Greedy phylogeny-based orthology assignment and its application to the evolutionary analysis of metabolic coupling

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Sabine Anita Christiane Thuß
aus Zwickau

Düsseldorf, April 2011

aus dem Institut für
der Heinrich-Heine Universität Düsseldorf

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

Referent: Prof. Dr. Martin Lercher
Koreferent: Prof. Dr. William Martin

Tag der mündlichen Prüfung: 27. 05. 2011

Declaration

This thesis is submitted for the degree of Doctor rerum naturalium at the Heinrich-Heine-University Düsseldorf. It has not been submitted to any other university for a degree. I agree that the University library may lend out or copy this thesis freely.

Sabine Anita Thuß.
April, 2011.

Acknowledgements:

At first I would like to thank my supervisor Prof. Dr. Martin Lercher for giving me the opportunity to work in my favourite scientific field of evolutionary biology in combination with bioinformatics. Thank you, Martin, for the scientific support over all this time and the interesting and motivating discussions we had.

Additionally, I wish to thank Prof. Dr. William Martin for reading and evaluating my thesis as a second reviewer.

I also wish to thank all my colleagues in the Martin Lercher lab, past and present, for the friendly and helpful atmosphere and great discussions: Gabriel Gelius-Dietrich, Christian Eßer, Wolfgang Kaisers, Wei-Hua Chen, Jan Wolfertz, Milan Majtanik, Na Gao, Janina Maß, Guang-Zhong Wang, Thomas Laubach, David Heckmann and Bastian Pfeiffer.

Special acknowledgements go to the systems administrators Jochen Kohl and Lutz Voigt for their valuable technical and scientific support and to the heart of the department, our secretary Anja Walge, who was always nice and helpful.

Furthermore, I would like to thank the students in our lab for their cheers and will to discuss biological and computer science issues, with special thanks to Claus Jonathan Fritzemeier for data supply and discussion and Ulrich Wittelsbürger for administrative support.

Another thank you goes to Thomas Mullick and Nina Levar for participating in my research topic and to the members of the William Martin lab Mayo Röttger and Nicole Grünheit for methodological support and data analysis and anyone I might have forgotten.

Finally, I want to thank my dear mom and grandma for their love and support over all those years, my friends and my dear husband Hazem for his constant scientific and emotional support and all the critical discussions that helped me a lot during the years of my studies.

Abstract

Orthologous proteins descend from a common ancestral protein via a speciation event and often keep their ancestral functions. Therefore, orthology assignment is often applied to identify gene content and functions in newly sequenced species. No commonly accepted gold standard exists so far for orthology assignment. One reason for this is a preference of different evolutionary mechanisms in different phylogenetic clades. Eukaryotic genomes often evolve via gene duplication, while LGT (Lateral Gene Transfer) is more frequent in prokaryotes. The development of orthology assignment methods is therefore often based on the research aim and requires more or less detailed resolution of different types of homology.

In this work I developed phyloCOP (phylogeny-based Clusters of Orthologous Proteins), a new greedy phylogeny- and reference-based orthology assignment method that detects transitive orthologous relationships in prokaryotes, while simultaneously excluding paralogy. PhyloCOP was designed to create orthologous clusters without one-to-many relations (paralogous genes) that can be directly used for function prediction and evolutionary studies. PhyloCOP provides customizable parameters to adjust the algorithm to the requirements of various datasets and research aims. The user defines the reference genome on which her or his comparative research is based. The degree of transitivity between orthologs within a cluster is also user-specified, which makes phyloCOP adjustable to prokaryotic datasets that include genomes with various phylogenetic distances. In order to evaluate phyloCOP, clusters generated from 14 and 539 prokaryotic genomes were compared to similar sequence similarity-based algorithms. PhyloCOP clusters that correspond to universally distributed Clusters of Orthologous Genes included genes from nearly all analyzed genomes, which is a proof for good orthology assignment quality.

Metabolic networks consist of metabolites connected by reactions, which are catalyzed by enzymes. Complex network connections are resolved best by regarding simpler units within the system. Coupled reaction subsets, basic functional modules of metabolic networks, in which reactions are connected in a common anabolic, catabolic or transport pathway, are used in this work to get insights into the evolution of metabolic networks in prokaryotes. If metabolic network reactions and catalytic enzyme composition of the reference genome are established, metabolic network composition of other genomes can be resolved via transitive orthology prediction.

I applied comparative analysis to enzymes that catalyze fully coupled reaction pairs to investigate metabolic network evolution using *Escherichia coli* K12 MG1655 as reference. Ancestral relations between 14 *E. coli* genomes were reconstructed from phyloCOP clusters and topologically displayed in a phylogenetic tree. Genomes were assigned to specific evolutionary times based on their last common ancestor with the reference genome. The existence of corresponding enzymes was checked at each ancestral time for each pair of coupled reaction enzymes. In order to resolve loss of reaction couplings and the occurrence of gene loss or LGT at specific evolutionary times, fractions of coupled and non-coupled enzyme pairs were calculated at each ancestral time point. I detected a correlation between gene loss and reaction coupling. All metabolic couplings turned out to be ancient and likely existed already in the common ancestor of the species analysed. However, there was a trend of increased loss of couplings in individual species with increasing phylogenetic distance. Previously documented gene loss in *E. coli* DH10B a substrain of *E. coli* K12 MG1655 was verified, which further supports the good quality of the clusters generated with phyloCOP. In order to get deeper insights into the evolution of metabolic coupling, further studies with larger datasets of more distantly related genomes are recommended.

Zusammenfassung

Orthologe Proteine entstehen aus einem gemeinsamen Vorgängerprotein bei der Artenbildung und behalten oft ihre ursprüngliche Funktion. Die Bestimmung orthologer Proteine wird daher häufig verwendet um die Genzusammensetzung und Genfunktionen in neu sequenzierten Arten zu ermitteln. Es gibt bisher keine gemeinhin akzeptierte Standardmethode zur Bestimmung von Orthologie. Ein Grund dafür ist, dass verschiedene phylogenetische Stämme unterschiedliche Evolutionsmechanismen bevorzugen. Eukaryotische Genome evolvieren häufig durch Genduplikation, während LGT (Lateraler Gen Transfer) häufiger in Prokaryoten vorkommt. Methoden zur Bestimmung von Orthologie werden deshalb oft für ein bestimmtes Forschungsziel entwickelt und es wird eine mehr oder weniger detaillierte Auflösung verschiedener Arten von Homologie benötigt.

