes, and several shorter clusters in the draft genomes. We found an interesting large cluster on a 97,191-bp contig of the Bacteroidales-associated draft genome ‘AGa' (see Additional file 12: Figure S1), which was partly matched by the protein families of M1. The gene cluster (NODE_457020_ORF_01660 to NODE_457020_ORF_01710; protein sequences in Additional file 13) includes three cellulases, based on the assignments of the GH5 family, and a cellobiose phosphorylase (GH94; EC 2.4.1.20) with an attached putative CBM (PF06204). The GH94 family was not assigned to the consensus module of M1 but it was contained in the M1 modules in 7 out of 18 LDA runs. Depending on the presence or absence of GH94 in the M1 modules of different runs, the gene cluster was identified either partly or completely. The cluster was flanked by genes annotated with GH3 (gene 01650) and GH5 (01630) on the left-hand side of the cluster, and Pfam annotations that do not appear to be associated with lignocellulose degradation for the genes 01720 and 01730 on the right-hand side. The gene 01640 lies downstream of the identified gene cluster and was annotated with two members of the major facilitator superfamily (MFS), PF07690 and PF13347. The broadly defined major facilitator

superfamily has a variety of functions and includes proteins which are active in sugar uptake [46]. In particular, PF13347 is characterized as a 'MFS/sugar transport protein' in the Pfam database. As additional evidence, we observed that the pentose transporter gene BACOVA_04388, which is part of a conserved xylan hydrolase gene cluster in *Bacteroides ovatus* (described by Dodd *et al*. [47]), also contained a MFS annotation (PF13347) according to IMG. Based on this evidence, the predicted AGa gene cluster might be involved in the uptake of sugars. The gene 01720, which flanks the predicted cluster from the upstream side, includes the uncharacterized family PF13585. Blasting the protein sequence of 01720 yielded many hits for hypothetical proteins, but among the top hits, we found a 'C-type lectin domain-containing protein' from *Flavobacterium sp.* F52. Lectins are required for sugar binding [48]. Finally, BLAST searches for the genes 01680 and 01690, which are the uncharacterized candidate genes of interest (green box in Figure S1, Additional file 12), resulted in hits to a 'putative outer-membrane insertion C-signal' domain for gene 01680 and a 'partial iron chelating ABC transporter substrate' (binding) domain for gene 01690. In summary, we found evidence for functionally related genes in the vicinity of the cellulase genes in the cluster. Similar to some PUL systems, the cluster seems to encode proteins with catalytic activities targeting cellulose or hemicellulose, as well as proteins with functions located in the membrane, such as sugar binding or sugar transport.

11. Ranking results with two different choices of genome-specific module weights

The ranking depends on our definition of the 'genome-specific' module weights. We compared two different definitions of the weights. First of all, the LDA model already provides probabilities that we could use as weights. But, as outlined in the supplementary Methods (see Additional file 3: Section 2), these weights rely on assumptions that are not appropriate in our case. Indeed, we observed that the weights of modules may depend on the presence of other, more dominant modules in a genome and cannot be compared across genomes. In some cases, this resulted in high ranks for modules that were not relevant for the phenotype of interest. For our analyses, we

therefore used alternative weights based on the completeness of modules that measure the percentage of the elements of a module that can be found in a (meta-)genome.

Despite the drawbacks of the probability weights, the results obtained with the two different types of weights agreed well. This indicates the robustness of the ranking with respect to the specific choice of weights. Both approaches mostly placed the same module at the top rank and the modules M1 to M5 were generally among the top 15 of ranked modules (in the top 10 in most cases).

References

1. Sweeney MD, Xu F: **Biomass converting enzymes as industrial biocatalysts for fuels and chemicals: Recent developments.** *Catalysts* 2012, **2:**244-263.
2. Singh D, Zeng JJ, Laskar DD, Deobald L, Hiscox WC, Chen SL: **Investigation of wheat straw biodegradation by *Phanerochaete chrysosporium*.** *Biomass Bioenerg* 2011, **35:**1030-1040.
3. Wolin MJ, Miller TL, Collins MD, Lawson PA: **Formate-dependent growth and homoacetogenic fermentation by a bacterium from human feces: description of *Bryantella formatexigens gen. nov., sp. nov*.** *Appl Environ Microbiol* 2003, **69:**6321-6326.
4. Anderson I, Abt B, Lykidis A, Klenk HP, Kyrpides N, Ivanova N: **Genomics of aerobic cellulose utilization systems in actinobacteria.** *PLoS ONE* 2012, **7:**e39331.
5. Gilad R, Rabinovich L, Yaron S, Bayer EA, Lamed R, Gilbert HJ, Shoham Y: **CelI, a noncellulosomal family 9 enzyme from *Clostridium thermocellum*, is a processive endoglucanase that degrades crystalline cellulose.** *J Bacteriol* 2003, **185:**391-398.
6. Doi RH, Kosugi A: **Cellulosomes: plant-cell-wall-degrading enzyme complexes.** *Nat Rev Microbiol* 2004, **2:**541-551.
7. Gaudin C, Belaich A, Champ S, Belaich JP: **CelE, a multidomain cellulase from *Clostridium cellulolyticum*: a key enzyme in the cellulosome?** *J Bacteriol* 2000, **182:**1910-1915.
8. Su X, Mackie RI, Cann IK: **Biochemical and mutational analyses of a multidomain cellulase/mannanase from *Caldicellulosiruptor bescii*.** *Appl Environ Microbiol* 2012, **78:**2230-2240.
9. Sakon J, Irwin D, Wilson DB, Karplus PA: **Structure and mechanism of endo/exocellulase E4 from *Thermomonospora fusca*.** *Nat Struct Biol* 1997, **4:**810-818.
10. Tolonen AC, Chilaka AC, Church GM: **Targeted gene inactivation in *Clostridium phytofermentans* shows that cellulose degradation requires the family 9 hydrolase Cphy3367.** *Mol Microbiol* 2009, **74:**1300-1313.

11. Devillard E, Goodheart DB, Karnati SK, Bayer EA, Lamed R, Miron J, Nelson KE, Morrison M: ***Ruminococcus albus* 8 mutants defective in cellulose degradation are deficient in two processive endocellulases, Cel48A and Cel9B, both of which possess a novel modular architecture.** *J Bacteriol* 2004, **186:**136-145.
12. Blouzard J-C, Coutinho PM, Fierobe H-P, Henrissat B, Lignon S, Tardif C, Pagès S, de Philip P: **Modulation of cellulosome composition in *Clostridium cellulolyticum*: Adaptation to the polysaccharide environment revealed by proteomic and carbohydrate-active enzyme analyses.** *Proteomics* 2010, **10:**541-554.
13. Gilbert HJ, Stalbrand H, Brumer H: **How the walls come crumbling down: Recent structural biochemistry of plant polysaccharide degradation.** *Curr Opin Plant Biol* 2008, **11:**338-348.
14. Larsbrink J, Izumi A, Ibatullin FM, Nakhai A, Gilbert HJ, Davies GJ, Brumer H: **Structural and enzymatic characterization of a glycoside hydrolase family 31 α-xylosidase from Cellvibrio japonicus involved in xyloglucan saccharification.** *Biochem J* 2011, **436:**567-580.
15. Gloster TM, Ibatullin FM, Macauley K, Eklof JM, Roberts S, Turkenburg JP, Bjornvad ME, Jorgensen PL, Danielsen S, Johansen KS, et al: **Characterization and three-dimensional structures of two distinct bacterial xyloglucanases from families GH5 and GH12.** *J Biol Chem* 2007, **282:**19177-19189.
16. Wilson D: **Evidence for a novel mechanism of microbial cellulose degradation.** *Cellulose* 2009, **16:**723-727.
17. Duan C-J, Feng J-X: **Mining metagenomes for novel cellulase genes.** *Biotechnol Lett* 2010, **32:**1765-1775.
18. Warnecke F, Luginbuhl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, Cayouette M, McHardy AC, Djordjevic G, Aboushadi N, et al: **Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite.** *Nature* 2007, **450:**560-565.
19. Pope PB, Mackenzie AK, Gregor I, Smith W, Sundset MA, McHardy AC, Morrison M, Eijsink VG: **Metagenomics of the Svalbard reindeer rumen microbiome reveals abundance of polysaccharide utilization loci.** *PLoS ONE* 2012, **7:**e38571.
20. Deboy RT, Mongodin EF, Fouts DE, Tailford LE, Khouri H, Emerson JB, Mohamoud Y, Watkins K, Henrissat B, Gilbert HJ, Nelson KE: **Insights into plant cell wall degradation from the genome sequence of the soil bacterium *Cellvibrio japonicus*.** *J Bacteriol* 2008, **190:**5455-5463.
21. Taylor LE, 2nd, Henrissat B, Coutinho PM, Ekborg NA, Hutcheson SW, Weiner RM: **Complete cellulase system in the marine bacterium *Saccharophagus degradans* strain 2-40T.** *J Bacteriol* 2006, **188:**3849-3861.
22. Taupp M, Mewis K, Hallam SJ: **The art and design of functional metagenomic screens.** *Curr Opin Biotechnol* 2011, **22:**465-472.
23. Schwarz WH: **The cellulosome and cellulose degradation by anaerobic bacteria.** *Appl Microbiol Biotechnol* 2001, **56:**634-649.
24. Kumar M, Khanna S: **Shift in microbial population in response to crystalline cellulose degradation during enrichment with a semi-desert soil.** *Int Biodeterior Biodegradation* 2014, **88:**134-141.
25. Olson DG, Tripathi SA, Giannone RJ, Lo J, Caiazza NC, Hogsett DA, Hettich RL, Guss AM, Dubrovsky G, Lynd LR: **Deletion of the Cel48S cellulase from Clostridium thermocellum.** *Proc Natl Acad Sci U S A* 2010, **107:**17727-17732.
26. Yoshida S, Hespen CW, Beverly RL, Mackie RI, Cann IK: **Domain analysis of a modular α-L-arabinofuranosidase with a unique carbohydrate binding strategy from the fiber-degrading bacterium *Fibrobacter succinogenes S85*.** *J Bacteriol* 2010, **192:**5424-5436.

27. Yoshida S, Mackie RI, Cann IK: **Biochemical and domain analyses of *FSUAxe6B*, a modular acetyl xylan esterase, identify a unique carbohydrate binding module in *Fibrobacter succinogenes S85*.** *J Bacteriol* 2010, **192:**483-493.
28. Wilson DB: **Three microbial strategies for plant cell wall degradation.** *Ann N Y Acad Sci* 2008, **1125:**289-297.
29. Lykidis A, Mavromatis K, Ivanova N, Anderson I, Land M, DiBartolo G, Martinez M, Lapidus A, Lucas S, Copeland A, et al: **Genome sequence and analysis of the soil cellulolytic actinomycete *Thermobifida fusca YX*.** *J Bacteriol* 2007, **189:**2477-2486.
30. Chapon V, Czjzek M, El Hassouni M, Py B, Juy M, Barras F: **Type II protein secretion in gram-negative pathogenic bacteria: The study of the structure/secretion relationships of the *cellulase Cel5* (formerly EGZ) from *Erwinia chrysanthemi*.** *J Mol Biol* 2001, **310:**1055-1066.
31. Blumer-Schuette SE, Lewis DL, Kelly RM: **Phylogenetic, microbiological, and glycoside hydrolase diversities within the extremely thermophilic, plant biomass-degrading genus *Caldicellulosiruptor*.** *Appl Environ Microbiol* 2010, **76:**8084-8092.
32. Peer A, Smith SP, Bayer EA, Lamed R, Borovok I: **Noncellulosomal cohesin- and dockerin-like modules in the three domains of life.** *FEMS Microbiol Lett* 2009, **291:**1-16.
33. Blumer-Schuette SE, Giannone RJ, Zurawski JV, Ozdemir I, Ma Q, Yin Y, Xu Y, Kataeva I, Poole FL, 2nd, Adams MW, et al: ***Caldicellulosiruptor* core and pangenomes reveal determinants for noncellulosomal thermophilic deconstruction of plant biomass.** *J Bacteriol* 2012, **194:**4015-4028.
34. Flint HJ, Bayer EA, Rincon MT, Lamed R, White BA: **Polysaccharide utilization by gut bacteria: Potential for new insights from genomic analysis.** *Nat Rev Microbiol* 2008, **6:**121-131.
35. Naas AE, Mackenzie AK, J. M, Schückel J, Willats WGT, Eijsink VGH, Pope PB: **Do rumen Bacteroidetes utilize an alternative mechanism for cellulose degradation?** *mBio* 2014, **5:**e01401-01414.
36. Pope PB, Denman SE, Jones M, Tringe SG, Barry K, Malfatti SA, McHardy AC, Cheng JF, Hugenholtz P, McSweeney CS, Morrison M: **Adaptation to herbivory by the Tammar wallaby includes bacterial and glycoside hydrolase profiles different from other herbivores.** *Proc Natl Acad Sci U S A* 2010, **107:**14793-14798.
37. Kudo H, Cheng KJ, Costerton JW: **Interactions between *Treponema bryantii* and cellulolytic bacteria in the in vitro degradation of straw cellulose.** *Can J Microbiol* 1987, **33:**244-248.
38. Nouaille R, Matulova M, Delort AM, Forano E: **Oligosaccharide synthesis in *Fibrobacter succinogenes* S85 and its modulation by the substrate.** *FEBS J* 2005, **272:**2416-2427.
39. Graham JE, Clark ME, Nadler DC, Huffer S, Chokhawala HA, Rowland SE, Blanch HW, Clark DS, Robb FT: **Identification and characterization of a multidomain hyperthermophilic cellulase from an archaeal enrichment.** *Nat Commun* 2011, **2:**375.
40. Anderson I, Scheuner C, Göker M, Mavromatis K, Hooper SD, Porat I, Klenk H-P, Ivanova N, Kyrpides N: **Novel insights into the diversity of catabolic metabolism from ten haloarchaeal genomes.** *PLoS ONE* 2011, **6:**e20237.
41. Martens EC, Lowe EC, Chiang H, Pudlo NA, Wu M, McNulty NP, Abbott DW, Henrissat B, Gilbert HJ, Bolam DN, Gordon JI: **Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts.** *PLoS Biol* 2011, **9:**e1001221.
42. Reeves AR, Wang GR, Salyers AA: **Characterization of four outer membrane proteins that play a role in utilization of starch by *Bacteroides thetaiotaomicron*.** *J Bacteriol* 1997, **179:**643-649.
43. Mackenzie AK, Pope PB, Pedersen HL, Gupta R, Morrison M, Willats WG, Eijsink VG: **Two SusD-like proteins encoded within a polysaccharide utilization locus of an uncultured ruminant bacteroidetes phylotype bind strongly to cellulose.** *Appl Environ Microbiol* 2012, **78:**5935-5937.

44. Martens EC, Koropatkin NM, Smith TJ, Gordon JI: **Complex glycan catabolism by the human gut microbiota: The Bacteroidetes Sus-like paradigm.** *J Biol Chem* 2009, **284:**24673-24677.
45. Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T, et al: **Metagenomic discovery of biomass-degrading genes and genomes from cow rumen.** *Science* 2011, **331:**463-467.
46. Pao SS, Paulsen IT, Saier MH, Jr.: **Major facilitator superfamily.** *Microbiol Mol Biol Rev* 1998, **62:**1-34.
47. Dodd D, Mackie RI, Cann IKO: **Xylan degradation, a metabolic property shared by rumen and human colonic bacteroidetes.** *Mol Microbiol* 2011, **79:**292-304.
48. Boraston AB, Tomme P, Amandoron EA, Kilburn DG: **A novel mechanism of xylan binding by a lectin-like module from *Streptomyces lividans* xylanase 10A.** *Biochem J* 2000, **350 Pt 3:**933-941.

D.3.2 Supplementary Methods

The following section includes the Supplemenatary Methods accompanying Konietzny *et al.* (2014).

Inference of phenotype-defining functional modules of protein families for microbial plant biomass degraders

Sebastian G.A. Konietzny, Phillip B. Pope, Aaron Weimann, Alice C. McHardy

– Supplementary Methods –

1. Preparing the input for LDA

From the set of protein family annotations for the (meta-)genomes, a suitable input collection for LDA was created. We limited our consideration to metagenomic sequences with taxonomic assignments because our aim was to process individual taxonomic bins, rather than whole metagenome samples. This strategy of splitting large metagenome samples into multiple taxonomic bins sharpened the co-occurrence signals, as it increased the number of input documents for LDA while reducing the effective sizes of the corresponding documents.

Protein families occurring in more than 75% of the genomes/taxonomic bins were removed from the annotation set, as these tend to be uninformative. To reduce the impact of very abundant families (e.g. those found by unspecific HMMs), the maximum number of hits considered for any family per genome or taxonomic bin was limited to 10, even if we found more than 10 hits for this particular family.

After filtering, for each genome and taxonomic bin, we defined a ‘document’ as the list of all protein families present in the corresponding annotation set. We therefore treated

protein family identifiers (e.g. GH16, PF00150) as equivalents to words in a natural language in text documents, and used these identifiers to define the input vocabulary V for LDA. Each created document can be seen as a multiset of protein families, meaning that single families may occur more than once. The resulting vocabulary comprised 8,413 family identifiers, 8,141 Pfam-A terms and 272 CAZy/dbCAN terms. In summary, we created an input collection of 3,216 documents, representing 2,884 prokaryotic genomes and 332 taxonomic bins from 18 metagenomes.

2. Two definitions of genome-specific module weights

We tested two definitions of weights: The probability weights $\theta_d(t)$ of the LDA model and 'completeness scores'.

In the generative LDA model, $\theta_d(t) = P(t \mid d)$ describes the probability that a newly sampled word for document d originates from topic t. Thus the word content of a document is most likely shaped by the topics with high probability values $\theta_d(t)$, and $\theta_d(t)$ reflects the relative importance of a topic.

It might appear intuitive by analogy to use the inferred $\theta_d(t)$ probabilities of topics to estimate the importance of modules for (meta-)genomes in the context of the phenotype prediction problem. This would correspond to setting the weight matrix W to be equivalent to the matrix Θ of the inferred LDA model, where single values are computed according to the formula:

$$\theta_t^d = \frac{n_t^{(d)} + \alpha}{\sum_{k=1}^{T} n_k^{(d)} + T\alpha} = \frac{n_t^{(d)} + \alpha}{|d| + T\alpha}$$

where $n_t^{(d)}$ is the number of words in document d that were assigned to topic t after the Gibbs sampling, T is the number of topics, and α is one of the two Dirichlet hyperparameters of the LDA model. However, one potential problem is that the values

reflect the relative amount of words that were assigned to the respective topic, which means the values are relative with respect to the importance of other topics that compete for the word assignments in the same document. Thus if more dominant topics exist in a document, they are likely to decrease the weights of all the other topics. This effect could be problematic when we compare the weights of a particular functional module in two different (meta-)genomes. The dominant topics of the individual documents are usually different from each other and also have varying relevance for the particular phenotype of interest. In both cases, the result is a document-specific shift of the weights of the non-dominant topics, which, in some cases, might hinder unbiased comparisons between different documents. We therefore aimed to define an intuitive and direct measure for the presence of a module in a (meta-)genome that does not depend on the document size or on other modules. The 'completeness score' of a module is the percentage of a module's protein families that occur in a specific genome or taxonomic bin. More precisely, we defined the weight of a module M_t in document d of the (meta-)genome collection based on completeness as:

$$weight_t(d) := \frac{|M_t \cap d|}{|M_t|} \times 100\%,$$

where $|M_t \cap d|$ is the size of the intersection of the protein family sets of module M_t and document d, and $|M_t|$ is the number of protein families contained in M_t. Note that the thresholds for these weights are more intuitive than those for weights based on the probability values $\theta_d(t)$.

3. The F-measure

The F-measure [1] is a combined measure of precision and recall, and is commonly used for assessing the performance in classification tasks [2] and for attribute ranking [3-5]. Since the process of lignocellulose degradation can be decomposed into different paradigms, we are searching for functional modules that are exclusive to some (but not all) of the lignocellulose degraders. Therefore, we should not require a module to

achieve perfect recall, although we are generally interested in modules that are specific to lignocellulose degraders. The F-measure is the weighted harmonic mean of precision and recall, where the weight factor β controls the relative importance of precision (P) and recall (R) [6]:

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

Setting $\beta = 1.0$ gives equal importance to precision and recall, and F_β reduces to F. We used $\beta = 0.5$ to give more importance to precision than to recall (i.e. recall half as important as precision) as suggested by Lewis [7].

4. Identification of similar modules across runs using Bron-Kerbosch

The Bron–Kerbosch algorithm [8] operates on an undirected graph and detects all maximal cliques, i.e. all complete subgraphs that cannot be extended by edges to additional vertices of the graph without losing their complete connectivity. In a complete subgraph, any two vertices are connected by an edge.

We constructed the input graph for the Bron-Kerbosch algorithm by inserting edges that represented the pairwise optimal mappings between the modules of any two LDA runs, as identified by the Hungarian algorithm [9]. Depending on the choice of the pairwise distance measure, we required KL-distances ≤ 5 or Jaccard distances ≤ 0.75 for the inclusion of edges. Cliques in this graph, found by the Bron-Kerbosch algorithm, represent sets of functional modules from different LDA runs with highly similar protein family content, as it holds true for all modules of a clique that they were identified as being their pairwise best matches by the Hungarian algorithm.

For the high-ranking modules, we found overlapping cliques of sizes ≥ 9, which shared one or more modules. Such overlaps establish (transitive) similarities between the modules of the overlapping cliques. Therefore, we merged the module sets of

overlapping cliques to define the set of functional modules used for generating a consensus module.

5. LDA model stability

Model inference with LDA depends on Gibbs sampling and needs to be monitored for successful convergence. We monitored the progression of the likelihood to assess the convergence of the Gibbs sampling procedure and specified 2000 iterations as the burn-in phase of a run, though a tendency for convergence could be observed much earlier. After burn-in, we collected 50 Gibbs samples from the posterior distribution over the model's parameters by taking a sample after every 10 iterations. Each sample was defined by an instance of the model's $\vec{z}$ vector, and we computed the distributions $\phi_t(w)$ and $\theta_d(t)$ for all documents and topics by averaging over the 50 samples. In general, averaging over samples allows us to summarize information from Gibbs samples [10, 11], but in the case of LDA, there is a theoretical risk of topic identity switching, even between samples from the same Markov chain [12]. We therefore used pairwise KL distances to track topics across the 50 samples and found no evidence for switching.

References

1. Van Rijsbergen CJ: *Information retrieval.* 2nd edn. London ; Boston: Butterworths; 1979.
2. Witten IH, Frank E: *Data mining: Practical machine learning tools and techniques, second edition (Morgan Kaufmann series in data management systems)* San Francisco, USA: Morgan Kaufmann Publishers Inc.; 2005.
3. Wang Z, Willard HF: **Evidence for sequence biases associated with patterns of histone methylation.** *BMC Genomics* 2012, **13:**367.
4. Akay MF: **Support vector machines combined with feature selection for breast cancer diagnosis.** *Expert Systems with Applications* 2009, **36:**3240-3247.
5. Chen YK, Li KB: **Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition.** *J Theor Biol* 2013, **318:**1-12.
6. Shatkay H, Feldman R: **Mining the biomedical literature in the genomic era: an overview.** *J Comput Biol* 2003, **10:**821-855.

7. Lewis DD: **Evaluating and optimizing autonomous text classification systems.** *Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 1995**:**246-254.
8. Bron C, Kerbosch J: **Algorithm 457: Finding all cliques of an undirected graph.** *Commun ACM* 1973, **16:**575-577.
9. Kuhn HW: **The Hungarian method for the assignment problem.** *Nav Res Log* 1955, **2:**83-97.
10. Gilks WR, Richardson S, Spiegelhalter DJ: *Markov Chain Monte Carlo in Practice* Boca Raton, Florida, USA: Chapman and Hall/CRC; 1999.
11. Heinrich G: *Technical report: Parameter estimation for text analysis* Darmstadt, Germany: Fraunhofer IGD; 2009.
12. Griffiths TL, Steyvers M: **Finding scientific topics.** *Proc Natl Acad Sci U S A* 2004, **101 Suppl 1:**5228-5235.

D.3.3 Overview of additional files

Additional file 1: Protein families of the consensus plant biomass degradation modules (PDMs). The Tables S1A-5A show each consensus module as a list of Pfam/CAZy terms. The consensus modules summarize highly similar modules from the 18 LDA runs and contain all elements that occurred in nine runs or more. The tables S1B-5B contain information about all additional Pfam/CAZy families that occurred in the similar modules in less than nine runs. Tables S6A and S6B list the families of the additional PUL module (see Results section). (Tables S1A-S6A for modules M1 to M5, and the PUL module.) [Additional tables S1B-S6B included on CD-ROM, not printed]

Additional file 2: Supplementary Note. Additional details about the results of the main manuscript. (Subsection D.3.1)

Additional file 3: Supplementary Methods. Additional details about the methods and preparation of the input data. (Subsection D.3.2)

Additional file 4: Plant biomass degradation module (PDM) assignments to the genomes of the learning set. PDM assignments to the genomes of known plant biomass degraders and non-degraders are visualized in Figure S1, and were obtained by leave-one-out classification. [Included on CD-ROM, not printed]

Additional file 5: Single predictions of the consensus modules on the learning set of genomes. Each sheet of the Excel file lists the predictions of one of the consensus modules with respect to the learning set of genomes (Tables S1-5). Table S6 contains the predictions of the additional PUL module. We used different colors to mark

true-positive (TP), true-negative (TN), false-positive (FP) and false-negative (FN) predictions (a description of the color coding is contained in the first sheet of the file). For each classified sample, we have provided several details (for example, name and phylum), as well as the genome-specific module weight (completeness score) of the respective consensus module. [Included on CD-ROM, not printed]

Additional file 6: Hemicellulolytic gene cluster in Fibrobacter *succinogenes* S85. The gene cluster in the figure encodes more than 10 hemicellulose-targeting enzymes in the genome of F. *succinogenes* S85. The protein domain architecture of the cluster genes has been described by Yoshida et al. [63,64]. F. *succinogenes* does not use a cellulosome-based degradation strategy, but rather a degradation paradigm that is still uncharacterized [92,93]. (**Note:** Reference numbers refer to the References section of the published article.) (Figure D.17)

Additional file 7: Co-occurrence profiles of the M1 protein families and GH6/GH48 across the learning set. Two heat maps display the combined co-occurrence profiles of the M1 protein families and two additional cellulases, GH6 and GH48, across the sets of the known phenotype-positive (Figure D.18) and phenotype-negative (Figure D.19) genomes, respectively. GH6 and GH48 were not assigned to module M1. The colors of the heat map cells represent the number of instances of each family in the respective genomes of the organisms (see legends and note that the counted number of instances was limited to a maximum of 10 per genome, as described in Methods). The phylogenetic relationships of the genomes are indicated by dendrograms alongside the rows of the heat maps.

(Figure D.18 and Figure D.19, respectively)

Additional file 8: Co-occurrence profiles of protein families that were weakly associated with M5 (across the learning set). Two heat maps displaying protein family co-occurrence profiles across the known phenotype-positive (Figure D.20) and phenotype-negative (Figure D.21) genomes. The columns of the heat maps represent the families that were weakly associated with the M5 module. These families did not satisfy the required threshold $C = 0.01$, but they belonged to the 50 protein families with the highest probabilities in the 16 topics that were used to create the M5 consensus module. Because the families failed to match the threshold, they were not counted for the consensus of M5. One example for such a family is GH48 (discussed in the main text). Families are ordered from left to right according to the number of topics in which they occurred. Families that occurred in less than 9 of the 16 topics are not displayed. We also added the cohesin and dockerin domains of M5 for a comparison. The colors of the heat map cells encode the number of instances of each family in the respective genomes of the organisms. (Figure D.20 and Figure D.21, respectively)

Additional file 9: Table S1 Leave-one-out results for the consensus plant biomass degradation modules (PDMs) obtained with the threshold $C = 0.005$. The threshold $C = 0.01$ was used to convert the discrete topic probability distributions of the LDA model into potential functional modules (see Methods). In additional tests, we used the threshold $C = 0.005$ instead. This cutoff level is less strict and allowed families with smaller probabilities to be included in the potential functional modules. This also resulted in enlarged consensus modules for the PDMs. Table S1 summarizes

the results obtained in leave-one-out validation for the PDMs M1 to M5 based on the threshold $C = 0.005$. (Figure D.22)

Additional file 10: Single predictions of the consensus modules on the remaining set of genomes and metagenome bins. Each sheet of the Excel file lists the predictions of one of the consensus modules with respect to all genomes and metagenome bins, except for the 120 known (non-)degraders used for learning (Tables S1-5). Table S6 contains the predictions of the additional PUL module. For each classified sample, we provide several details (for example, name and phylum), as well as the genome-specific module weight (completeness score) of the respective consensus module. [Included on CD-ROM, not printed]

Additional file 11: Venn diagram of the predicted occurrences of the modules M1 to M4. The diagram displays the overlap between the genomes and metagenome bins with predicted occurrences of the modules M1, M2, M3, and M4. Genomes from the learning set were excluded. (Figure D.23)

Additional file 12: Gene cluster in the cow rumen draft genome AGa. The red box in the figure marks a gene cluster (NODE_457020_ORF_01660 to NODE_457020_ORF_01710), which was identified based on the families assigned to highest scoring module (M1). The cluster is located on a 97,191-bp contig of the draft genome AGa (Bacteroidales) from the cow rumen metagenome [19]. The cluster includes three cellulases, based on assignments of the GH5 family, and a cellobiose phosphorylase (GH94; EC 2.4.1.20) with an attached putative carbohydrate binding domain (PF06204). The GH94 family was not assigned to the consensus module of M1, but it was contained in the M1 modules in 7 of 18 LDA runs.

Depending on the presence or absence of GH94 in the M1 modules of different runs, the gene cluster was identified either partly or completely. The cluster includes two genes (genes 01680 and 01690; green rectangle) without any annotated functional domains; these are uncharacterized genes that may be relevant for the degradation of lignocellulose. The presence of two Pfam families related to the major facilitator superfamily in gene 01640 (marked by the yellow box) indicates a link between the (hemi)cellulases of the GH5 and GH94 families, and sugar-binding or transport proteins located in the outer membrane (see Subsection D.3.1: Section 10). (**Note:** Reference numbers refer to the References section of the published article.) (Figure D.24)

Additional file 13: Protein sequences of the identified gene cluster in the cow rumen draft genome AGa. Protein sequences (NODE_457020 _ORF_01620 to NODE_457020_ORF_01740) from the cow rumen metagenome representing a gene cluster in the draft genome AGa (discussed in the main manuscript), as well as its surrounding genes. [Included on CD-ROM, not printed]

Additional file 14: Table S1 Microbial isolate strains (lignocellulose degraders and non-degraders) that were used as the learning set. Table S1 represents a manually curated list of 120 phenotype-positive or phenotype-negative prokaryotic genomes, including the respective literature references. [Included on CD-ROM, not printed]

Additional file contents

M1 Consensus

Domain	# runs	Description
GH5	17	chitosanase (EC 3.2.1.132); beta-mannosidase (EC 3.2.1.25); Cellulase (EC 3.2.1.4); glucan 1,3-beta-glucosidase (EC 3.2.1.58); licheninase (EC 3.2.1.73); glucan endo-1,6-beta-glucosidase (EC 3.2.1.75); mannan endo-beta-1,4-mannosidase (EC 3.2.1.78); endo-beta-1,4-xylanase (EC 3.2.1.8); cellulose beta-1,4-cellobiosidase (EC 3.2.1.91); [...]
PF00150	18	Cellulase (glycosyl hydrolase family 5)
GH9	18	endoglucanase (EC 3.2.1.4); cellobiohydrolase (EC 3.2.1.91); beta-glucosidase (EC 3.2.1.21); exo-beta-glucosaminidase (EC 3.2.1.165)
PF00759	17	Glycosyl hydrolase family 9
GH10	17	endo-1,4-beta-xylanase (EC 3.2.1.8); endo-1,3-beta-xylanase (EC 3.2.1.32)
PF00331	17	Glycosyl hydrolase family 10
GH26	17	beta-mannanase (EC 3.2.1.78); beta-1,3-xylanase (EC 3.2.1.32)
PF02156	17	Glycosyl hydrolase family 26
GH43	18	beta-xylosidase (EC 3.2.1.37); beta-1,3-xylosidase (EC 3.2.1.-); alpha-L-arabinofuranosidase (EC 3.2.1.55); arabinanase (EC 3.2.1.99); xylanase (EC 3.2.1.8); galactan 1,3-beta-galactosidase (EC 3.2.1.145)
PF04616	17	Glycosyl hydrolases family 43
CBM4	16	Modules of approx. 150 residues found in bacterial enzymes. Binding of these modules has been demonstrated with xylan, beta-1,3-glucan, beta-1,3-1,4-glucan, beta-1,6-glucan and amorphous cellulose but not with crystalline cellulose.
PF02018	17	Carbohydrate binding domain (CBM_4_9)
CBM6	18	Modules of approx. 120 residues. The cellulose-binding function has been demonstrated in one case on amorphous cellulose and beta-1,4-xylan. Some of these modules also bind beta-1,3-glucan, beta-1,3-1,4-glucan, and beta-1,4-glucan.
PF03422	18	Carbohydrate binding module (family 6)
PF02927	17	N-terminal ig-like domain of cellulase
CBM35	17	Modules of approx. 130 residues. A module that is conserved in three Cellvibrio xylan-degrading enzymes binds to xylan and the interaction is calcium dependent, while a module from a Cellvibrio mannanase binds to decorated soluble mannans and mannooligosaccharides. [...]
PF00756	13	Putative esterase
PF13472	10	GDSL-like Lipase/Acylhydrolase family

Figure D.11: Protein families of the consensus plant biomass degradation module (PDM) M1. Consensus modules summarize highly similar modules from the 18 LDA runs and contain all elements that occurred in nine runs or more.