In dieser Arbeit habe ich phyloCOP (phylogeniebasierte Cluster Orthologer Proteine) entwickelt, eine neue gierige phylogenie- und referenzbasierte Methode zur Bestimmung von Orthologie, die transitive Orthologieverhältnisse in Prokaryoten detektiert und gleichzeitig Paralogie ausschließt. PhyloCOP wurde entwickelt, um Cluster mit einfachen Eins-zu-Eins-Verhältnissen der orthologen Proteine untereinander zu finden (ohne paraloge Proteine), die direkt für Funktionsvorhersagen und Evolutionsanalysen verwendet werden können. Der phyloCOP Algorithmus kann durch benutzerdefinierte Parameter an die Erfordernisse verschiedener Datensätze und Forschungsziele angepasst werden. Die Nutzerin oder der Nutzer bestimmt das Referenzgenom auf dem ihre oder seine vergleichenden Forschungen basieren. Der Grad der Transitivität zwischen den Orthologen Proteinen innerhalb eines Clusters wird ebenfalls durch den Benutzer festgelegt. Dadurch können die Eigenschaften von phyloCOP an prokaryotische Datensätze mit Genomen unterschiedlicher phylogenetischer Distanz angepasst werden. Um phyloCOP zu bewerten, wurden Cluster für 14 und 539 prokaryotische Genome erstellt und mit den Ergebnissen ähnlichen Algorithmen, die auch auf Sequenzähnlichkeiten basieren, verglichen. PhyloCOP Cluster, die universell vorkommenden Clustern Orthologer Gene entsprechen, enthielten ein Gen von fast jedem untersuchten Genom, was ein Beleg für die gute Qualität der Orthologiebestimmung ist.

Metabolische Netzwerke bestehen aus Metaboliten, die durch Reaktionen miteinander verbunden sind die wiederrum von Enzymen katalysiert werden. Komplexe Netzwerkverbindungen können am besten aufgelöst werden indem man einfachere

Einheiten innerhalb des Systems betrachtet. Gruppen gekoppelter Reaktionen, grundlegende Funktionsmodule metabolischer Netzwerke, in denen Reaktionen in einem gemeinsamen anabolischen oder katabolischen Pfad oder einem Transportweg verbunden sind, werden in dieser Arbeit verwendet um einen Einblick in die Evolution prokaryotischer metabolischer Netzwerke zu gewinnen. Wenn Reaktionen des metabolischen Netztwerks und die Zusammensetzung der katalytischen Enzyme eines Referenzgenoms bekannt sind, kann die Zusammensetzung metabolischer Netzwerke anderer Genome durch transitive Vorhersage von orthologen Proteinen ermittelt werden.

Zur Untersuchung der Evolution metabolischer Netzwerke habe ich eine vergleichende Analyse mit Enzymen durchgeführt, die vollständig gekoppelte Reaktionspaare katalysieren und dabei *Escherichia coli* K12 MG1665 als Referenz verwendet. Die Verwandtschaftsverhätnisse von 14 *E. coli* Genomen wurden aus phyloCOP Clustern rekonstruiert und als phylogenetischer Baum dargestellt. Basierend auf ihrem letzten gemeinsamen Vorfahren mit dem Referenzgenom wurden die Genome bestimmten evolutionären Zeitpunkten zugeordnet. An jedem evolutionären Urzeitpunkt wurde das Vorkommen orthologer Enzyme für jedes Paar gekoppelter Reaktionen überprüft. Um den Verlust von Reaktionskopplungen sowie das Auftreten von Genverlust oder LGT an bestimmten evolutionären Zeitpunkten aufzulösen, wurden die Anteile gekoppelter und ungekoppelter Enzympaare an jedem Urzeitpunkt berechnet. Dabei habe ich eine Zusammenhang zwischen Genverlust und Reaktionskopplung detektiert. Es stellte sich heraus, dass alle metabolischen Kopplungen ursprünglich sind und vermutlich bereits im gemeinsamen Vorfahren aller untersuchter Arten vorkamen. Allerdings gab es die Tendenz eines vermehrten Verlustes an Kopplungen in einzelnen Arten. Ein im Vorfeld dokumentierter Genverlust in *E. coli* DH10B, einem Unterstamm von *E. coli* K12 MG1655 wurde bestätigt, ein weiterer Nachweis für die gute Qualität der Cluster die mit phyloCOP erstellt wurden. Um einen tieferen Einblick in die Evolution metabolischer Kopplung zu gewinnen, werden weiterführende Studien mit größeren Datensätzen weiter entfernt verwandter Genome empfohlen.

Abbreviations:

BLAST	Basic Local Alignment Search Tool
COG	Clusters of Orthologous Genes
DNA	Desoxy Ribonucleic Acid
EcoCyc	encyclopedia of *Escherichia coli* genes and metabolism (EcoCyc)
E. coli	*Escherichia coli*
FBA	Flux Balance Analysis
FCA	Flux Coupling Analysis
FCF algorithm	Flux Coupling Finder algorithm
KEGG	Kyoto Encyclopedia of Genes and Genomes
LGT	Lateral Gene Transfer
MCL	Markov Cluster Algorithm
MMB-FCF	Minimal Metabolic Behaviour FCF
MUSCLE	Multiple Sequence Comparison by Log-Expectation
PHYLIP	phylogeny inference package
RBH	Reciprocal Best sequence similarity Hits
rRNA	ribosomal Ribo Nucleic Acid
starCOG	star wired Clusters of Orthologous Genes
SOG	Synteny-based Orthologous Genes

Table of Contents

Chapter 1 Introduction

Genome analysis aims to understand the molecular origin of an organism's features and behavior. Genome analysis includes cellular function prediction and determination of phylogenetic roots of a species. Comparative analysis is frequently used for gene identification and function prediction during genome analysis. Gained information about the genome and proteome can subsequently be used to trace back the evolution of a cells physiology.

The following chapters explain the concept and advantages of analyzing newly sequenced genomes via comparisons to other genomes. They give a general overview about comparative function prediction and its application on evolutionary analyses and metabolic network reconstruction, which leads to the scientific interest of this work: tracing back the evolution of prokaryotic metabolic networks via comparative genomics.