M2 Consensus

Domain	# runs	Description
GH16	16	xyloglucan:xyloglucosyltransferase (EC 2.4.1.207); keratan-sulfate endo-1,4-beta-galactosidase (EC 3.2.1.103); endo-1,3-beta-glucanase (EC 3.2.1.39); endo-1,3(4)-beta-glucanase (EC 3.2.1.6); licheninase (EC 3.2.1.73); beta-agarase (EC 3.2.1.81); kappa;-carrageenase (EC 3.2.1.83); xyloglucanase (EC 3.2.1.151)
PF00722	17	Glycosyl hydrolases family 16
GH30	10	glucosylceramidase (EC 3.2.1.45); beta-1,6-glucanase (EC 3.2.1.75); beta-xylosidase (EC 3.2.1.37); beta-fucosidase (EC 3.2.1.38); beta-glucosidase (3.2.1.21); endo-beta-1,6-galactanase (EC:3.2.1.164)
CBM6	18	Modules of approx. 120 residues. The cellulose-binding function has been demonstrated in one case on amorphous cellulose and beta-1,4-xylan. Some of these modules also bind beta-1,3-glucan, beta-1,3-1,4-glucan, and beta-1,4-glucan.
PF03422	18	Carbohydrate binding module (family 6)
CBM16	16	Carbohydrate-binding module 16. Binding to cellulose and glucomannan demonstrated [B. Bae et al (2008) J Biol Chem. 283:12415-25 (PMID: 18025086)]
CBM35	18	Modules of approx. 130 residues. A module that is conserved in three Cellvibrio xylan-degrading enzymes binds to xylan and the interaction is calcium dependent, while a module from a Cellvibrio mannanase binds to decorated soluble mannans and mannooligosaccharides. [...]
CBM61	17	Modules of approx. 150 residues found appended to GH16, GH30, GH31, GH43, GH53 and GH66 catalytic domains. A beta-1,4-galactan binding function has been demonstrated for the CBM60 of Thermotoga maritima GH53 galactanase [PMID: 20826814].
CBM47	17	Modules of approx 150 residues. Fucose-binding activity demonstrated
CBM32	17	Binding to galactose and lactose has been demonstrated for the module of Micromonospora viridifaciens sialidase (PMID: 16239725). Binding to polygalacturonic acid has been shown for a Yersinia member (PMID: 17292916). [...]
CBM13	11	Modules of approx. 150 residues which always appear as a threefold internal repeat. The only apparent exception to this, xylanase II of Actinomadura sp. FC7 (GenBank U08894), is in fact not completely sequenced. These modules were first identified in several plant lectins such as ricin or agglutinin of Ricinus communis which bind galactose residues. [...]
PF14200	11	Ricin-type beta-trefoil lectin domain-like
PF00652	11	Ricin-type beta-trefoil lectin domain
GH87	17	mycodextranase (EC 3.2.1.61); alpha-1,3-glucanase (EC 3.2.1.59)
PF00754	17	F5/8 type C domain
PF00041	17	Fibronectin type III domain
GH119	17	alpha-amylase (EC 3.2.1.1) (Distantly related to family GH57)
PF12708	14	Pectate lyase superfamily protein
PF02311	13	AraC-like ligand binding domain
PF02018	13	Carbohydrate binding domain (CBM_4_9)
GH55	12	exo-beta-1,3-glucanase (EC 3.2.1.58); endo-beta-1,3-glucanase (EC 3.2.1.39).
PF13483	9	Beta-lactamase superfamily domain

Figure D.12: Protein families of the consensus plant biomass degradation module (PDM) M2. Consensus modules summarize highly similar modules from the 18 LDA runs and contain all elements that occurred in nine runs or more.

M3 Consensus

Domain	# runs	Description
GH5	18	chitosanase (EC 3.2.1.132); beta-mannosidase (EC 3.2.1.25); Cellulase (EC 3.2.1.4); glucan 1,3-beta-glucosidase (EC 3.2.1.58); licheninase (EC 3.2.1.73); glucan endo-1,6-beta-glucosidase (EC 3.2.1.75); mannan endo-beta-1,4-mannosidase (EC 3.2.1.78); endo-beta-1,4-xylanase (EC 3.2.1.8); cellulose beta-1,4-cellobiosidase (EC 3.2.1.91); [...]
GH43	12	beta-xylosidase (EC 3.2.1.37); beta-1,3-xylosidase (EC 3.2.1.-); alpha-L-arabinofuranosidase (EC 3.2.1.55); arabinanase (EC 3.2.1.99); xylanase (EC 3.2.1.8); galactan 1,3-beta-galactosidase (EC 3.2.1.145)
PF04616	12	Glycosyl hydrolases family 43
PF03629	17	Domain of unknown function (DUF303)
PF01095	18	Pectinesterase
PL1	18	pectate lyase (EC 4.2.2.2); exo-pectate lyase (EC 4.2.2.9); pectin lyase (EC 4.2.2.10).
PF12708	18	Pectate lyase superfamily protein
GH28	18	polygalacturonase (EC 3.2.1.15); exo-polygalacturonase (EC 3.2.1.67); exo-polygalacturonosidase (EC 3.2.1.82); rhamnogalacturonase (EC 3.2.1.-); endo-xylogalacturonan hydrolase (EC 3.2.1.-); rhamnogalacturonan alpha-L-rhamnopyranohydrolase (EC 3.2.1.40)
PF00295	18	Glycosyl hydrolases family 28
CE6	12	acetyl xylan esterase (EC 3.1.1.72).
CE7	18	acetyl xylan esterase (EC 3.1.1.72); cephalosporin-C deacetylase (EC 3.1.1.41).
CE8	18	pectin methylesterase (EC 3.1.1.11).
CE12	18	pectin acetylesterase (EC 3.1.1.-); rhamnogalacturonan acetylesterase (EC 3.1.1.-); acetyl xylan esterase (EC 3.1.1.72)
PL9	13	pectate lyase (EC 4.2.2.2); exopolygalacturonate lyase (EC 4.2.2.9); thiopeptidoglycan lyase (EC 4.2.2.-).
GH106	16	alpha-L-rhamnosidase (EC 3.2.1.40)
GH88	14	d-4,5 unsaturated beta-glucuronyl hydrolase (EC 3.2.1.-)
PF07470	18	Glycosyl Hydrolase Family 88
GH105	18	unsaturated rhamnogalacturonyl hydrolase (EC 3.2.1.-)
PF00657	18	GDSL-like Lipase/Acylhydrolase
PF13229	18	Right handed beta helix region
GH95	17	alpha-1,2-L-fucosidase (EC 3.2.1.63); alpha-L-fucosidase (EC 3.2.1.51)
PF13472	15	GDSL-like Lipase/Acylhydrolase family
PF13524	13	Glycosyl transferases group 1

Figure D.13: Protein families of the consensus plant biomass degradation module (PDM) M3. Consensus modules summarize highly similar modules from the 18 LDA runs and contain all elements that occurred in nine runs or more.

M4 Consensus

Domain	# runs	Description
GH5	18	chitosanase (EC 3.2.1.132); beta-mannosidase (EC 3.2.1.25); Cellulase (EC 3.2.1.4); glucan 1,3-beta-glucosidase (EC 3.2.1.58); licheninase (EC 3.2.1.73); glucan endo-1,6-beta-glucosidase (EC 3.2.1.75); mannan endo-beta-1,4-mannosidase (EC 3.2.1.78); endo-beta-1,4-xylanase (EC 3.2.1.8); cellulose beta-1,4-cellobiosidase (EC 3.2.1.91); [...]
PF00150	13	Cellulase (glycosyl hydrolase family 5)
GH43	16	beta-xylosidase (EC 3.2.1.37); beta-1,3-xylosidase (EC 3.2.1.-); alpha-L-arabinofuranosidase (EC 3.2.1.55); arabinanase (EC 3.2.1.99); xylanase (EC 3.2.1.8); galactan 1,3-beta-galactosidase (EC 3.2.1.145)
PF04616	15	Glycosyl hydrolases family 43
GH2	18	beta-galactosidase (EC 3.2.1.23) ; beta-mannosidase (EC 3.2.1.25); beta-glucuronidase (EC 3.2.1.31); mannosylglycoprotein endo-beta-mannosidase (EC 3.2.1.152); exo-beta-glucosaminidase (EC 3.2.1.165)
PF02836	18	Glycosyl hydrolases family 2, TIM barrel domain
PF00703	18	Glycosyl hydrolases family 2
PF02837	18	Glycosyl hydrolases family 2, sugar binding domain
GH3	18	beta-glucosidase (EC 3.2.1.21); xylan 1,4-beta-xylosidase (EC 3.2.1.37); beta-N-acetylhexosaminidase (EC 3.2.1.52); glucan 1,3-beta-glucosidase (EC 3.2.1.58); glucan 1,4-beta-glucosidase (EC 3.2.1.74); exo-1,3-1,4-glucanase (EC 3.2.1.-); alphalpha-L-arabinofuranosidase (EC 3.2.1.55).
PF01915	18	Glycosyl hydrolase family 3 C-terminal domain
PF00933	18	Glycosyl hydrolase family 3 N terminal domain
GH35	9	beta-galactosidase (EC 3.2.1.23); exo-beta-glucosaminidase (EC 3.2.1.165)
PF02449	10	Beta-galactosidase
GH42	10	beta-galactosidase (EC 3.2.1.23)
PF02065	17	Melibiase (GH27) [GH-D clan, a superfamily of alpha-galactosidases]
GH31	18	alpha-glucosidase (EC 3.2.1.20); alpha-1,3-glucosidase (EC 3.2.1.84); sucrase-isomaltase (EC 3.2.1.48) (EC 3.2.1.10); alpha-xylosidase (EC 3.2.1.-); alpha-glucan lyase (EC 4.2.2.13); isomaltosyltransferase (EC 2.4.1.-). [GH-D clan, a superfamily of alpha-galactosidases]
PF01055	18	Glycosyl hydrolases family 31
GH36	17	alpha-galactosidase (EC 3.2.1.22); alpha-N-acetylgalactosaminidase (EC 3.2.1.49); stachyose synthase (EC 2.4.1.67); raffinose synthase (EC 2.4.1.82) [GH-D clan, a superfamily of alpha-galactosidases]
PF14310	18	Fibronectin type III-like domain
PF07859	15	alpha/beta hydrolase fold
CE10	14	arylesterase (EC 3.1.1.-); carboxyl esterase (EC 3.1.1.3); acetylcholinesterase (EC 3.1.1.7); cholinesterase (EC 3.1.1.8); sterol esterase (EC 3.1.1.13); brefeldin A esterase (EC 3.1.1.-).
GH32	12	invertase (EC 3.2.1.26); endo-inulinase (EC 3.2.1.7); beta-2,6-fructan 6-levanbiohydrolase (EC 3.2.1.64); endo-levanase (EC 3.2.1.65); exo-inulinase (EC 3.2.1.80); fructan beta-(2,1)-fructosidase/1-exohydrolase (EC 3.2.1.153); fructan beta-(2,6)-fructosidase/6-exohydrolase (EC 3.2.1.154); sucrose:sucrose 1-fructosyltransferase (EC 2.4.1.99); [...]
PF00135	10	Carboxylesterase family
PF13802	9	Galactose mutarotase-like
GH106	9	alpha-L-rhamnosidase (EC 3.2.1.40)

Figure D.14: Protein families of the consensus plant biomass degradation module (PDM) M4. Consensus modules summarize highly similar modules from the 18 LDA runs and contain all elements that occurred in nine runs or more.

M5 Consensus

Domain	# runs	Description
PF00942	16	Cellulose binding domain
GH124	16	endoglucanase (EC 3.2.1.4)
CBM3	16	Modules of approx. 150 residues found in bacterial enzymes. The cellulose-binding function has been demonstrated in many cases. In one instance binding to chitin has been reported.
dockerin		
PF00404	16	Dockerin type I repeat
cohesin		
PF00963	16	Cohesin domain
PF07591	16	Pretoxin HINT domain
PF13186	16	Domain of unknown function (DUF4008)
CBM36	15	Modules of approx. 120-130 residues displaying structural similarities to CBM6 modules. The only CBM36 currently characterised, that from Paenbacillus polymyxa xylanase 43A, shows calcium-dependent binding of xylans and xylooligosaccharides. [...]
PF05593	12	RHS Repeat
PF07238	10	PilZ domain
PF13403	9	Hint domain

Figure D.15: Protein families of the consensus plant biomass degradation module (PDM) M5. Consensus modules summarize highly similar modules from the 18 LDA runs and contain all elements that occurred in nine runs or more.

PUL module consensus

Domain	# runs	Description
PF07980	17	SusD family
PF00593	17	TonB dependent receptor
PF07715	17	TonB-dependent Receptor Plug Domain
PF14322	17	Starch-binding associating with outer membrane
PF13715	17	Cna protein B-type domain
PF13620	16	Carboxypeptidase regulatory-like domain
PF13568	13	Outer membrane protein beta-barrel domain
PF00691	13	OmpA family
PF13505	12	Outer membrane protein beta-barrel domain
PF02321	11	Outer membrane efflux protein

Figure D.16: Protein families of the additional polysaccharide utilization locus (PUL) consensus module. Consensus modules summarize highly similar modules from the 18 LDA runs and contain all elements that occurred in nine runs or more.

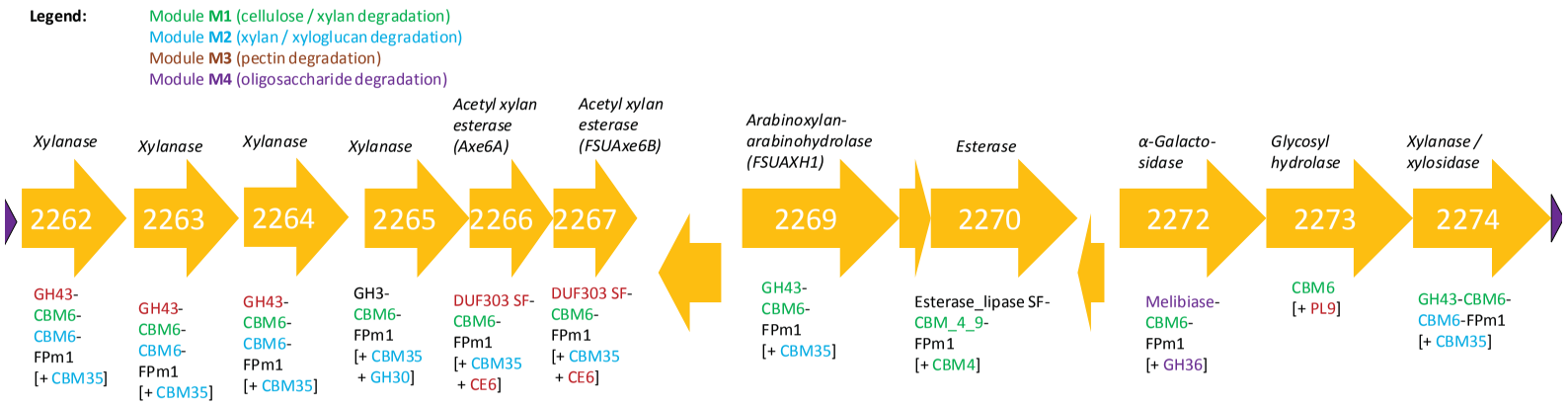


Figure D.17: Hemicellulolytic gene cluster in Fibrobacter *succinogenes* S85. The gene cluster encodes more than 10 hemicellulose-targeting enzymes in the genome of F. *succinogenes* S85. The protein domain architecture of the cluster genes has been described by Yoshida et al. [63,64]. F. *succinogenes* does not use a cellulosome-based degradation strategy, but rather a degradation paradigm that is still uncharacterized [92,93]. (**Note:** Reference numbers refer to the References section of the publication.)

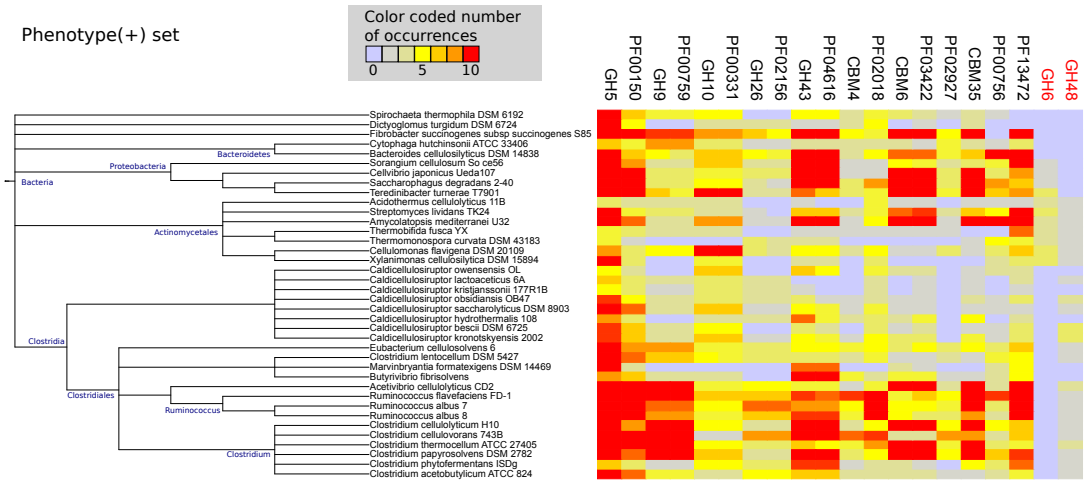


Figure D.18: Co-occurrence profiles of the M1 protein families and GH6/GH48 across the learning set. (Part 1 of 2) The heat maps displays the combined co-occurrence profiles of the M1 protein families and two additional cellulases, GH6 and GH48, across the set of the known phenotype-positive genomes. GH6 and GH48 were not assigned to module M1. The colors of the heat map cells represent the number of instances of each family in the respective genomes of the organisms (see legends and note that the counted number of instances was limited to a maximum of 10 per genome, as described in Methods). The phylogenetic relationships of the genomes are indicated by dendrograms alongside the rows of the heat maps.

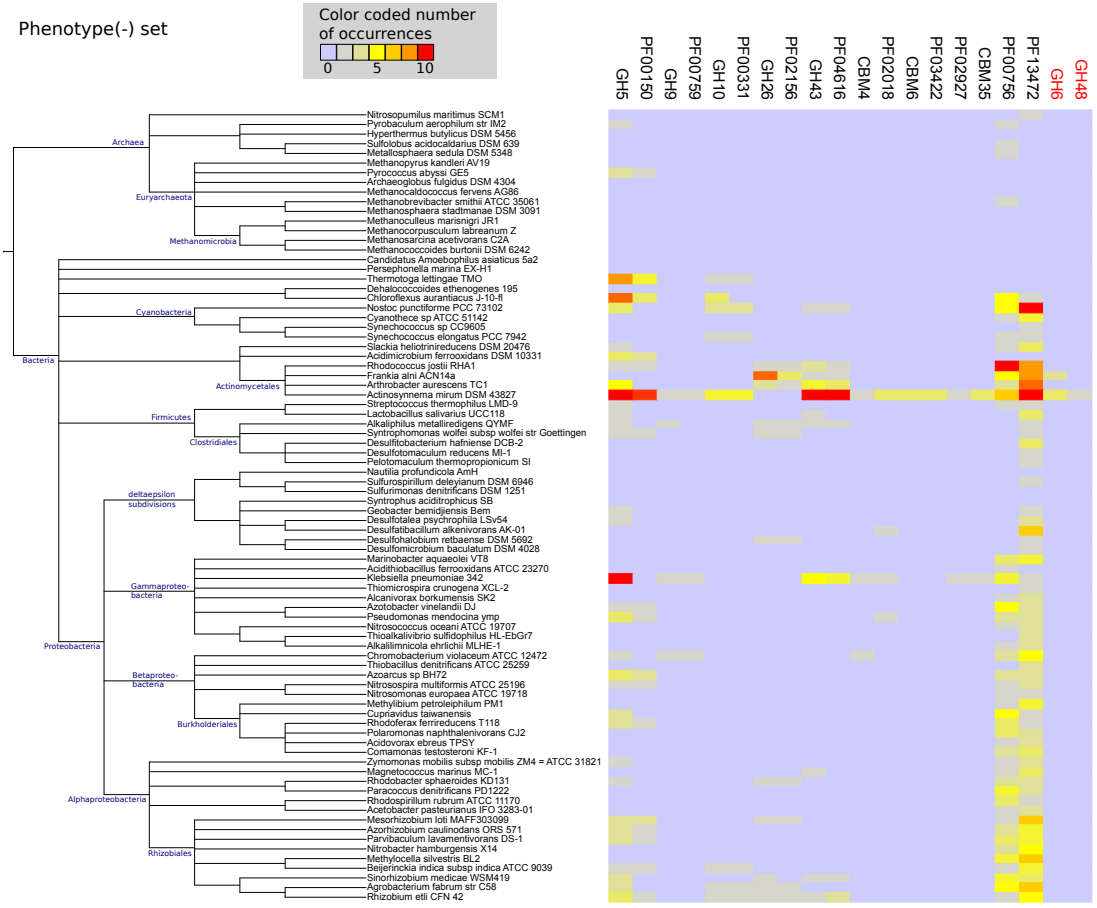


Figure D.19: Co-occurrence profiles of the M1 protein families and GH6/GH48 across the learning set. (Part 2 of 2) The heat maps displays the combined co-occurrence profiles of the M1 protein families and two additional cellulases, GH6 and GH48, across the set of the known phenotype-negative genomes. GH6 and GH48 were not assigned to module M1. The colors of the heat map cells represent the number of instances of each family in the respective genomes of the organisms (see legends and note that the counted number of instances was limited to a maximum of 10 per genome, as described in Methods). The phylogenetic relationships of the genomes are indicated by dendrograms alongside the rows of the heat maps.

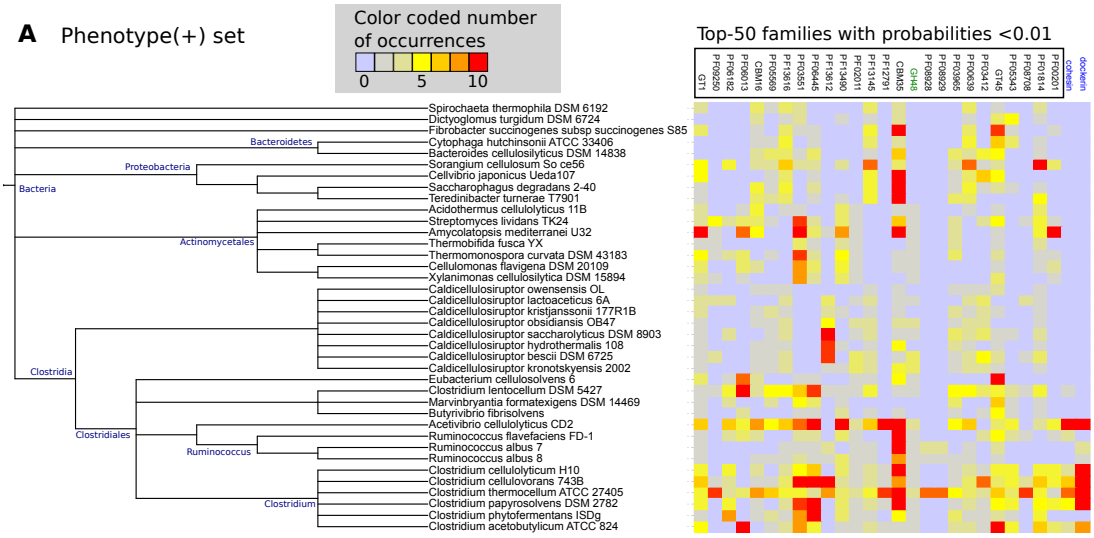


Figure D.20: Co-occurrence profiles of protein families that were weakly associated with M5 (across the learning set). (Part 1 of 2) The heat map displays protein family co-occurrence profiles across the known phenotype-positive genomes. The columns of the heat map represent the families that were weakly associated with the M5 module. These families did not satisfy the required threshold $C = 0.01$, but they belonged to the 50 protein families with the highest probabilities in the 16 topics that were used to create the M5 consensus module. Because the families failed to match the threshold, they were not counted for the consensus of M5. One example for such a family is GH48 (discussed in the main text). Families are ordered from left to right according to the number of topics in which they occurred. Families that occurred in less than 9 of the 16 topics are not displayed. We also added the cohesin and dockerin domains of M5 for a comparison. The colors of the heat map cells encode the number of instances of each family in the respective genomes of the organisms.

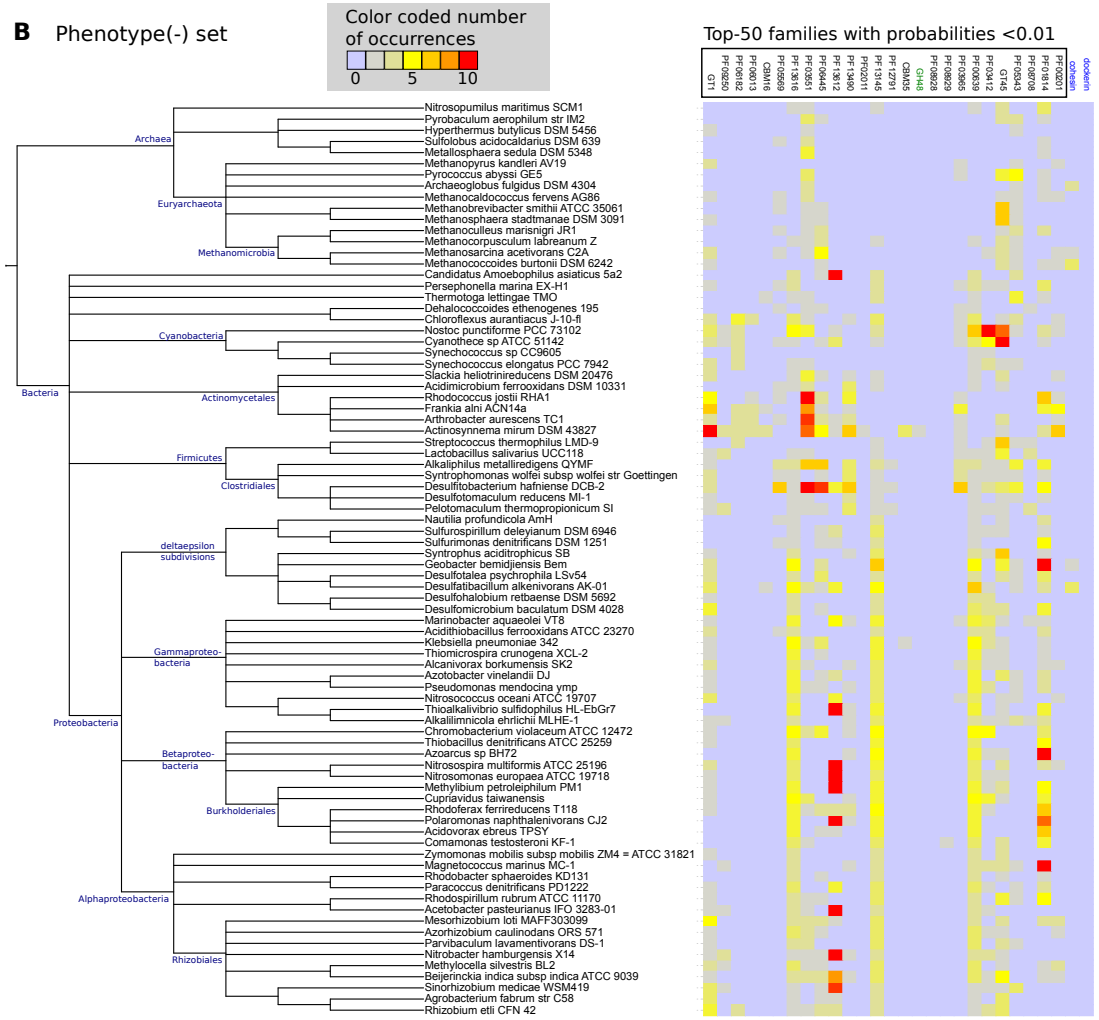


Figure D.21: Co-occurrence profiles of protein families that were weakly associated with M5 (across the learning set). (Part 2 of 2) The heat map displays protein family co-occurrence profiles across the known phenotype-negative genomes. The columns of the heat map represent the families that were weakly associated with the M5 module. These families did not satisfy the required threshold $C = 0.01$, but they belonged to the 50 protein families with the highest probabilities in the 16 topics that were used to create the M5 consensus module. Because the families failed to match the threshold, they were not counted for the consensus of M5. One example for such a family is GH48 (discussed in the main text). Families are ordered from left to right according to the number of topics in which they occurred. Families that occurred in less than 9 of the 16 topics are not displayed. We also added the cohesin and dockerin domains of M5 for a comparison. The colors of the heat map cells encode the number of instances of each family in the respective genomes of the organisms.

Table S1 Leave-one-out results for the consensus plant biomass degradation modules (PDMs) obtained with the threshold $C = 0.005$.

		M1	M2	M3	M4	M5
Set of recurring modules (18 repetitions of analyses)	**Number of modules in set**	18	18	18	18	16
	Average rank	1.3 (±0.57)	2.4 (±0.61)	4.2 (±1.5)	6 (±1.6)	7.5 (±3.4)
Consensus PDM	**Size**	45	37	53	37	34
Performance evaluation	**LOO $F_{0.5}$-score (%)**	89.08	93.37	83.33	85.86	80
	LOO recall (%)	81.58	81.58	76.32	89.47	63.16
	LOO precision (%)	91.18	96.88	85.29	85	85.71
	LOO accuracy (%)	91.67	93.33	88.33	91.67	85
	LOO MAC (%)	88.96	90.18	85.11	91.08	79.14
	Weight threshold (mean)	57.77	45.96	56.47	56.77	41.27

Figure D.22: Table S1 Leave-one-out results for the consensus plant biomass degradation modules (PDMs) obtained with the threshold *C = 0.005*. The threshold $C = 0.01$ was used to convert the discrete topic probability distributions of the LDA model into potential functional modules (see Methods). In additional tests, we used the threshold $C = 0.005$ instead. This cutoff level is less strict and allowed families with smaller probabilities to be included in the potential functional modules. This also resulted in enlarged consensus modules for the PDMs. Table S1 summarizes the results obtained in leave-one-out validation for the PDMs M1 to M5 based on the threshold $C = 0.005$.

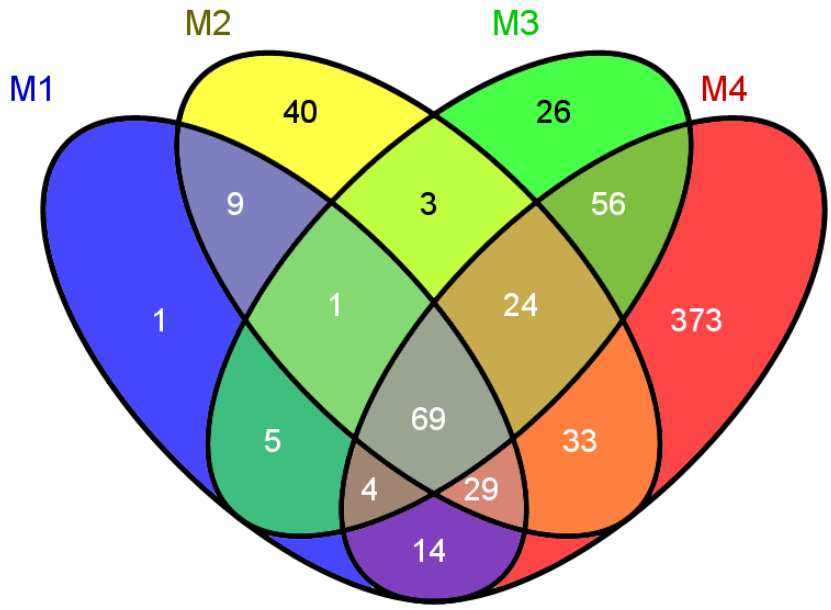


Figure D.23: Venn diagram of the predicted occurrences of the modules M1 to M4. The diagram displays the overlap between the genomes and metagenome bins with predicted occurrences of the modules M1, M2, M3, and M4. Genomes from the learning set were excluded.