1.1 Genome analysis

The entire heritable information of a species is called **genome**. In most organisms it is encoded as DNA (Desoxyribonucleic acid), with the exception of some viruses that encode their genomes as RNA (Ribonucleic acid). Genome analysis is therefore the key to understand a life form's origin, habits and behavior. DNA is a double-stranded polymer with α-helical secondary structure. Its primary structure, a single stranded polymer, consists of a sequence of four nucleotide molecules. The sugar (DNA = deoxyribose and RNA = ribose) and phosphate component of a nucleotide build the backbone of the polymer (Ghosh *et al.*, 2003). Since the phosphate groups connect two sugar molecules by asymmetric phosphodiester bonds between the third and fifth sugar carbon atoms, DNA strands are directed with a 5' and a 3' end. By convention, the primary structure of a DNA or RNA molecule is written from the 5' end to the 3' end. Nucleotides differ in their base molecular component only. The sequence of the four existing nucleotide base abbreviations – A (Adenine), T (Thymine), G (Guanine) and C (Cytosine) – is used as simplified representation of the DNA molecule. The sequence of one DNA strand is sufficient to describe the entire primary and secondary structure of the DNA molecule, since double-strands are formed via complementary hydrogen bonds between the bases A and T as well as G and C (Berg *et al.*, 2002). The entire genome consists of one or more DNA (or RNA) molecules and can therefore be described by the nucleotide sequence. Genome analysis is a complex time-consuming task with variable workflow that depends on available resources and cost-benefit considerations. Basic steps are genome sequencing, assembly and genome annotation (Koonin *et al.*, 1996).

1.2 Comparative genome analysis

After regulatory sequences and potential genes have been identified via sequence pattern search, it is of interest to differentiate pseudogenes from expressed genes and determine their functions. **Function prediction**, the second basic part of genome annotation, is done either experimentally (e.g. gene knock-out experiments) or via **comparative analyses**, on gene, transcript and protein level. However, experimental function prediction is time-consuming and comparative analysis is often faster and simpler because of the existence of automated computational implementations. Various features of a gene can be compared, like structural similarities or similar interaction with other cellular components (Clifen *et al.*, 2003; Kellis *et al.*, 2003).

Sequence similarity comparisons between putative genes via BLAST and FASTA provide fast and relatively distinct function prediction. Genes, in particular genomic sequences which are flanked by gene specific regions, are compared to sequences of genes with known function (Koski and Golding 2001).

1.2.1 Genes, proteins and function prediction

Proteins, complex polymeric molecules, can be viewed as executive versions of genes. Most cellular components are made of proteins, which have various functions. Some proteins form cell structure while others are catalytic enzymes, which perform metabolism and maintain cell homeostasis. The nucleotide sequence of a protein-coding gene determines the primary structure of a protein, a sequence of 20 different covalently bonded amino acids. Hence, the terms gene and protein will be used interchangeably in the following chapters. Comparative function prediction at **protein sequence level** is more sensitive, because of the redundancy of the genetic code. Different nucleotide triplets specify one amino acid. Genes with identical sequence on the protein level might therefore not have the same nucleotide sequence. Nucleotide substitutions that do not affect the amino acid sequence of proteins are called synonymous nucleotide substitutions. Synonymous substitution rates are genome specific. In addition, there is a gnome-specific codon bias. Different species prefer different codons for the same amino acid (Grosjean *et al.*, 1982, Ermolaeva, 2001). Therefore, amino acid sequence comparison leads to more significant gene function prediction than nucleotide sequence comparison. Various algorithms, e.g. BLAST, FASTA or Smith-Waterman, exist for pairwise protein sequence similarity search (detailed explanation of BLAST in Chapter 2). In contrast to global alignment, algorithms that perform local sequence alignment do not compare the whole sequence in an instant, but seed their starting points at different positions in the sequence to get the longest local similarities (Altschul *et al.*, 1990).

Definite knowledge about a protein's function can only be achieved by often time consuming experimental analysis. Previous comparative sequence similarity detection can help to curtail possible functions. However, proteins with similar sequence must not have similar functions. For a more distinct function prediction and also for evolutionary studies based on sequence similarity comparison, phylogenetic relationships between genes and species have to be taken into account. The following subchapter will give an overview about evolutionary mechanisms, which drive gene and species diversification and explain differences between prokaryotes and eukaryotes.

1.2.2 Genome evolution in prokaryotes and eukaryotes

Just as species originate from ancestral species, genes originate from ancestral genes. Genes descend from one another via **speciation events, gene duplications** or gene transfer from other species, so called **LGT (Lateral Gene Transfer)** (Bapteste *et al.*, 2009).

Gene and genome duplications are important evolutionary driving forces for the development of new genes, especially in **eukaryotes** (Wapinski *et al.*, 2007a). Selective pressure on function maintenance of duplicated genes differs from non-duplicated ones, which leads to the accumulation of functional changes in duplicates.

Gene duplications happen with notably lower frequency in **prokaryotes**. Bacterial or archaeal genomes evolve more often via direct or indirect genomic DNA exchange between individuals of different species (LGT) (Koonin *et al.*, 2009). Three main transfer types exist: 1) Transformation, where environmental DNA is absorbed by the cell, 2) conjugation, by direct cellular contact and 3) transduction via bacteriophages, which act as transfer vectors (Davison, 1999; Bapteste *et al.*, 2009).

The classical depiction of reconstructed **phylogenetic relations** is a tree like structure. It is assumed, that two genes or species descend from one common ancestor. Gene trees are a topological illustration of the evolutionary history of a particular group of genes with common ancestry. Phylogenetic species trees encompass the evolutionary history of entire genomes or superordinate taxonomic groups and are usually generated from selected vertically inherited genes (often rRNA genes), which exist in all genomes. The leaves of the tree refer to investigated species or genes. Branches connect leaves with common ancestors. Ancestral nodes are further connected down to the root of the tree, which corresponds to the last common ancestor of all investigated species or genes. Some phylogenetic trees depict a summarized evolutionary history of clades, monophyletic organism groups with common origin (Figure 1.1).

A phylogenetic tree is a reliable illustration of eukaryotic genome evolution, since most genes are vertically inherited. However, phylogenetic trees are only approximations for prokaryotic genome evolution because of the high number of LGT events even between distantly related genomes. Figure 1.1 shows a phylogenetic tree that was reconstructed from the rRNA operon that exists in all depicted prokaryotic taxa (Dagan and Martin, 2009).