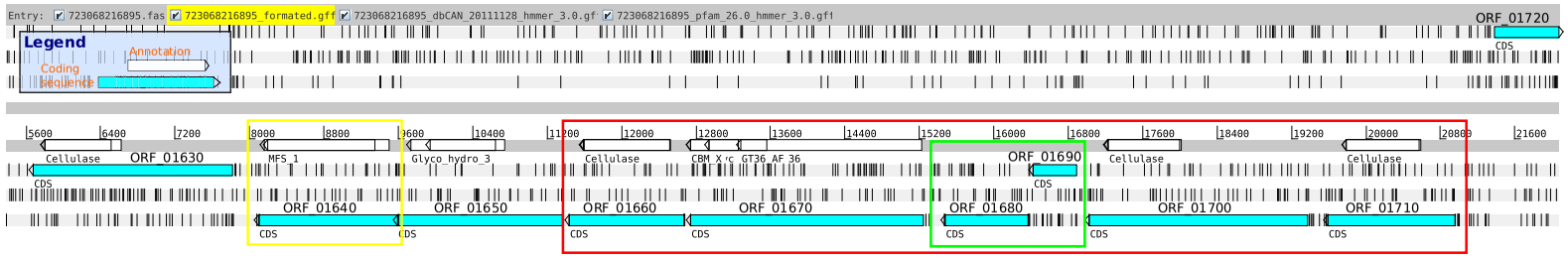


Figure D.24: Gene cluster in the cow rumen draft genome AGa. The red box marks a gene cluster (NODE_457020_ORF_01660 to NODE_457020_ORF_01710), which was identified based on the families assigned to highest scoring module (M1). The cluster is located on a 97,191-bp contig of the draft genome AGa (Bacteroidales) from the cow rumen metagenome [19]. The cluster includes three cellulases, based on assignments of the GH5 family, and a cellobiose phosphorylase (GH94; EC 2.4.1.20) with an attached putative carbohydrate binding domain (PF06204). The GH94 family was not assigned to the consensus module of M1, but it was contained in the M1 modules in 7 of 18 LDA runs. Depending on the presence or absence of GH94 in the M1 modules of different runs, the gene cluster was identified either partly or completely. The cluster includes two genes (genes 01680 and 01690; green rectangle) without any annotated functional domains; these are uncharacterized genes that may be relevant for the degradation of lignocellulose. The presence of two Pfam families related to the major facilitator superfamily in gene 01640 (marked by the yellow box) indicates a link between the (hemi)cellulases of the GH5 and GH94 families, and sugar-binding or transport proteins located in the outer membrane (see Subsection D.3.1: Section 10). (**Note:** Reference numbers refer to the References section of the publication.)

APPENDIX E

Licenses for figure reprints

This chapter includes the licenses for images from external sources that were reprinted in this thesis.

Title:	Determination of the Core of a Minimal Bacterial Gene Set
Author:	Rosario Gil, Francisco J. Silva, Juli Peretó et al.
Publication:	Microbiology and Molecular Biology Reviews
Publisher:	American Society for Microbiology
Date:	Sep 1, 2004

Permissions Request

ASM authorizes an advanced degree candidate to republish the requested material in his/her doctoral thesis or dissertation. If your thesis, or dissertation, is to be published commercially, then you must reapply for permission.

BACK CLOSE WINDOW

Citation: Duan J, Jiang W, Cheng Z, Heikkila JJ, Glick BR (2013) The Complete Genome Sequence of the Plant Growth-Promoting Bacterium *Pseudomonas* sp. UW4. PLoS ONE 8(3): e58640. doi:10.1371/journal.pone.0058640

Editor: Mark Willem John van Passel, Wageningen University, Netherlands

Received: October 25, 2012; **Accepted:** February 5, 2013; **Published:** March 13, 2013

Funding: Funding for this study was provided by the Natural Science and Engineering Research Council of Canada. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Title: A large-scale evaluation of computational protein function prediction

Author: Predrag Radivojac, Wyatt T Clark, Tal Ronnen Oron, Alexandra M Schnoes, Tobias Wittkop, Artem Sokolov

Publication: Nature Methods

Publisher: Nature Publishing Group

Date: Jan 27, 2013

Creative Commons

The request you have made is considered to be non-commercial/educational. As the article you have requested has been distributed under a Creative Commons license (Attribution-Noncommercial), you may reuse this material for non-commercial/educational purposes without obtaining additional permission from Nature Publishing Group, providing that the author and the original source of publication are fully acknowledged(please see the article itself for the license version number). You may reuse this material without obtaining permission from Nature Publishing Group, providing that the author and the original source of publication are fully acknowledged, as per the terms of the license. For license terms, please see http://creativecommons.org/

NATURE PUBLISHING GROUP LICENSE TERMS AND CONDITIONS

Nov 25, 2015

This is a License Agreement between Sebastian GA Konietzny ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3755861202630
License date	Nov 25, 2015
Licensed content publisher	Nature Publishing Group
Licensed content publication	Nature Reviews Genetics
Licensed content title	Network biology: understanding the cell's functional organization
Licensed content author	Albert-Laszlo Barabasi and Zoltan N. Oltvai
Licensed content date	Feb 1, 2004
Volume number	5
Issue number	2
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Figures	Figure 2 | Yeast protein interaction network.
Author of this NPG article	no
Your reference number	None
Title of your thesis / dissertation	Detection of functional modules in (meta-)genomic datasets
Expected completion date	Jun 2015
Estimated size (number of pages)	200
Total	0.00 EUR
Terms and Conditions	

Terms and Conditions for Permissions

Nature Publishing Group hereby grants you a non-exclusive license to reproduce this material for this purpose, and for no other use,subject to the conditions below:

1. NPG warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to Nature Publishing Group and does not carry the copyright of another entity (as credited in the published version). If the credit line on any part of the material you have requested

indicates that it was reprinted or adapted by NPG with permission from another source, then you should also seek permission from that source to reuse the material.

2. Permission granted free of charge for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to the work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.Where print permission has been granted for a fee, separate permission must be obtained for any additional, electronic re-use (unless, as in the case of a full paper, this has already been accounted for during your initial request in the calculation of a print run).NB: In all cases, web-based use of full-text articles must be authorized separately through the 'Use on a Web Site' option when requesting permission.

3. Permission granted for a first edition does not apply to second and subsequent editions and for editions in other languages (except for signatories to the STM Permissions Guidelines, or where the first edition permission was granted for free).

4. Nature Publishing Group's permission must be acknowledged next to the figure, table or abstract in print. In electronic form, this acknowledgement must be visible at the same time as the figure/table/abstract, and must be hyperlinked to the journal's homepage.

5. The credit line should read:
Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)
For AOP papers, the credit line should read:
Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

   Note: For republication from the *British Journal of Cancer,* the following credit lines apply.
Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)For AOP papers, the credit line should read:
Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

6. Adaptations of single figures do not require NPG approval. However, the adaptation should be credited as follows:

   Adapted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)

   Note: For adaptation from the *British Journal of Cancer,* the following credit line applies.
Adapted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

7. Translations of 401 words up to a whole article require NPG approval. Please visit http://www.macmillanmedicalcommunications.com for more information.Translations of up to a 400 words do not require NPG approval. The translation should be credited as follows:

   Translated by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication).

   Note: For translation from the *British Journal of Cancer,* the following credit line applies.
Translated by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

We are certain that all parties will benefit from this agreement and wish you the best in the use of this material. Thank you.
Special Terms:

v1.1

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

NATURE PUBLISHING GROUP LICENSE
TERMS AND CONDITIONS

May 02, 2015

This is a License Agreement between Sebastian GA Konietzny ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3620801144824
License date	May 02, 2015
Licensed content publisher	Nature Publishing Group
Licensed content publication	Nature Reviews Microbiology
Licensed content title	Metagenomics and industrial applications
Licensed content author	Patrick Lorenz and Jurgen Eck
Licensed content date	Jun 1, 2005
Volume number	3
Issue number	6
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables /illustrations	1
High-res required	no
Figures	Figure 1 | Multi-parameter footprint analysis
Author of this NPG article	no
Your reference number	None
Title of your thesis / dissertation	Detection of functional modules in (meta-)genomic datasets
Expected completion date	Jun 2015
Estimated size (number of pages)	200
Total	0.00 EUR
Terms and Conditions	

Terms and Conditions for Permissions

Nature Publishing Group hereby grants you a non-exclusive license to reproduce this material for this purpose, and for no other use,subject to the conditions below:

1. NPG warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to Nature Publishing Group and does not carry the copyright of another entity (as credited in the published version). If the credit line on any part of the material you have requested indicates that it was reprinted or adapted by NPG with permission from another source, then you should also seek permission from that source to reuse the material.

2. Permission granted free of charge for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to the work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.Where print permission has been granted for a fee, separate permission must be obtained for any additional, electronic re-use (unless, as in the case of a full paper, this has already been accounted for during your initial request in the calculation of a print run).NB: In all cases, web-based use of full-text articles must be authorized separately through the 'Use on a Web Site' option when requesting permission.

3. Permission granted for a first edition does not apply to second and subsequent editions and for editions in other languages (except for signatories to the STM Permissions Guidelines, or where the first edition permission was granted for free).

4. Nature Publishing Group's permission must be acknowledged next to the figure, table or abstract in print. In electronic form, this acknowledgement must be visible at the same time as the figure/table/abstract, and must be hyperlinked to the journal's homepage.

5. The credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)
   For AOP papers, the credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

   Note: For republication from the *British Journal of Cancer,* the following credit lines apply.
   Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)For AOP papers, the credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

6. Adaptations of single figures do not require NPG approval. However, the adaptation should be credited as follows:

   Adapted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)

   Note: For adaptation from the *British Journal of Cancer,* the following credit line applies.
   Adapted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

7. Translations of 401 words up to a whole article require NPG approval. Please visit http://www.macmillanmedicalcommunications.com for more information.Translations of up to a 400 words do not require NPG approval. The translation should be credited as follows:

   Translated by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication).

   Note: For translation from the *British Journal of Cancer,* the following credit line applies.
   Translated by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

We are certain that all parties will benefit from this agreement and wish you the best in the use of this material. Thank you.

Special Terms:

v1.1

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

NATURE PUBLISHING GROUP LICENSE
TERMS AND CONDITIONS

Jan 26, 2015

This is a License Agreement between Sebastian GA Konietzny ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3556491499284
License date	Jan 26, 2015
Licensed content publisher	Nature Publishing Group
Licensed content publication	Nature Chemical Biology
Licensed content title	Heterogeneity in the chemistry, structure and function of plant cell walls
Licensed content author	Rachel A Burton, Michael J Gidley, Geoffrey B Fincher
Licensed content date	Sep 17, 2010
Volume number	6
Issue number	10
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables /illustrations	1
High-res required	no
Figures	Figure 2: Heterogeneity in structures of wall polysaccharides in plants
Author of this NPG article	no
Your reference number	None
Title of your thesis / dissertation	Detection of functional modules in (meta-)genomic datasets
Expected completion date	Mar 2015
Estimated size (number of pages)	180
Total	0.00 EUR
Terms and Conditions	

Terms and Conditions for Permissions

Nature Publishing Group hereby grants you a non-exclusive license to reproduce this material for this purpose, and for no other use,subject to

the conditions below:

1. NPG warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to Nature Publishing Group and does not carry the copyright of another entity (as credited in the published version). If the credit line on any part of the material you have requested indicates that it was reprinted or adapted by NPG with permission from another source, then you should also seek permission from that source to reuse the material.

2. Permission granted free of charge for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to the work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.Where print permission has been granted for a fee, separate permission must be obtained for any additional, electronic re-use (unless, as in the case of a full paper, this has already been accounted for during your initial request in the calculation of a print run).NB: In all cases, web-based use of full-text articles must be authorized separately through the 'Use on a Web Site' option when requesting permission.

3. Permission granted for a first edition does not apply to second and subsequent editions and for editions in other languages (except for signatories to the STM Permissions Guidelines, or where the first edition permission was granted for free).

4. Nature Publishing Group's permission must be acknowledged next to the figure, table or abstract in print. In electronic form, this acknowledgement must be visible at the same time as the figure/table/abstract, and must be hyperlinked to the journal's homepage.

5. The credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)
   For AOP papers, the credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

   Note: For republication from the *British Journal of Cancer,* the following credit lines apply.
   Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)For AOP papers, the credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

6. Adaptations of single figures do not require NPG approval. However, the adaptation should be credited as follows:

   Adapted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)

   Note: For adaptation from the *British Journal of Cancer,* the following credit line applies.
   Adapted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

7. Translations of 401 words up to a whole article require NPG approval. Please visit http://www.macmillanmedicalcommunications.com for more information.Translations of up to a 400 words do not require NPG approval. The translation should be credited as follows:

   Translated by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication).

   Note: For translation from the *British Journal of Cancer,* the following credit line applies.

Translated by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK:
[JOURNAL NAME] (reference citation), copyright (year of publication)

We are certain that all parties will benefit from this agreement and wish you the best in the use of this material. Thank you.

Special Terms:

v1.1

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

NATURE PUBLISHING GROUP LICENSE
TERMS AND CONDITIONS

Mar 02, 2015

This is a License Agreement between Sebastian GA Konietzny ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3580830110051
License date	Mar 02, 2015
Licensed content publisher	Nature Publishing Group
Licensed content publication	Nature
Licensed content title	Genomics of cellulosic biofuels
Licensed content author	Edward M. Rubin
Licensed content date	Aug 14, 2008
Volume number	454
Issue number	7206
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables /illustrations	1
High-res required	no
Figures	FIGURE 1. Biology of bioconversion of solar energy into biofuels.
Author of this NPG article	no
Your reference number	None
Title of your thesis / dissertation	Detection of functional modules in (meta-)genomic datasets
Expected completion date	Mar 2015
Estimated size (number of pages)	180
Total	0.00 EUR
Terms and Conditions	

Terms and Conditions for Permissions

Nature Publishing Group hereby grants you a non-exclusive license to reproduce this material for this purpose, and for no other use,subject to the conditions below:

1. NPG warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to Nature Publishing Group and does not carry the copyright of another entity (as credited in the published version). If the credit line on any part of the material you have requested indicates that it was reprinted or adapted by NPG with permission from another source, then you should also seek permission from that source to reuse the material.

2. Permission granted free of charge for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to the work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.Where print permission has been granted for a fee, separate permission must be obtained for any additional, electronic re-use (unless, as in the case of a full paper, this has already been accounted for during your initial request in the calculation of a print run).NB: In all cases, web-based use of full-text articles must be authorized separately through the 'Use on a Web Site' option when requesting permission.

3. Permission granted for a first edition does not apply to second and subsequent editions and for editions in other languages (except for signatories to the STM Permissions Guidelines, or where the first edition permission was granted for free).

4. Nature Publishing Group's permission must be acknowledged next to the figure, table or abstract in print. In electronic form, this acknowledgement must be visible at the same time as the figure/table/abstract, and must be hyperlinked to the journal's homepage.

5. The credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)
   For AOP papers, the credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

   Note: For republication from the *British Journal of Cancer,* the following credit lines apply.
   Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)For AOP papers, the credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

6. Adaptations of single figures do not require NPG approval. However, the adaptation should be credited as follows:

   Adapted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)

   Note: For adaptation from the *British Journal of Cancer,* the following credit line applies.
   Adapted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

7. Translations of 401 words up to a whole article require NPG approval. Please visit http://www.macmillanmedicalcommunications.com for more information.Translations of up to a 400 words do not require NPG approval. The translation should be credited as follows:

   Translated by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication).

   Note: For translation from the *British Journal of Cancer,* the following credit line applies.
   Translated by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

We are certain that all parties will benefit from this agreement and wish you the best in the use of this material. Thank you.

Special Terms:

v1.1

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

Subject: FW: RE: Request for copyright permisson (ID-[837921/837921]-LJ)
From: "Thomas, Karen" <kthomas@apa.org>
Date: 04/06/2015 09:02 PM
To: "sebastian.konietzny@hhu.de" <Sebastian.Konietzny@hhu.de>

File: Konietzny, Sebastian (author)

Reproduce Figure 15, p. 235, from Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211-244. http://dx.doi.org/10.1037/0033-295X.114.2.211

Sebastian,

Thank you for contacting APA.

APA's policies on copyright and permissions can be found by visiting the Copyright and Permissions Information page located at http://www.apa.org/about/contact/copyright/index.aspx. In reading through our Policy, you will see that there are some instances under which formal APA permission is not required.

This is one of those instances. However, an appropriate credit line is required (as outlined in our Policy). The attribution and credit line requirements can be found at http://www.apa.org/about/contact/copyright/index.aspx#attribution.

I hope this helps. We appreciate your mindful concern for copyright and permissions matters.

Regards,

Karen Thomas| Permissions Supervisor
Sales, Licensing, Marketing, and Exhibitions
Publications & Databases
American Psychological Association
750 First Street NE, Washington, DC 20002-4242
Tel: 202.336.5541 | Fax: 202.336.5633
email: kthomas@apa.org | www.apa.org

From: SM PsycINFO Permissions
Sent: Monday, March 02, 2015 5:35 PM
To: Sebastian Konietzny
Cc: Permissions
Subject: RE: RE: Request for copyright permisson (ID-[837921/837921]-LJ)

Montag, 16. März 2015

Gregor Heinrich
Schumannstr. 14
64287 Darmstadt
Email: gregor@arbylon.net

Bewilligung von Nutzungsrechten

Hiermit erkläre ich, dass ich der alleinige Rechteinhaber für sämtliche Inhalte (Text und Grafik) im nachfolgend genannten technischen Report bin:

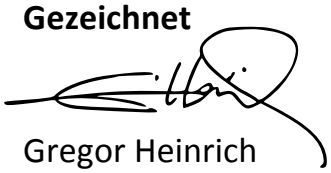
"Parameter estimation for text analysis",
Technical Report, 15. Sept. 2009.*

Mit dieser Erklärung gewähre ich Herrn Sebastian Gil Anthony Konietzny die Nutzungsrechte für die Verwendung sämtlicher Abbildungen und Tabellen aus dem Report in der von ihm voraussichtlich unter nachfolgend genanntem Titel geplanten Veröffentlichung seiner Doktorarbeit, sowohl in elektronischer wie auch in gedruckter Form:

Detection of functional modules in genomic and
metagenomic datasets

Darüber hinaus gilt diese Erklärung ebenfalls für die Verwendung von Abbildungen in Vorträgen, die mit der Doktorarbeit im Zusammenhang stehen.

Gezeichnet



Gregor Heinrich

*) URL: http://www.arbylon.net/publications/text-est2.pdf

ASSOCIATION FOR COMPUTING MACHINERY, INC. LICENSE TERMS AND CONDITIONS

Mar 16, 2015

This is a License Agreement between Sebastian GA Konietzny ("You") and Association for Computing Machinery, Inc. ("Association for Computing Machinery, Inc.") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Association for Computing Machinery, Inc., and the payment terms and conditions.

License Number	3590780688949
License date	Mar 16, 2015
Licensed content publisher	Association for Computing Machinery, Inc.
Licensed content publication	Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining
Licensed content title	Probabilistic author-topic models for information discovery
Licensed content author	Mark Steyvers, et al
Licensed content date	Aug 22, 2004
Type of Use	Thesis/Dissertation
Requestor type	Academic
Format	Print and electronic
Portion	figure/table
Number of figures/tables	1
Will you be translating?	No
Order reference number	None
Title of your thesis/dissertation	Detection of functional modules in (meta-)genomic datasets
Expected completion date	Mar 2015
Estimated size (pages)	180
Billing Type	Credit Card
Credit card info	Visa ending in 2039
Credit card expiration	10/2015
Total	7.60 EUR
Terms and Conditions	

Rightslink Terms and Conditions for ACM Material

1. The publisher of this copyrighted material is Association for Computing Machinery, Inc. (ACM). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at).

2. ACM reserves all rights not specifically granted in the combination of (i) the license

details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

3. ACM hereby grants to licensee a non-exclusive license to use or republish this ACM-copyrighted material* in secondary works (especially for commercial distribution) with the stipulation that consent of the lead author has been obtained independently. Unless otherwise stipulated in a license, grants are for one-time use in a single edition of the work, only with a maximum distribution equal to the number that you identified in the licensing process. Any additional form of republication must be specified according to the terms included at the time of licensing.

*Please note that ACM cannot grant republication or distribution licenses for embedded third-party material. You must confirm the ownership of figures, drawings and artwork prior to use.

4. Any form of republication or redistribution must be used within 180 days from the date stated on the license and any electronic posting is limited to a period of six months unless an extended term is selected during the licensing process. Separate subsidiary and subsequent republication licenses must be purchased to redistribute copyrighted material on an extranet. These licenses may be exercised anywhere in the world.

5. Licensee may not alter or modify the material in any manner (except that you may use, within the scope of the license granted, one or more excerpts from the copyrighted material, provided that the process of excerpting does not alter the meaning of the material or in any way reflect negatively on the publisher or any writer of the material).

6. Licensee must include the following copyright and permission notice in connection with any reproduction of the licensed material: "[Citation] © YEAR Association for Computing Machinery, Inc. Reprinted by permission." Include the article DOI as a link to the definitive version in the ACM Digital Library. Example: Charles, L. "How to Improve Digital Rights Management," Communications of the ACM, Vol. 51:12, © 2008 ACM, Inc. http://doi.acm.org/10.1145/nnnnnn.nnnnnn (where nnnnnn.nnnnnn is replaced by the actual number).

7. Translation of the material in any language requires an explicit license identified during the licensing process. Due to the error-prone nature of language translations, Licensee must include the following copyright and permission notice and disclaimer in connection with any reproduction of the licensed material in translation: "This translation is a derivative of ACM-copyrighted material. ACM did not prepare this translation and does not guarantee that it is an accurate copy of the originally published work. The original intellectual property contained in this work remains the property of ACM."

8. You may exercise the rights licensed immediately upon issuance of the license at the end of the licensing transaction, provided that you have disclosed complete and accurate details of your proposed use. No license is finally effective unless and until full payment is received from you (either by CCC or ACM) as provided in CCC's Billing and Payment terms and conditions.

9. If full payment is not received within 90 days from the grant of license transaction, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted.

10. Use of materials as described in a revoked license, as well as any use of the materials

beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

11. ACM makes no representations or warranties with respect to the licensed material and adopts on its own behalf the limitations and disclaimers established by CCC on its behalf in its Billing and Payment terms and conditions for this licensing transaction.

12. You hereby indemnify and agree to hold harmless ACM and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

13. This license is personal to the requestor and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.

14. This license may not be amended except in a writing signed by both parties (or, in the case of ACM, by CCC on its behalf).

15. ACM hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and ACM (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

16. This license transaction shall be governed by and construed in accordance with the laws of New York State. You hereby agree to submit to the jurisdiction of the federal and state courts located in New York for purposes of resolving any disputes that may arise in connection with this licensing transaction.

17. There are additional terms and conditions, established by Copyright Clearance Center, Inc. ("CCC") as the administrator of this licensing service that relate to billing and payment for licenses provided through this service. Those terms and conditions apply to each transaction as if they were restated here. As a user of this service, you agreed to those terms and conditions at the time that you established your account, and you may see them again at any time at http://myaccount.copyright.com

18. Thesis/Dissertation: This type of use requires only the minimum administrative fee. It is not a fee for permission. Further reuse of ACM content, by ProQuest/UMI or other document delivery providers, or in republication requires a separate permission license and fee. Commercial resellers of your dissertation containing this article must acquire a separate license.

Special Terms:

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

List of Figures

List of Tables

References

Achtman M (2012). Insights from genomic comparisons of genetically monomorphic bacterial pathogens. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, **367**(1590):860–7.
URL http://www.ncbi.nlm.nih.gov/pubmed/22312053 8

Achtman M and Wagner M (2008). Microbial diversity and the genetic nature of microbial species. *Nature Reviews Microbiology*, **6**(6):431–40.
URL http://www.ncbi.nlm.nih.gov/pubmed/18461076 8, 17, 48

Adrio JL and Demain AL (2014). Microbial enzymes: tools for biotechnological processes. *Biomolecules*, **4**(1):117–39.
URL http://www.ncbi.nlm.nih.gov/pubmed/24970208 7

Aittokallio T and Schwikowski B (2006). Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics*, **7**(3):243–55.
URL http://www.ncbi.nlm.nih.gov/pubmed/16880171 26

Altaf-Ul-Amin M, Shinbo Y, Mihara K, Kurokawa K, and Kanaya S (2006). Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics*, **7**:207.
URL http://www.ncbi.nlm.nih.gov/pubmed/16613608 36

Altincicek B, Kollas AK, Sanderbrand S, *et al.* (2001). GcpE is involved in the 2-C-methyl-D-erythritol 4-phosphate pathway of isoprenoid biosynthesis in Escherichia coli. *Journal of Bacteriology*, **183**(8):2411–6.
URL http://www.ncbi.nlm.nih.gov/pubmed/11274098 54

Altman T, Travers M, Kothari A, Caspi R, and Karp PD (2013). A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC bioinformatics*, **14**(1):112. 26

Andersson DI and Hughes D (2009). Gene amplification and adaptive evolution in bacteria. *Annual Review of Genetics*, **43**:167–95.
URL http://www.ncbi.nlm.nih.gov/pubmed/19686082 17

Andrzejewski D and Zhu X (2009). Latent Dirichlet allocation with topic-in-set knowledge. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 43–48. Association for Computational Linguistics. 144

Aravind L (2000). Guilt by association: Contextual information in genome analysis. *Genome Research*, **10**(8):1074–1077. 10, 34, 58

Arnold FH (2001). Combinatorial and computational challenges for biocatalyst design. *Nature*, **409**(6817):253–7.
URL http://www.ncbi.nlm.nih.gov/pubmed/11196654 3

Ashburner M, Ball CA, Blake JA, *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, **25**(1):25–9.
URL http://www.ncbi.nlm.nih.gov/pubmed/10802651 24, 53

Aso T and Eguchi K (2009). Predicting protein-protein relationships from literature using latent topics. *Genome Informatics*, **23**(1):3–12. 61

Aspeborg H, Coutinho PM, Wang Y, Brumer H, and Henrissat B (2012). Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). *BMC Evolutionary Biology*, **12**(1):186. 172, 173

Aziz RK, Bartels D, Best AA, *et al.* (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, **9**:75.
URL http://www.ncbi.nlm.nih.gov/pubmed/18261238 22

Bader GD, Cary MP, and Sander C (2006). Pathguide: a pathway resource list. *Nucleic Acids Research*, **34**(Database issue):D504–6.
URL http://www.ncbi.nlm.nih.gov/pubmed/16381921 25

Bansal AK (2005). Bioinformatics in microbial biotechnology–a mini review. *Microbial Cell Factories*, **4**:19.
URL http://www.ncbi.nlm.nih.gov/pubmed/15985162 8

Barabasi AL and Oltvai ZN (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, **5**(2):101–13.
URL http://www.ncbi.nlm.nih.gov/pubmed/14735121 30, 36

Barker D and Pagel M (2005). Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Computational Biology*, **1**(1):e3.
URL http://www.ncbi.nlm.nih.gov/pubmed/16103904 48, 52

Bateman A, Coin L, Durbin R, *et al.* (2004). The Pfam protein families database. *Nucleic Acids Research*, **32**(Database issue):D138–41.
URL http://www.ncbi.nlm.nih.gov/pubmed/14681378 22, 24

Bentley SD and Parkhill J (2004). Comparative genomic structure of prokaryotes. *Annual Review of Genetics*, **38**:771–92.
URL http://www.ncbi.nlm.nih.gov/pubmed/15568993 8, 16, 17

Blei DM, Ng AY, and Jordan MI (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**:993–1022. 62, 140, 141

Blouzard JC, Coutinho PM, Fierobe HP, *et al.* (2010). Modulation of cellulosome composition in Clostridium cellulolyticum: adaptation to the polysaccharide environment revealed by proteomic and carbohydrate-active enzyme analyses. *Proteomics*, **10**(3):541–554. 172

Bork P, Dandekar T, Diaz-Lazcoz Y, *et al.* (1998). Predicting function: from genes to genomes and back1. *Journal of Molecular Biology*, **283**(4):707 – 725.
URL http://www.sciencedirect.com/science/article/pii/S0022283698921441 7

Bowers PM, Cokus SJ, Eisenberg D, and Yeates TO (2004). Use of logic relationships to decipher protein network organization. *Science*, **306**(5705):2246–9.
URL http://www.ncbi.nlm.nih.gov/pubmed/15618515 46, 54, 55

Brohee S and van Helden J (2006). Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, **7**:488.
URL http://www.ncbi.nlm.nih.gov/pubmed/17087821 26, 58

Burton RA, Gidley MJ, and Fincher GB (2010). Heterogeneity in the chemistry, structure and function of plant cell walls. *Nature Chemical Biology*, **6**(10):724–732. 100

Caetano-Anolles G, Yafremava LS, Gee H, *et al.* (2009). The origin and evolution of modern metabolism. *The International Journal of Biochemistry & Cell Biology*, **41**(2):285–97.
URL http://www.ncbi.nlm.nih.gov/pubmed/18790074 51

Campillos M, von Mering C, Jensen LJ, and Bork P (2006). Identification and analysis of evolutionarily cohesive functional modules in protein networks. *Genome Research*, **16**(3):374–82.
URL http://www.ncbi.nlm.nih.gov/pubmed/16449501 38, 39, 90

Caspi R, Foerster H, Fulcher CA, *et al.* (2008). The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, **36**(Database issue):D623–31.
URL http://www.ncbi.nlm.nih.gov/pubmed/17965431 25, 26, 40

Clarke J, Wu HC, Jayasinghe L, *et al.* (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, **4**(4):265–270. 18

Cohen O, Ashkenazy H, Burstein D, and Pupko T (2012). Uncovering the co-evolutionary network among prokaryotic genes. *Bioinformatics*, **28**(18):i389–i394.
URL http://www.ncbi.nlm.nih.gov/pubmed/22962457 55

Cordero OX, Snel B, and Hogeweg P (2008). Coevolution of gene families in prokaryotes. *Genome Research*, **18**(3):462–8.
URL http://www.ncbi.nlm.nih.gov/pubmed/18230804 36

Cortes C and Vapnik V (1995). Support-vector networks. *Machine Learning*, **20**(3):273–297.
URL http://dx.doi.org/10.1007/BF00994018 154

Cunningham FX, Lafond TP, and Gantt E (2000). Evidence of a role for LytB in the nonmevalonate pathway of isoprenoid biosynthesis. *Journal of bacteriology*, **182**(20):5841–5848. 54

Dale JM, Popescu L, and Karp PD (2010). Machine learning methods for metabolic pathway prediction. *BMC Bioinformatics*, **11**:15.
URL http://www.ncbi.nlm.nih.gov/pubmed/20064214 40

Date SV and Marcotte EM (2003). Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nature Biotechnology*, **21**(9):1055–62.
URL http://www.ncbi.nlm.nih.gov/pubmed/12923548 10, 52, 54, 55, 56

De Filippo C, Ramazzotti M, Fontana P, and Cavalieri D (2012). Bioinformatic approaches for functional annotation and pathway inference in metagenomics data.