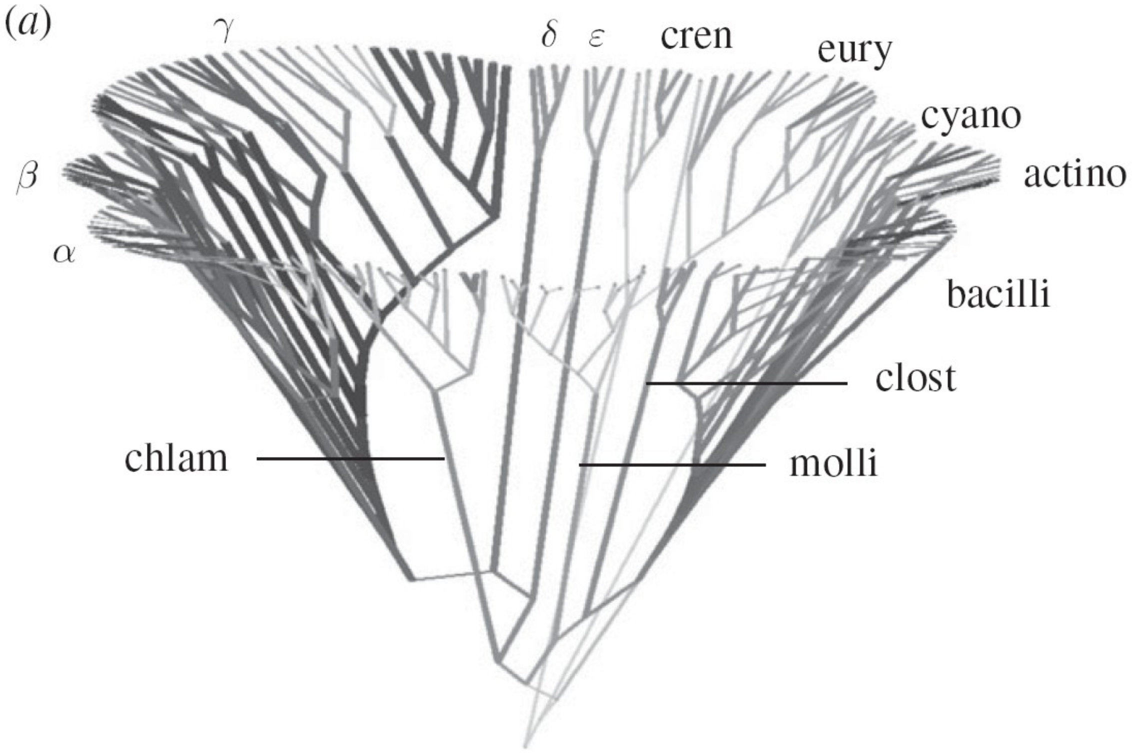


Figure 1.1: Phylogenetic tree of proteobacteria (Dagan and Martin, 2009).
The phylogeny was reconstructed from the rRNA operon that exists in all included genomes. It should be considered as an approximation that does not completely resolve prokaryotic genome evolution.

The gene network in Figure 1.2 is a better and more realistic illustration of prokaryotic genome evolution. Connections between the branches of a phylogenetic tree show LGT events that connect different, even distantly related taxonomic clades (Dagan and Martin 2009).

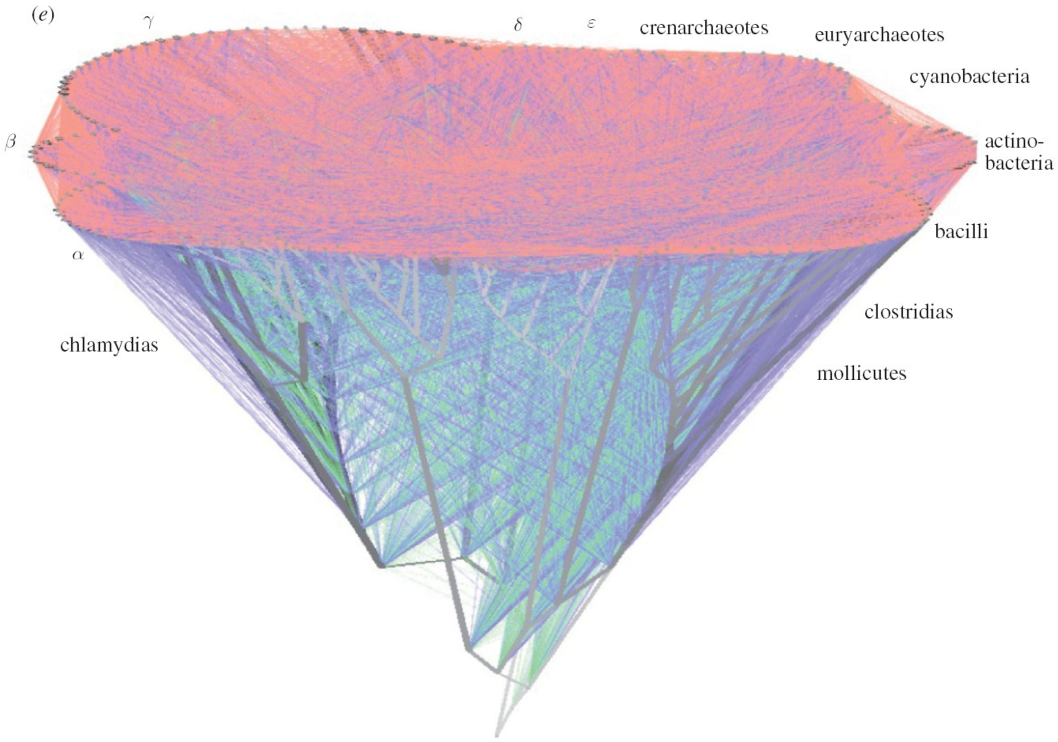


Figure 1.2: Phylogenetic relations between proteobacteria (Dagan and Martin, 2009)**.**
This network representation of protobacterial phylogeny is based on the phylogenetic tree in Figure 1.1 that served as backbone representation of vertical inheritance. The branches are connected by LGT events.

1.2.3 Comparative phylogeny reconstruction

Genome annotation via comparative sequence similarity search to already annotated genomes resolves most gene identities, because of the allover high protein uniformity among prokaryotes. To trace back the evolutionary history of a species, it's relatedness to other species has to be obtained. Phylogenetic relations among species, which describe their origin and development, are resolved from comparisons of common parts of the genome or sequence comparisons of homologous proteins. Sequence selection depends on examined species and available data. Mixed datasets including prokaryotes and eukaryotes require the comparison of universally distributed genes. Such genes usually have a function that is necessary in all life forms. Most of them are related to basic cellular mechanisms - for example, functions or structures that are involved in gene expression, like ribosomal proteins (Chapter 2; Ciccarelli *et al.*, 2007).