Briefings in Bioinformatics, **13**(6):696–710.
URL http://www.ncbi.nlm.nih.gov/pubmed/23175748 10

Di Ventura B, Lemerle C, Michalodimitrakis K, and Serrano L (2006). From in vivo to in silico biology and back. *Nature*, **443**(7111):527–33.
URL http://www.ncbi.nlm.nih.gov/pubmed/17024084 21

Dröge J and McHardy AC (2012). Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Briefings in Bioinformatics*, **13**(6):646–655.
URL http://bib.oxfordjournals.org/content/13/6/646.abstract 20

Duan CJ and Feng JX (2010). Mining metagenomes for novel cellulase genes. *Biotechnology Letters*, **32**(12):1765–1775.
URL http://dx.doi.org/10.1007/s10529-010-0356-z 98

Duan J, Jiang W, Cheng Z, Heikkila JJ, and Glick BR (2013). The complete genome sequence of the plant growth-promoting bacterium Pseudomonas sp. UW4. *PLoS One*, **8**(3):e58640. 21

Durot M, Bourguignon PY, and Schachter V (2009). Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiology Reviews*, **33**(1):164–90.
URL http://www.ncbi.nlm.nih.gov/pubmed/19067749 34

Dykhuizen D (2005). Species Numbers in Bacteria. *Proceedings of the California Academy of Sciences*, **56**(6 Suppl 1):62–71.
URL http://www.ncbi.nlm.nih.gov/pubmed/21874075 9

Enright AJ, Iliopoulos I, Kyrpides NC, and Ouzounis CA (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**(6757):86–90.
URL http://www.ncbi.nlm.nih.gov/pubmed/10573422 41

Ferrer L, Dale JM, and Karp PD (2010). A systematic study of genome context methods: calibration, normalization and combination. *BMC Bioinformatics*, **11**:493.
URL http://www.ncbi.nlm.nih.gov/pubmed/20920312 52, 54

Fondi M, Emiliani G, and Fani R (2009). Origin and evolution of operons and metabolic pathways. *Research in Microbiology*, **160**(7):502–12.
URL http://www.ncbi.nlm.nih.gov/pubmed/19465116 17, 39

Friedberg I (2006). Automated protein function prediction–the genomic challenge. *Briefings in Bioinformatics*, **7**(3):225–42.
URL http://www.ncbi.nlm.nih.gov/pubmed/16772267 23, 28

Friedman N (2004). Inferring cellular networks using probabilistic graphical models. *Science*, **303**(5659):799–805. 49

Gaasterland T and Ragan MA (1998). Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microbial and Comparative Genomics*, **3**(4):199–217.
URL http://www.ncbi.nlm.nih.gov/pubmed/10027190 45

Gil R, Silva FJ, Pereto J, and Moya A (2004). Determination of the core of a minimal bacterial gene set. *Microbiology and Molecular Biology Reviews*, **68**(3):518–37, table of contents.
URL http://www.ncbi.nlm.nih.gov/pubmed/15353568 17, 18, 19, 20, 281

Gilks WR, Richardson S, and Spiegelhalter DJ (1999). *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, Boca Raton, Florida, USA. 63

Gillis J and Pavlidis P (2012). "Guilt by association" is the exception rather than the rule in gene networks. *PLoS Computational Biology*, **8**(3):e1002444.
URL http://www.ncbi.nlm.nih.gov/pubmed/22479173 58

Glazko GV and Mushegian AR (2004). Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. *Genome Biology*,

5(5):R32.
URL http://www.ncbi.nlm.nih.gov/pubmed/15128446 38, 48, 55, 56, 57

Godzik A (2011). Metagenomics and the protein universe. *Current Opinion in Structural Biology*, **21**(3):398–403.
URL http://www.ncbi.nlm.nih.gov/pubmed/21497084 27

Goh CS, Bogan AA, Joachimiak M, Walther D, and Cohen FE (2000). Co-evolution of proteins with their interaction partners. *Journal of Molecular Biology*, **299**(2):283–93.
URL http://www.ncbi.nlm.nih.gov/pubmed/10860738 46

Green ML and Karp PD (2006). The outcomes of pathway database computations depend on pathway ontology. *Nucleic Acids Research*, **34**(13):3687–97.
URL http://www.ncbi.nlm.nih.gov/pubmed/16893953 26, 27

Griffiths TL, Steyvers M, and Tenenbaum JB (2007). Topics in semantic representation. *Psychological Review*, **114**(2):211. 87, 88, 89

Hartwell LH, Hopfield JJ, Leibler S, and Murray AW (1999). From molecular to modular cell biology. *Nature*, **402**(6761 Suppl):C47–C52.
URL http://dx.doi.org/10.1038/35011540 7, 36

Hawkins RD, Hon GC, and Ren B (2010). Next-generation genomics: an integrative approach. *Nature Reviews Genetics*, **11**(7):476–86.
URL http://www.ncbi.nlm.nih.gov/pubmed/20531367 4

Heinrich G (2009). *Technical report: Parameter estimation for text analysis*. Fraunhofer IGD, Darmstadt, Germany.
URL http://faculty.cs.byu.edu/~ringger/CS601R/papers/Heinrich-GibbsLDA.pdf 63, 183, 185, 186, 187

Henry C, Overbeek R, and Stevens RL (2010). Building the blueprint of life. *Biotechnology Journal*, **5**(7):695–704.
URL http://www.ncbi.nlm.nih.gov/pubmed/20665643 17, 18, 39

Hieter P and Boguski M (1997). Functional genomics: it's all how you read it. *Science*, **278**(5338):601–2.
URL http://www.ncbi.nlm.nih.gov/pubmed/9381168 7

Horner-Devine MC, Carney KM, and Bohannan BJ (2004). An ecological perspective on bacterial biodiversity. *Proceedings of the Royal Society of London B: Biological Sciences*, **271**(1535):113–22.
URL http://www.ncbi.nlm.nih.gov/pubmed/15058386 8

Hugenholtz P (2002). Exploring prokaryotic diversity in the genomic era. *Genome Biology*, **3**(2):REVIEWS0003.
URL http://www.ncbi.nlm.nih.gov/pubmed/11864374 9

Huisman GW and Gray D (2002). Towards novel processes for the fine-chemical and pharmaceutical industries. *Current Opinion in Biotechnology*, **13**(4):352–8.
URL http://www.ncbi.nlm.nih.gov/pubmed/12323358 3

Huynen M, Snel B, Lathe r W, and Bork P (2000). Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Research*, **10**(8):1204–10.
URL http://www.ncbi.nlm.nih.gov/pubmed/10958638 10, 58

Janga SC, Diaz-Mejia JJ, and Moreno-Hagelsieb G (2011). Network-based function prediction and interactomics: the case for metabolic enzymes. *Metabolic Engineering*, **13**(1):1–10.
URL http://www.ncbi.nlm.nih.gov/pubmed/20654726 58, 174

Jaroszewski L, Li Z, Krishna SS, *et al.* (2009). Exploration of uncharted regions of the protein universe. *PLoS Biology*, **7**(9):e1000205.
URL http://www.ncbi.nlm.nih.gov/pubmed/19787035 32

Jarrell KA (2009). Synthetic biology and the sustainable chemistry revolution. *Industrial Biotechnology*, **5**(4):210–212. 7

Jeffery CJ (1999). Moonlighting proteins. *Trends in Biochemical Sciences*, **24**(1):8–11. URL http://www.ncbi.nlm.nih.gov/pubmed/10087914 25, 51

Jensen LJ, Julien P, Kuhn M, *et al.* (2008). eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Research*, **36**(Database issue):D250–4. URL http://www.ncbi.nlm.nih.gov/pubmed/17942413 24, 50

Jothi R, Przytycka TM, and Aravind L (2007). Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. *BMC Bioinformatics*, **8**:173. URL http://www.ncbi.nlm.nih.gov/pubmed/17521444 48, 49, 52

Kanehisa M, Araki M, Goto S, *et al.* (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, **36**(Database issue):D480–4. URL http://www.ncbi.nlm.nih.gov/pubmed/18077471 26

Kanehisa M, Goto S, Kawashima S, Okuno Y, and Hattori M (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Research*, **32**(Database issue):D277–80. URL http://www.ncbi.nlm.nih.gov/pubmed/14681412 24, 25, 56

Karimpour-Fard A, Leach SM, Gill RT, and Hunter LE (2008). Predicting protein linkages in bacteria: which method is best depends on task. *BMC Bioinformatics*, **9**:397. URL http://www.ncbi.nlm.nih.gov/pubmed/18816389 42, 54

Karr J, Sanghvi J, Macklin D, *et al.* (2012). A whole-cell computational model predicts phenotype from genotype. *Cell*, **150**(2):389 – 401. URL http://www.sciencedirect.com/science/article/pii/S0092867412007763 35

Kastenmüller G, Schenk ME, Gasteiger J, and Mewes HW (2009). Uncovering metabolic pathways relevant to phenotypic traits of microbial genomes. *Genome Biology*, **10**(3):R28. 100

Kell DB (2004). Metabolomics and systems biology: making sense of the soup. *Current Opinion in Microbiology*, **7**(3):296–307.
URL http://www.ncbi.nlm.nih.gov/pubmed/15196499 23

Kensche PR, van Noort V, Dutilh BE, and Huynen MA (2008). Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. *Journal of the Royal Society Interface*, **5**(19):151–70.
URL http://www.ncbi.nlm.nih.gov/pubmed/17535793 45, 46, 47, 48, 50, 54, 57

Kepes F, Jester BC, Lepage T, *et al.* (2012). The layout of a bacterial genome. *FEBS Letters*, **586**(15):2043–8.
URL http://www.ncbi.nlm.nih.gov/pubmed/22483986 17

Keseler IM, Collado-Vides J, Santos-Zavaleta A, *et al.* (2011). EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Research*, **39**(Database issue):D583–90.
URL http://www.ncbi.nlm.nih.gov/pubmed/21097882 26

Klemm K and Bornholdt S (2005). Topology of biological networks and reliability of information processing. *Proceedings of the National Academy of Sciences*, **102**(51):18414–9.
URL http://www.ncbi.nlm.nih.gov/pubmed/16339314 26

Kohse-Hinghaus K, Owald P, Cool T, *et al.* (2010). Biofuel Combustion Chemistry: From Ethanol to Biodiesel. *Angewandte Chemie International Edition*, **49**(21):3572–3597.
URL http://dx.doi.org/10.1002/anie.200905335 98

Koller D and Friedman N (2009). *Probabilistic graphical models: principles and techniques*. MIT press. 62

Konietzny SG, Dietz L, and McHardy AC (2011). Inferring functional modules of protein families with probabilistic topic models. *BMC Bioinformatics*, **12**(1):141. viii, 5, 6, 67, 87, 88, 91, 142, 149, 154, 162, 170, 198

Konietzny SG, Pope PB, Weimann A, and McHardy AC (2014). Inference of phenotype-defining functional modules of protein families for microbial plant biomass degraders. *Biotechnology for Biofuels*, **7**(1):124. viii, 5, 143, 154, 155, 163, 170, 172, 216, 233

Koonin EV, Mushegian AR, and Bork P (1996). Non-orthologous gene displacement. *Trends in Genetics*, **12**(9):334–6.
URL http://www.ncbi.nlm.nih.gov/pubmed/8855656 51

Kreimer A, Borenstein E, Gophna U, and Ruppin E (2008). The evolution of modularity in bacterial metabolic networks. *Proceedings of the National Academy of Sciences*, **105**(19):6976–81.
URL http://www.ncbi.nlm.nih.gov/pubmed/18460604 36

Krieger CJ, Zhang P, Mueller LA, *et al.* (2004). MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Research*, **32**(Database issue):D438–42.
URL http://www.ncbi.nlm.nih.gov/pubmed/14681452 26

Kumar R, Singh S, and Singh OV (2008). Bioconversion of lignocellulosic biomass: biochemical and molecular perspectives. *Journal of Industrial Microbiology & Biotechnology*, **35**(5):377–391. 98

Kyrpides NC (2009). Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream. *Nature Biotechnology*, **27**(7):627–32.
URL http://www.ncbi.nlm.nih.gov/pubmed/19587669 8

Landauer TK and Dumais ST (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, **104**(2):211. 5

Landauer TK, Foltz PW, and Laham D (1998). An introduction to latent semantic analysis. *Discourse Processes*, **25**(2-3):259–284. 5

Lawrence JG (1999). Gene transfer, speciation, and the evolution of bacterial genomes. *Current Opinion in Microbiology*, **2**(5):519 – 523.
URL http://www.sciencedirect.com/science/article/pii/S1369527499000107 7

Lawrence JG and Roth JR (1996). Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics*, **143**(4):1843–60.
URL http://www.ncbi.nlm.nih.gov/pubmed/8844169 39

Lee D, Redfern O, and Orengo C (2007). Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology*, **8**(12):995–1005.
URL http://www.ncbi.nlm.nih.gov/pubmed/18037900 28, 41, 45

Levitt M (2009). Nature of the protein universe. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(27):11079–84.
URL http://www.ncbi.nlm.nih.gov/pubmed/19541617 27

Li H, Kristensen DM, Coleman MK, and Mushegian A (2009). Detection of biochemical pathways by probabilistic matching of phyletic vectors. *PLoS ONE*, **4**(4):e5326.
URL http://www.ncbi.nlm.nih.gov/pubmed/19390636 55, 56

Liberles DA, Thorn A, Heijne G, and Elofsson A (2002). The use of phylogenetic profiles for gene predictions. *Current Genomics*, **3**(3):131–137. 46

Lingner T, Mühlhausen S, Gabaldón T, Notredame C, and Meinicke P (2010). Predicting phenotypic traits of prokaryotes from protein domain frequencies. *BMC Bioinformatics*, **11**(1):481. 100

Liolios K, Chen IMA, Mavromatis K, *et al.* (2010). The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, **38**(suppl 1):D346–D354.
URL http://nar.oxfordjournals.org/content/38/suppl_1/D346.abstract 142

Liu Y, Li J, Sam L, *et al.* (2006). An integrative genomic approach to uncover molecular mechanisms of prokaryotic traits. *PLoS Computational Biology*, **2**(11):e159. URL http://www.ncbi.nlm.nih.gov/pubmed/17112314 47

Lorenz P and Eck J (2005). Metagenomics and industrial applications. *Nature Reviews Microbiology*, **3**(6):510–516. 7, 31

Luisi PL, Oberholzer T, and Lazcano A (2002). The notion of a DNA minimal cell: a general discourse and some guidelines for an experimental approach. *Helvetica Chimica Acta*, **85**(6):1759–1777. 17

Mao X, Cai T, Olyarchuk JG, and Wei L (2005). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics*, **21**(19):3787–93. URL http://www.ncbi.nlm.nih.gov/pubmed/15817693 24, 53

Marcotte EM (2000). Computational genetics: finding protein function by nonhomology methods. *Current Opinion in Structural Biology*, **10**(3):359–65. URL http://www.ncbi.nlm.nih.gov/pubmed/10851184 46

Mardis ER (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, **24**(3):133–41. URL http://www.ncbi.nlm.nih.gov/pubmed/18262675 4, 9

Mardis ER (2011). A decade's perspective on DNA sequencing technology. *Nature*, **470**(7333):198–203. URL http://www.ncbi.nlm.nih.gov/pubmed/21307932 4, 9

Markowitz VM, Chen IM, Palaniappan K, *et al.* (2012). IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research*, **40**(Database issue):D115–22. URL http://www.ncbi.nlm.nih.gov/pubmed/22194640 22

Markowitz VM, Ivanova NN, Szeto E, *et al.* (2008). IMG/M: a data management and analysis system for metagenomes. *Nucleic acids research*, **36**(suppl 1):D534–D538. 22

Martin MJ, Herrero J, Mateos A, and Dopazo J (2003). Comparing bacterial genomes through conservation profiles. *Genome Research*, **13**(5):991–8.
URL http://www.ncbi.nlm.nih.gov/pubmed/12695324 36

McAuliffe JD and Blei DM (2008). Supervised Topic Models. In Platt J, Koller D, Singer Y, and Roweis S, eds., *Advances in Neural Information Processing Systems 20*, pages 121–128. Curran Associates, Inc.
URL http://papers.nips.cc/paper/3328-supervised-topic-models.pdf 139

McCarthy A (2010). Third generation DNA sequencing: pacific biosciences' single molecule real time technology. *Chemistry & Biology*, **17**(7):675–676. 18

McHardy AC and Rigoutsos I (2007). What's in the mix: phylogenetic classification of metagenome sequence samples. *Current Opinion in Microbiology*, **10**(5):499 – 503. Antimicrobials/Genomics.
URL http://www.sciencedirect.com/science/article/pii/S136952740700118X 20

Medini D, Donati C, Tettelin H, Masignani V, and Rappuoli R (2005). The microbial pan-genome. *Current opinion in genetics & development*, **15**(6):589–594. 17

Medini D, Serruto D, Parkhill J, *et al.* (2008). Microbiology in the post-genomic era. *Nature Reviews Microbiology*, **6**(6):419–30.
URL http://www.ncbi.nlm.nih.gov/pubmed/18475305 8

von Mering C, Huynen M, Jaeggi D, *et al.* (2003a). STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research*, **31**(1):258–61.
URL http://www.ncbi.nlm.nih.gov/pubmed/12519996 43, 52

von Mering C, Jensen LJ, Snel B, *et al.* (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, **33**(Database issue):D433–7.
URL http://www.ncbi.nlm.nih.gov/pubmed/15608232 43

von Mering C, Zdobnov EM, Tsoka S, *et al.* (2003b). Genome evolution reveals biochemical networks and functional modules. *Proceedings of the National Academy of Sciences*, **100**(26):15428–33.
URL http://www.ncbi.nlm.nih.gov/pubmed/14673105 42

Meyer F, Goesmann A, McHardy AC, *et al.* (2003). GenDB–an open source genome annotation system for prokaryote genomes. *Nucleic Acids Research*, **31**(8):2187–95.
URL http://www.ncbi.nlm.nih.gov/pubmed/12682369 22

Meyer F, Paarmann D, D'Souza M, *et al.* (2008). The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**(1):386. 22

Mikolov T, Yih Wt, and Zweig G (2013). Linguistic Regularities in Continuous Space Word Representations. In *HLT-NAACL*, pages 746–751. 171

Minguez P and Dopazo J (2010). Functional genomics and networks: new approaches in the extraction of complex gene modules. *Expert Review of Proteomics*, **7**(1):55–63.
URL http://www.ncbi.nlm.nih.gov/pubmed/20121476 25

Moore GE *et al.* (1998). Cramming more components onto integrated circuits. *Proceedings of the IEEE*, **86**(1):82–85. 171