Reconstruction of phylogenetic relationships among multiple species is often based on **multiple sequence alignment** in which more than two DNA or amino acid sequences are aligned simultaneously. Various multiple alignment algorithms exist, like CLUSTAL, MAFFT or MUSCLE, which are based on heuristics and differ mainly in their accuracy and speed (Edgar, 2004). These algorithms use a progressive hierarchical strategy for phylogenetic tree reconstruction, in which pairwise optimal alignments are subsequently used as guide tree for step-wise multiple alignment. Ancestral relations are afterwards obtained from the alignment results via statistical methods like Maximum Parsimony or Maximum Likelihood (Guindon *et al*., 2003, compare Chapter 2).

Although phylogenetic trees are only an approximation of prokaryotic evolutionary history, they are useful to describe phylogenetic relations between eukaryotes and prokaryotes. Furthermore, the classical tree phylogeny is easy to retrieve from universally distributed genes, while the detection of LGT events is a rather complex task that may require multiple gene tree comparisons.

However, discrimination between different types of ancestral gene relations is important for function prediction and exact phylogeny reconstruction via comparative analysis.

1.2.4 Homology and analogy

The biological term homology describes similarity between features with common ancestry of two or more taxonomic groups. In contrast, similar inter-taxonomic features without ancestral relation are called analogous. According to that, similar genes with common ancestral origin are referred to as **homologs**, while evolutionary unrelated genes with high sequence similarity are called **analogs** (Fitch, 1970).

The similarity of analogous genes is often based on specific functional sequence motifs or domains, like e.g. the zinc-finger motif, which evolved several times independently. Some analogous genes have similar function due to these domains. Simple secondary protein structures like α-helix or β-sheet can similarly be placed around one another by chance. In contrast, **complex similarities** of protein sequences and higher level protein structure indicate homologous relations. Identification of homologous genes forms the basis of virtually all comparative genomic analyses (Krishna et al. 2004).

In order to gain information about the evolution of genes and species analogous and homologous genes must be distinguished.

1.2.5 Orthology, Paralogy and Xenology

Homologous genes evolved from a common ancestral gene. It is assumed that functional similarity between homologous proteins is associated with their evolutionary history. Different types of homologous relationships exist, which are differentiated by the evolutionary event that induced their occurrence.

Orthologs are genes in different species that originate via speciation from a single gene in a common ancestral species (see Figure 1.3). It is often assumed that orthologous genes retain their function during the course of evolution, as a loss of function would cause selective disadvantages in most cases (Kuzinar *et al.*, 2008).

In contrast, **paralogs** are copies of the same ancestral gene within one genome (see Figure 1.3). Paralogs arise from **gene duplications**, which occur independently from speciation events. As explained previously, the frequency of genome duplications in prokaryotic genomes is much lower than in eukaryotes. The amount of paralogs is therefore lower in prokaryotes. Paralogs rarely retain identical functions. While the original function is frequently conserved – especially if its loss affects the organism’s fitness – different mutations may fully or partially change the function of one or both paralogous copies. Neofunctionalization, for instance, leads to functional changes in only one of the two paralogs. Subfunctionalization splits the original function among both copies (Rastogi *et al.*, 2005). Advantageous mutational changes accumulate faster in duplicated genes, because two copies increase the organism’s flexibility to respond to selective pressure on the original function.

However, correct assignment of orthology and paralogy is required for accurate function prediction. Comparisons of intra-genomic with inter-genomic sequence similarities help to distinguish between these two types of homology (Tatusov *et al.*, 2000).

The existence of paralogs causes one-to-many or many-to-many orthology relationships (see Figure 1.3). With respect to the **order of evolutionary events**, paralogy can further be differentiated into: 1) **in-paralogy**, if duplication happened after speciation and 2) **out-paralogy** if duplication happened before speciation. These notations have been introduced by Remm et al., 2001 to distinguish between recent and ancient gene duplication. They developed a program Inparanoid, which detects orthologous clusters between a pair of genomes, including all common in-paralogs. In-paralogs are co-orthologous to the same genes in other species, while out-paralogous are not (O'Brian *et al.*, 2005). In-paralogs descend all from the same protein in the common ancestor of both species, while this is not the case for out-paralogs. As mentioned before, functional similarity among paralogs is less likely than among orthologs. However, because of their closer phylogenetic relatedness, functional similarity among co-orthologs/in-paralogs is more likely than among out-paralogs. Inparanoid detects co-orthologs and differentiate them from out-paralogs to get a complete picture of all orthologous relationships between a pair of species, which is especially important for function prediction in eukaryotes, were duplicated genes occur with high frequency (see Chapter 1.2.6). The example in Figure 1.3 illustrates that the existence of paralogous relationships increases **function prediction complexity** (Jensen, 2001): All genes are co-orthologous to A1. With respect to speciation event 2 (and the corresponding ancestral species), genes B1 and B2 are intra-species out-paralogs, while C2 and C3 are in-paralogs to one another but out-paralogs to C1 and B1. Consequently, gene B1 is orthologous to C1 but not to C2 and C3. B1 and C2 are inter-species out-paralogs. With respect to speciation event 2 not all paralogous genes in species C are co-orthologous to the same gene in species B (Roth *et al.*, 2008). Various orthology assignment methods differentiate between in- and out-paralogy during the cluster process. Final clusters include in-paralogs of each step, which may lead to difficulties in gene function assignment if the phylogenetic relation between the cluster members is not resolved.

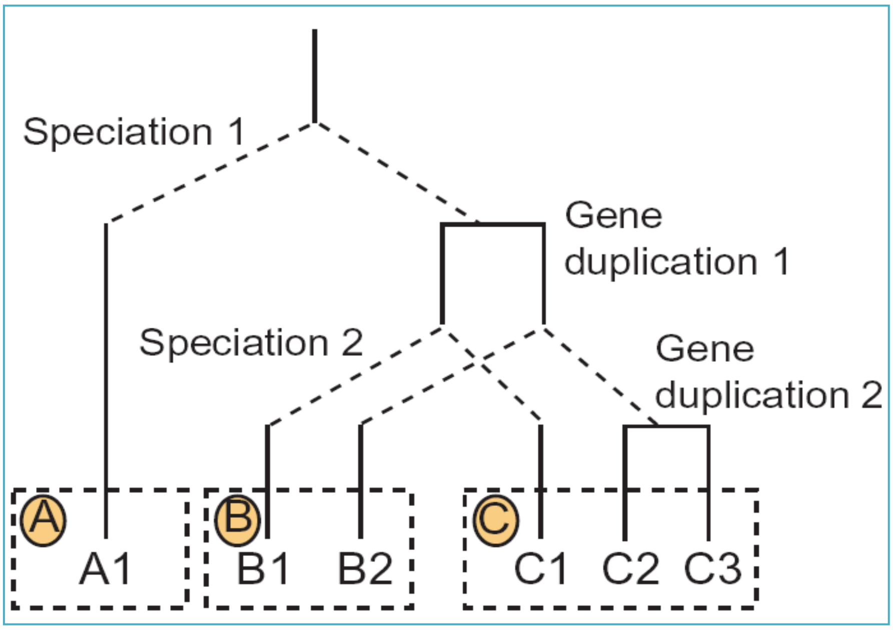


Figure 1.3: Orthologous and paralogous relations between genes (Jensen, 2001)**.** This homology subtype diagram illustrates the development of species A, B and C and their common homologous gene set. Numbers refer to intra-species paralogs. Y-shaped bifurcations symbolize speciation and horizontal lines duplication events.

Laterally transferred genes, so called **xenologs** are also homologous to genes from the donor genome. Recently transferred genes may have similar functions to the original in the donor, but they are not suitable for classical phylogeny predictions which take only vertical gene transfer into account, since they cross-connect phylogenetic line. Xenologs, if mistaken for orthologs or paralogs, violate the reconstruction of the phylogenetic history of genes and genomes. Since the tree like evolutionary history of prokaryotes is interrupted by LGT, xenologous genes should be either filtered for classical, vertical inheritance based phylogenetic tree reconstruction (approximation of the real prokaryotic phylogenetic history) or used to depict network-like evolutionary relations between the tree branches (Dagan and Martin 2009). It is possible to distinguish xenologs from orthologs in closely related genomes by **synteny** considerations. The synteny criterion assumes that if the chromosomal environment of two homologous genes is similar, they most probably arose via vertical inheritance and are orthologous rather than xenologous (Chapter 2; Wapinski *et al.*, 2007a+b).

1.2.6 Homology assignment methods

Because of the variability of evolutionary mechanisms various homology assignment strategies were designed to fulfill different requirements for subsequent analyses based on the assignment. Most homology assignment methods are therefore tailored for specific taxonomic groups and scientific tasks like genome analysis, function prediction or evolutionary studies. Homology prediction is based on sequence similarity in most of the cases. Homologous relationships can be captured **1) pairwise**, which means that orthologs (optionally including or excluding paralogs) from two species are determined due to their mutual sequence similarity independently from homologous relations to other species, or **2)** through a **clustering** process, in which orthology relations of all examined species are taken into account simultaneously or sequentially. The method of choice again depends on research questions and observed species. A pairwise approach can reach higher resolution of co-orthology relations for closely related species (O'Brian *et al.*, 2005; Roth *et al.*, 2008). In contrast, usage of cluster algorithms is advantageous for the detection of orthology in distant species, since orthologs with low mutual sequence similarity can be recognized via common orthology relations to genes from intermediate species. This property of orthology is called **transitivity** (Roth *et al.*, 2008).

On the other hand there are some homologous proteins, which are evolutionary distant and lost their similar functions. This is mostly true for genes from distantly related species or for homologous genes that occurred via gene duplication prior to the last common ancestor of a homologous cluster. For correct function assignment orthologous genes have to be distinguished from paralogs, which sometimes is not accurate when only based on inter-genomic sequence comparisons (see previous Chapter; Wapinski et al., 2007a). Additional intra-genomic sequence comparisons can help to detect paralogous genes. **Synteny** (similar chromosomal neighborhood) among closely related orthologs is a useful criterion to distinguish xenologs from other orthologs, (Wapinski *et al.*, 2007a).

1.2.6.1 Pairwise algorithms

The most simple approach for homology assignment is based on pairwise **reciprocal best** sequence similarity **hits** (**RBH**) between two genomes (Fitch, 1970; Wall *et al.*, 2003). If two genes from different species have the highest mutual sequence similarity they are considered to be orthologous. However, this procedure may lead to false assignments, if paralogous genes are involved. Since only best hits are taken into account, one-to-one pairs of orthologs are detected even if the true orthology relation is multiple, which may lead to incomplete or false subsequent function prediction. Some orthologs may even not be detected at all, if the reciprocal alignment to the duplicated gene is better. Paralogs can be detected if other sequence similarity hits are taken into account but for correct function prediction they must be distinguished from orthologs.

Inparanoid, for example, has been developed, to resolve the complex orthologous and paralogous relationships among pairs of eukaryotic species. I explain this method here to clarify that a complete resolution of orthologous and paralogous relations can only be achieved by a pairwise approach, since co-orthology and both types of paralogy are defined with respect to a specific speciation event (compare Chapter 1.2.5). On the other hand, clusters generated by this method cannot be directly used for subsequent phylogeny reconstruction or other evolutionary analyses, which is a striking drawback of pairwise orthology assignment. Inparanoid is an extended version of the standard RBH methods, which takes not only the best hits into account and detects many-to-many orthologous relations for a pair of genomes. The resulting clusters of Inparanoid only contain co-orthologous genes from two genomes, excluding out-paralogs (Sonnhammer *et al.*, 2004). Each co-orthologous cluster is based on an orthologous gene pair, which builds the core of each homologous cluster. At first, all mutual BLAST hits between a pair of genomes as well as their mean mutual scores are obtained. Gene pairs with a mean mutual score above a score cut-off (50 bits) and an overlap cut-off (50 % overlap) are putative orthologous pairs. Subsequently, co-orthologs are clustered around the pair of main orthologs, regarding each genome separately. Genes are only compared to main orthologs if they are from the same genome. A gene is considered in-paralogous and clustered to a corresponding orthologous group if its sequence similarity to the main ortholog is higher than to any gene from the other genome.

If orthologous clusters are overlapping, their grade of overlapping is checked by different criteria and clusters are either merged or deleted, or proteins assigned to the cluster with the best matching main ortholog.

Again, clusters obtained by Inparanoid resolve all recent homology relationships between a **pair of species** and may successfully be used for further evolutionary studies. However, pairwise clusters need to be further processed and merged to resolve the evolutionary history of each homologous gene set, which is a difficult task. Inparanoid marks the transition to more complex clustering algorithms, which assign orthologs for more than two genomes simultaneously but lack the differentiation between in- and out-paralogs in the final clusters.