Muley VY and Ranjan A (2012). Effect of reference genome selection on the performance of computational methods for genome-wide protein-protein interaction prediction. *PLoS ONE*, **7**(7):e42057.
URL http://www.ncbi.nlm.nih.gov/pubmed/22844541 49

Muley VY and Ranjan A (2013). Evaluation of physical and functional protein-protein interaction prediction methods for detecting biological pathways. *PLoS ONE*, **8**(1):e54325.
URL http://www.ncbi.nlm.nih.gov/pubmed/23349851 11, 42, 52, 54, 55, 56

Newman D, Asuncion A, Smyth P, and Welling M (2009). Distributed algorithms for topic models. *The Journal of Machine Learning Research*, **10**:1801–1828. 171

van Noort V, Snel B, and Huynen MA (2003). Predicting gene function by conserved co-expression. *Trends in Genetics*, **19**(5):238–42.
URL http://www.ncbi.nlm.nih.gov/pubmed/12711213 42

O'Driscoll A, Daugelaite J, and Sleator RD (2013). Big data, Hadoop and cloud computing in genomics. *Journal of Biomedical Informatics*, **46**(5):774 – 781.
URL http://www.sciencedirect.com/science/article/pii/S1532046413001007 171

Ogawa J and Shimizu S (1999). Microbial enzymes: new industrial applications from traditional screening methods. *Trends in Biotechnology*, **17**(1):13–21.
URL http://www.ncbi.nlm.nih.gov/pubmed/10098273 3

Oliver S (2000). Guilt-by-association goes global. *Nature*, **403**(6770):601–3.
URL http://www.ncbi.nlm.nih.gov/pubmed/10688178 58

Osterman A and Overbeek R (2003). Missing genes in metabolic pathways: a comparative genomics approach. *Current Opinion in Chemical Biology*, **7**(2):238–51.
URL http://www.ncbi.nlm.nih.gov/pubmed/12714058 58

Overbeek R, Fonstein M, DSouza M, Pusch GD, and Maltsev N (1999). The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences*, **96**(6):2896–2901.
URL http://www.pnas.org/content/96/6/2896.abstract 41

Pagel P, Wong P, and Frishman D (2004). A domain interaction map based on phylogenetic profiling. *Journal of Molecular Biology*, **344**(5):1331–46.
URL http://www.ncbi.nlm.nih.gov/pubmed/15561146 47

Paley SM and Karp PD (2002). Evaluation of computational metabolic-pathway predictions for Helicobacter pylori. *Bioinformatics*, **18**(5):715–24.
URL http://www.ncbi.nlm.nih.gov/pubmed/12050068 40

Papin JA, Stelling J, Price ND, *et al.* (2004). Comparison of network-based pathway analysis methods. *Trends in Biotechnology*, **22**(8):400–5.
URL http://www.ncbi.nlm.nih.gov/pubmed/15283984 26

Parro V, Moreno-Paz M, and Gonzalez-Toril E (2007). Analysis of environmental transcriptomes by DNA microarrays. *Environmental Microbiology*, **9**(2):453–64.
URL http://www.ncbi.nlm.nih.gov/pubmed/17222143 11

Pazos F, Ranea JA, Juan D, and Sternberg MJ (2005). Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *Journal of Molecular Biology*, **352**(4):1002–15.
URL http://www.ncbi.nlm.nih.gov/pubmed/16139301 46

Pazos F and Valencia A (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering*, **14**(9):609–14.
URL http://www.ncbi.nlm.nih.gov/pubmed/11707606 46

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, and Yeates TO (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(8):4285–8.
URL http://www.ncbi.nlm.nih.gov/pubmed/10200254 11, 41, 45

Peregrin-Alvarez JM, Tsoka S, and Ouzounis CA (2003). The phylogenetic extent of metabolic enzymes and pathways. *Genome Research*, **13**(3):422–7.
URL http://www.ncbi.nlm.nih.gov/pubmed/12618373 37, 38, 39

Pereira-Leal JB, Enright AJ, and Ouzounis CA (2004). Detection of functional modules from protein interaction networks. *Proteins*, **54**(1):49–57.
URL http://www.ncbi.nlm.nih.gov/pubmed/14705023 36

Porteous I, Newman D, Ihler A, *et al.* (2008). Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 569–577. ACM. 189

Qi Y, Bar-Joseph Z, and Klein-Seetharaman J (2006). Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, **63**(3):490–500.
URL http://www.ncbi.nlm.nih.gov/pubmed/16450363 51, 52

Radivojac P, Clark WT, Oron TR, *et al.* (2013). A large-scale evaluation of computational protein function prediction. *Nature Methods*, **10**(3):221–7.
URL http://www.ncbi.nlm.nih.gov/pubmed/23353650 24, 28, 29, 31

Raes J and Bork P (2008). Molecular eco-systems biology: towards an understanding of community function. *Nature Reviews Microbiology*, **6**(9):693–9.
URL http://www.ncbi.nlm.nih.gov/pubmed/18587409 35

Raman K and Chandra N (2009). Flux balance analysis of biological systems: applications and challenges. *Briefings in Bioinformatics*, **10**(4):435–49.
URL http://www.ncbi.nlm.nih.gov/pubmed/19287049 23, 34, 35

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, and Barabasi AL (2002). Hierarchical organization of modularity in metabolic networks. *Science*, **297**(5586):1551–5.
URL http://www.ncbi.nlm.nih.gov/pubmed/12202830 36

Reed JL, Famili I, Thiele I, and Palsson BO (2006). Towards multidimensional genome annotation. *Nature Reviews Genetics*, **7**(2):130–41.
URL http://www.ncbi.nlm.nih.gov/pubmed/16418748 22

Rinke C, Schwientek P, Sczyrba A, *et al.* (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, **499**(7459):431–7.
URL http://www.ncbi.nlm.nih.gov/pubmed/23851394 8

Rives AW and Galitski T (2003). Modular organization of cellular networks. *Proceedings of the National Academy of Sciences*, **100**(3):1128–33.
URL http://www.ncbi.nlm.nih.gov/pubmed/12538875 36

Rocap G, Larimer FW, Lamerdin J, *et al.* (2003). Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature*, **424**(6952):1042–1047. 7

Rocha EP (2008). The organization of the bacterial genome. *Annual Review of Genetics*, **42**:211–33.
URL http://www.ncbi.nlm.nih.gov/pubmed/18605898 16, 17, 39, 48

Rosen-Zvi M, Griffiths T, Steyvers M, and Smyth P (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press. 139, 140, 141, 143

Roth C, Rastogi S, Arvestad L, *et al.* (2007). Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *Journal of Experimental Zoology, Part B: Molecular and Developmental Evolution*, **308**(1):58–73.
URL http://www.ncbi.nlm.nih.gov/pubmed/16838295 50

Rubin EM (2008). Genomics of cellulosic biofuels. *Nature*, **454**(7206):841–845.
URL http://dx.doi.org/10.1038/nature07190 98, 99, 149

Saito N, Ohashi Y, Soga T, and Tomita M (2010). Unveiling cellular biochemical reactions via metabolomics-driven approaches. *Current Opinion in Microbiology*, **13**(3):358–62.
URL http://www.ncbi.nlm.nih.gov/pubmed/20430690 10, 32

Schloss PD and Handelsman J (2004). Status of the microbial census. *Microbiology and Molecular Biology Reviews*, **68**(4):686–91.
URL http://www.ncbi.nlm.nih.gov/pubmed/15590780 9

Schmidhuber J (2015). Deep learning in neural networks: An overview. *Neural Networks*, **61**:85–117. 171

Schneider A, Seidl MF, and Snel B (2013). Shared protein complex subunits contribute to explaining disrupted co-occurrence. *PLoS Computational Biology*, **9**(7):e1003124.
URL http://www.ncbi.nlm.nih.gov/pubmed/23874172 54, 55, 93

Schwikowski B, Uetz P, and Fields S (2000). A network of protein-protein interactions in yeast. *Nature Biotechnology*, **18**(12):1257–1261. 58

Seshadri R, Kravitz SA, Smarr L, Gilna P, and Frazier M (2007). CAMERA: a community resource for metagenomics. *PLoS Biology*, **5**(3):e75. 22

Simon C and Daniel R (2011). Metagenomic analyses: past and future trends. *Applied and Environmental Microbiology*, **77**(4):1153–61.
URL http://www.ncbi.nlm.nih.gov/pubmed/21169428 11, 42

Singh BK (2010). Exploring microbial diversity for biotechnology: the way forward. *Trends in Biotechnology*, **28**(3):111–6.
URL http://www.ncbi.nlm.nih.gov/pubmed/20005589 28

Slonim N, Elemento O, and Tavazoie S (2006). Ab initio genotype-phenotype association reveals intrinsic modularity in genetic networks. *Molecular Systems Biology*, **2**(2006.0005):1–14.
URL http://www.ncbi.nlm.nih.gov/pubmed/16732191 47

Snel B, Bork P, and Huynen MA (2002). Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Research*, **12**(1):17–25.
URL http://www.ncbi.nlm.nih.gov/pubmed/11779827 48

Snel B and Huynen MA (2004). Quantifying modularity in the evolution of biomolecular systems. *Genome Research*, **14**(3):391–7.
URL http://www.ncbi.nlm.nih.gov/pubmed/14993205 36, 37, 38, 51

Snitkin ES, Gustafson AM, Mellor J, Wu J, and DeLisi C (2006). Comparative assessment of performance and genome dependence among phylogenetic profiling methods. *BMC Bioinformatics*, **7**:420.
URL http://www.ncbi.nlm.nih.gov/pubmed/17005048 57

Sonnhammer EL and Koonin EV (2002). Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics*, **18**(12):619–20.
URL http://www.ncbi.nlm.nih.gov/pubmed/12446146 50

Spirin V, Gelfand MS, Mironov AA, and Mirny LA (2006). A metabolic network in the evolutionary context: multiscale structure and modularity. *Proceedings of the National Academy of Sciences*, **103**(23):8774–9.
URL http://www.ncbi.nlm.nih.gov/pubmed/16731630 39

Stein L (2001). Genome annotation: from sequence to biology. *Nature Reviews Genetics*, **2**(7):493–503.
URL http://www.ncbi.nlm.nih.gov/pubmed/11433356 21, 22

Steyvers M and Griffiths T (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, **427**(7):424–440. 62

Steyvers M, Smyth P, Rosen-Zvi M, and Griffiths T (2004). Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315. ACM. 141

Sun J, Xu J, Liu Z, *et al.* (2005). Refined phylogenetic profiles method for predicting protein-protein interactions. *Bioinformatics*, **21**(16):3409–15. 48, 49

Tatusov RL, Galperin MY, Natale DA, and Koonin EV (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, **28**(1):33–6.
URL http://www.ncbi.nlm.nih.gov/pubmed/10592175 52

Tatusov RL, Koonin EV, and Lipman DJ (1997). A genomic perspective on protein families. *Science*, **278**(5338):631–7.
URL http://www.ncbi.nlm.nih.gov/pubmed/9381173 24, 50

Teeling H and Glockner FO (2012). Current opportunities and challenges in microbial metagenome analysis–a bioinformatic perspective. *Briefings in Bioinformatics*, **13**(6):728–42.
URL http://www.ncbi.nlm.nih.gov/pubmed/22966151 28

Tettelin H, Riley D, Cattuto C, and Medini D (2008). Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology*, **11**(5):472–7.
URL http://www.ncbi.nlm.nih.gov/pubmed/19086349 17

Tian W and Skolnick J (2003). How well is enzyme function conserved as a function of pairwise sequence identity? *Journal of Molecular Biology*, **333**(4):863–82.
URL http://www.ncbi.nlm.nih.gov/pubmed/14568541 45

Tomita M (2001). Whole-cell simulation: a grand challenge of the 21st century. *Trends in Biotechnology*, **19**(6):205 – 210.
URL http://www.sciencedirect.com/science/article/pii/S0167779901016365 35

Valencia A (2005). Automatic annotation of protein function. *Current Opinion in Structural Biology*, **15**(3):267–74.
URL http://www.ncbi.nlm.nih.gov/pubmed/15922590 22

Vieites JM, Guazzaroni ME, Beloqui A, Golyshin PN, and Ferrer M (2009). Metagenomics approaches in systems microbiology. *FEMS Microbiology Reviews*, **33**(1):236–

55.

URL http://www.ncbi.nlm.nih.gov/pubmed/19054115 35

Wagner GP, Pavlicev M, and Cheverud JM (2007). The road to modularity. *Nature Reviews Genetics*, **8**(12):921–31.

URL http://www.ncbi.nlm.nih.gov/pubmed/18007649 36

Walsh C (2001). Enabling the chemistry of life. *Nature*, **409**(6817):226–31.

URL http://www.ncbi.nlm.nih.gov/pubmed/11196650 22

Warnecke F and Hugenholtz P (2007). Building on basic metagenomics with complementary technologies. *Genome Biology*, **8**(12):231.

URL http://www.ncbi.nlm.nih.gov/pubmed/18177506 11

Weimann A, Trukhina Y, Pope PB, Konietzny SG, and McHardy AC (2013). De novo prediction of the genomic components and capabilities for microbial plant biomass degradation from (meta-) genomes. *Biotechnology for Biofuels*, **6**(1):24. 153, 154, 163

Wilkinson DJ (2007). Bayesian methods in bioinformatics and computational systems biology. *Briefings in Bioinformatics*, **8**(2):109–116. 49

Wilson DB (2011). Microbial diversity of cellulose hydrolysis. *Current Opinion in Microbiology*, **14**(3):259–263. 98, 149, 172

Wilson MC and Piel J (2013). Metagenomic approaches for exploiting uncultivated bacteria as a resource for novel biosynthetic enzymology. *Chemistry and Biology*, **20**(5):636–47.

URL http://www.ncbi.nlm.nih.gov/pubmed/23706630 28

Wooley JC, Godzik A, and Friedberg I (2010). A primer on metagenomics. *PLoS Computational Biology*, **6**(2):e1000667.

URL http://www.ncbi.nlm.nih.gov/pubmed/20195499 18

Wren BW (2000). Microbial genome analysis: insights into virulence, host adaptation

and evolution. *Nature Reviews Genetics*, **1**(1):30–9.
URL http://www.ncbi.nlm.nih.gov/pubmed/11262871 8

Wu J, Hu Z, and DeLisi C (2006). Gene annotation and network inference by phylogenetic profiling. *BMC Bioinformatics*, **7**:80.
URL http://www.ncbi.nlm.nih.gov/pubmed/16503966 52, 55

Xiao H and Stibor T (2010). Efficient Collapsed Gibbs Sampling for Latent Dirichlet Allocation. In *ACML*, pages 63–78. 171, 189

Yamada T, Kanehisa M, and Goto S (2006). Extraction of phylogenetic network modules from the metabolic network. *BMC Bioinformatics*, **7**:130.
URL http://www.ncbi.nlm.nih.gov/pubmed/16533389 37, 38, 39, 54, 55, 56

Yan F, Xu N, and Qi Y (2009). Parallel Inference for Latent Dirichlet Allocation on Graphics Processing Units. In Bengio Y, Schuurmans D, Lafferty J, Williams C, and Culotta A, eds., *Advances in Neural Information Processing Systems 22*, pages 2134–2142. Curran Associates, Inc.
URL http://papers.nips.cc/paper/3788-parallel-inference-for-latent-diric
pdf 171

Yanai I and DeLisi C (2002). The society of genes: networks of functional links between genes from comparative genomics. *Genome Biology*, **3**(11):research0064.
URL http://www.ncbi.nlm.nih.gov/pubmed/12429063 46

Ye Y, Osterman A, Overbeek R, and Godzik A (2005). Automatic detection of subsystem/pathway variants in genome analysis. *Bioinformatics*, **21 Suppl 1**:i478–86.
URL http://www.ncbi.nlm.nih.gov/pubmed/15961494 51

Zhai K, Boyd-Graber J, Asadi N, and Alkhouja ML (2012). Mr. LDA: A flexible large scale topic modeling package using variational inference in mapreduce. In *Proceedings of the 21st international conference on World Wide Web*, pages 879–888. ACM. 171

Zhang X, Kim S, Wang T, and Baral C (2006). Joint learning of logic relationships for studying protein function using phylogenetic profiles and the Rosetta Stone method. *IEEE Transactions on: Signal Processing*, **54**(6):2427–2435. 46, 54, 55

Zheng B, McLean DC, and Lu X (2006). Identifying biological concepts from a protein-related corpus with a probabilistic topic model. *BMC Bioinformatics*, **7**(1):58. 61

Zhu J, Ahmed A, and Xing EP (2012). MedLDA: maximum margin supervised topic models. *The Journal of Machine Learning Research*, **13**(1):2237–2278. 139