1.2.6.2 Clustering algorithms and transitivity

Clustering algorithms detect homologous relations between genes from more than two species and assign genes into corresponding clusters. They are especially useful for further phylogenetic analysis, since multiple alignment of the cluster members can directly be used as input for phylogenetic tree reconstruction (Tatusov *et al.*, 2000; Wapinski *et al.*, 2007).

One of the simplest and most intuitive clustering techniques is reference-based, which means, that all genes with significant sequence similarity to a reference gene are connected to it. The basic idea is that well characterized functions of a reference genome can be assigned to corresponding orthologous genes from other genomes. Reference-based orthology assignment is frequently used for comparative functional analysis and further explained in Chapter 2 using the example of starCOG K12. This method is an expansion of the reciprocal best hit approach explained previously to more than two genomes. However, it is only useful if the compared genomes are phylogenetically close, sequence similarity between the reference genes and its orthologs is significantly high and the number of paralogous genes is low.

In contrast to pairwise or reference-based species comparison, orthologous genes from species, with low mutual sequence similarity, can be detected via common orthologous relations to genes from intermediate species. This property of orthology is called **transitivity**. However, transitivity is violated by paralogy, since not all paralogous genes are co-orthologous to the same genes (Figure 1.3, Roth *et al.*, 2008). Insufficient discrimination between recent and ancient paralogy may therefore lead to the assignment of phylogenetically distant paralogs, which have no **functional similarity** to the other orthologs. Furthermore, evolutionary gene or genome history may be resolved incorrectly if the order of duplication and speciation events is not resolved correctly. This is especially problematic for homology prediction in eukaryotic species.

In contrast, most vertically inherited homology relationships among prokaryotes are orthologous. Cluster algorithms based on transitivity therefore, lead to reliable phylogeny reconstruction and function prediction even for distantly related species. Most popular examples for transitivity based homology cluster databases are **COG** (**Clusters of**

Orthologous Groups of proteins) and eggNOG ('evolutionary genealogy of genes: Non-supervised Orthologous Groups') (Tatusov *et al.*, 2000; Muller *et al.*, 2010). These databases contain prokaryotic and eukaryotic species. COGs are initialized by three genes from distantly related genomes with mutual best sequence similarity, which is taken as proof for mutual orthology relation. If such low similarity is at least detected as the best inter-genomic similarity, genes most probably belong to the same family. It is also expected that orthologs from distantly related genomes have lower sequence similarity, which is difficult to distinguish from analogous similarity. Therefore low-complexity regions are masked before the BLAST run. Before building triangles, intra-genomic gene similarities, which are stronger than inter-genomic ones are detected and corresponding paralogs are merged and treated as one orthologous unit during the subsequent analysis. Seed **triangle clusters** are subsequently merged if two of the genes are the same. Each resulting COG is analyzed afterwards. In order to distinguish the different evolutionary development of single domains, multidomain proteins are split into segments and previous steps are repeated with them. The aim is to find COGs, in which all members descended from one gene in the ancestral genome of all species represented in the cluster. The chance that this might be violated because of a duplication event previous to the ancestral speciation is high in COGs with multiple members. The last step is therefore; to check if large COGs with multiple members have to be separated, in which alignments are checked manually with the help of phylogenetic trees (Tatusov *et al.*, 2000). Compared to Inparanoid, there is still a possibility to cluster ancient paralogs in the orthologous clusters. That's why not all COGs are usable for function prediction. It must be decided, which of the clusters may include ancient paralogs. On the other hand COGs, unlike Inparanoid, clusters, may be directly used for evolutionary analysis of genes from multiple genomes.

OrthoMCL clusters orthologous and paralogous genes from more than one **eukaryotic** species based on a graph clustering algorithm (Li *et al.*, 2003). The MCL algorithm is explained in Chapter 2 of this work. The basic idea is that sequence similarities among genes are displayed in a graph, in which nodes represent genes and edges between them similarity scores. Unlike COGs, OrthoMCL identifies at first reciprocal best BLAST hits between every **pair of genomes**. To **resolve recent and ancient paralogous relations**, in-paralogous genes for every inter-genomic reciprocal best BLAST hit pair of genes are determined. Subsequently, co-orthologs are merged to one edge in the graph, which is then clustered with the MCL algorithm.

1.2.6.3 Phylogeny and Synteny

All the methods, explained so far are fast and (except the last step of COG) fully automated. Although most of the previously mentioned algorithms aim to resolve orthologous and paralogous relations, assignment results may not be fully reliable for further analyses. One main reason is that paralogs may evolve with different rates as well as orthologs. In order to resolve homologous relations correctly, including the number of gene losses, duplications and LGT events, other methods assign homologous genes considering phylogenetic relations. Most of these methods are computationally expensive (Wapinski *et al.*, 2007).

SYNERGY is a **greedy** algorithm, which clusters orthologs (and paralogs) of several genomes into orthogroups, based on a phylogenetic species tree that serves as guideline for the clustering procedure. In every step, clusters are approximated based on sequence similarity hits and step by step refined by the simultaneous reconstruction of intermediate **gene trees**. The step-wise procedure is a combination of more exact phylogeny-based and computationally less expensive automated sequence similarity hit based methods. As described previously, the branches of phylogenetic trees of prokaryotes are connected due to multiple LGT events. In order to **exclude xenologs**, which violate the reconstruction of vertical inheritance, chromosomal neighborhood relations (**synteny**) are also taken into account. All genes in a resulting orthogroup descended from a single ancestral gene via speciation or duplication. This means that the gene tree of any orthogroup, when compared to the corresponding species tree has a gene in the ancestral species which is the origin of all other cluster members.

SYNERGY at first computes a **gene similarity graph** from all-versus-all genomes FASTA sequence comparisons to get the gene distances (Pearson and Lipman, 1988). Only sequence pairs with *significant similarity* (e-value < 0.1 and identity > 50 %) are taken into account. Edges are **weighted** due to: 1) the sequence **similarity** score of the gene pair, which is obtained based on the JTT amino acid substitution model, and 2) the gene pair's **synteny** similarity score, which is defined by the fraction of neighboring gene pairs with *significant similarity* (considered as putative orthologs) and describes how similar the chromosomal neighborhood is (Jones *et al.*, 1992; Chapter 2).

After this initial step, the algorithm captures orthologous clusters step-wise based on the phylogenetic species tree of all investigated genomes. Every **ancestral node** is regarded

sequentially starting from the leaves and ending at the root. At first candidate orthogroups are detected between the genes of the genome pair with the most recent common ancestor X, based on the gene similarity graph. Two genes with mutual reciprocal edges or reciprocal edges to a third party gene are assigned into a candidate orthogroup.

This is followed by a **phylogenetic gene tree reconstruction** for each candidate orthogroup. Simple one-to-one orthogroups are easy to reconstruct. Neighbor-Joining, based on a distance matrix of all genes in the candidate orthogroup is used for the reconstruction of more complex one-to-many or many-to-many orthologous relations. All internal branches are inspected in order to find the correct **root**. The rooting at each position is based on the assumption that all distances between leaves and root should be similar, and scored by a leaf-to-root variance proportional to sequence similarity and synteny scores. This is correct if all genes evolve with similar rate but violated by the different evolutionary rates of paralogs (see previous Subchapter). To resolve this, more likely root positions with lower occurrence of gene duplication or loss - with respect to the number of genes included in the orthogroup - are favored. The rooting score is thus retrieved by combined consideration of gene duplication and loss likelihood and sequence similarity and synteny variance. Rooting of the **phylogenetic gene tree** is a crucial step, since the position of the root reveals if the candidate orthogroup is sound - which means that all included genes descended from one common ancestral gene in the ancestral genome X. A duplication at the root of gene tree hints for a gene duplication that happened before X (or in X) and the candidate orthogroup is **split** into two distinct ones. This phylogeny-based step restricts and refines the coarse-grained homology search and determines final orthogroups for every internal species tree node. The alternation of coarse-grained orthogroups prediction with the phylogeny-based refinement aims 1) to detect complete clusters, which include all genes that descended from one gene in X, and 2) to reconstruct their phylogenetic history.

Gene tree reconstruction is followed in every iteration by an update of the **gene similarity graph**: a) genes, which belong to the same orthogroup are merged by replacing the gene nodes by orthogroups, b) distances between the new orthogroups and other nodes in the graph are recalculated with the standard formula for distance update (used for example by the neighbor-Joining algorithm):

$$d_{new,p} = \frac{1}{2}\left(d_{a,p} + d_{b,p} - d_{a,b}\right)$$

Here **p** refers to any gene or orthogroup in the graph and **a** and **b** are the original genes assigned to the **new** orthogroup.

All steps are repeated for every intermediate ancestral node until the root of the species tree. Each detected orthogroup from previous iterations is treated as a unit in the following steps. Like the genes in the beginning, they are step-by-step merged to superordinate orthogroups. This leads to a **complete** final depiction of the **phylogenetic history** of each orthogroup, including the correct sequence of ancestral gene duplication and speciation events. The evolutionary history of eukaryotic genomes can be subsequently reconstructed using the SYNERGY clusters. The same is true for prokaryotic genomes, but one has to keep in mind that a phylogeny reconstruction based on vertical inheritance is only an approximation of prokaryotic genome evolution (Dagan and Martin 2009).

This method is less computational expensive than other phylogeny-based approaches, since gene tree reconstruction is oriented at a phylogenetic species tree. However, SYNERGY's greedy algorithm requires gene tree calculations at every step. Additionally, function prediction based on the orthogroups is complex. The corresponding final gene trees of the orthogroups must be examined manually to decide, which of them do not include many ancient paralogs and are therefore reliable for subsequent function prediction. It is also likely that genes in orthogroups with multiple duplication events have diverse functions. The cost-benefit ratio of SYNERGY is good for complex gene family analysis but for comparative function prediction it might be more useful to exclude paralogy and generate pure orthologous clusters by a computationally less expensive procedure.

1.3 Systems biology and the evolution of metabolic networks

The ancestral origin of many newly sequenced genomes is unknown. Evolutionary studies focus on the origin of species, genes, cells and cellular components as well as their change and diversity over time, also referred to as their evolutionary history.

Huge, daily growing molecular biological databases include data about different cellular levels. Computer based evaluation is necessary to resolve and understand these data. In the past, biological knowledge was sparse, and scientific research focused on highly specialized knowledge about small parts of biological systems like cells, tissues, or organisms.

Hence a new trend of integrated investigation of biological systems crystallized among biologists, caused by accelerated experimental speed. Integrated research on complex systems requires the application of interdisciplinary tools. **Systems biology** is a modern interdisciplinary field, based on the idea of an integrated systemic biological analysis, in which experimental analyses and computer simulations alternate for systematic resolution of complex biological processes. It is a synonym for modern biological research covering various biological techniques (Palsson, 2006).

The concept of a **biological system** describes all components that keep a biological process working. It depends on the perspective and the hierarchical level of analysis, which components are taken into account and how detailed the system is observed.

1.3.1 Metabolic network analysis

As described previously, any cell or higher life form can be considered as complex hierarchical system, which consists of many lower ranked interacting systems. Cell **metabolism** is a system of chemical reactions that maintains cell homeostasis and cell function. The biological field of metabolomics deals with the existence and concentration of metabolites, while fluxomic studies focus on the chemical reactions that convert one metabolite into another.

Metabolic networks are interaction networks between metabolites, which are connected by reactions that are often catalyzed by enzymatic proteins. Each reaction proceeds with a specific speed, its **reaction rate or flux**. Different pathways through the entire network follow the transformation of specific metabolites. Two general types of pathways exist: 1) anabolic pathways, which synthesize more complex molecules like nucleic acids or amino acids from smaller components, and 2) catabolic pathways that decompose substrates into energy carrier molecules or building components, which are further used in anabolism (Palsson, 2006).

A typical illustration of a metabolic network is a **graph**; in which nodes represent metabolites connected by vertices that refer to reactions (see Figure 1.4). This simplified network representation does not include enzyme co-factors or other regulatory molecules, but can be easily reconstructed as stoichiometric matrix and therefore, directly be used as input for computational network analyses.

Metabolic network analysis is a growing systems biological branch for the investigation of metabolic pathways in biological systems like cells and tissues, which aims to identify the distribution of all involved reactions and their corresponding fluxes, to understand the whole system's metabolism. Metabolic networks are usually investigated by both experiments and computer simulation (Covert et al., 2001). For a simplified network analysis different hierarchical levels can be regarded step by step, starting from the whole cellular in- and output over the two metabolic sectors anabolism and catabolism, resolution of pathways, which fulfill a definite role until the resolution of individual reactions.

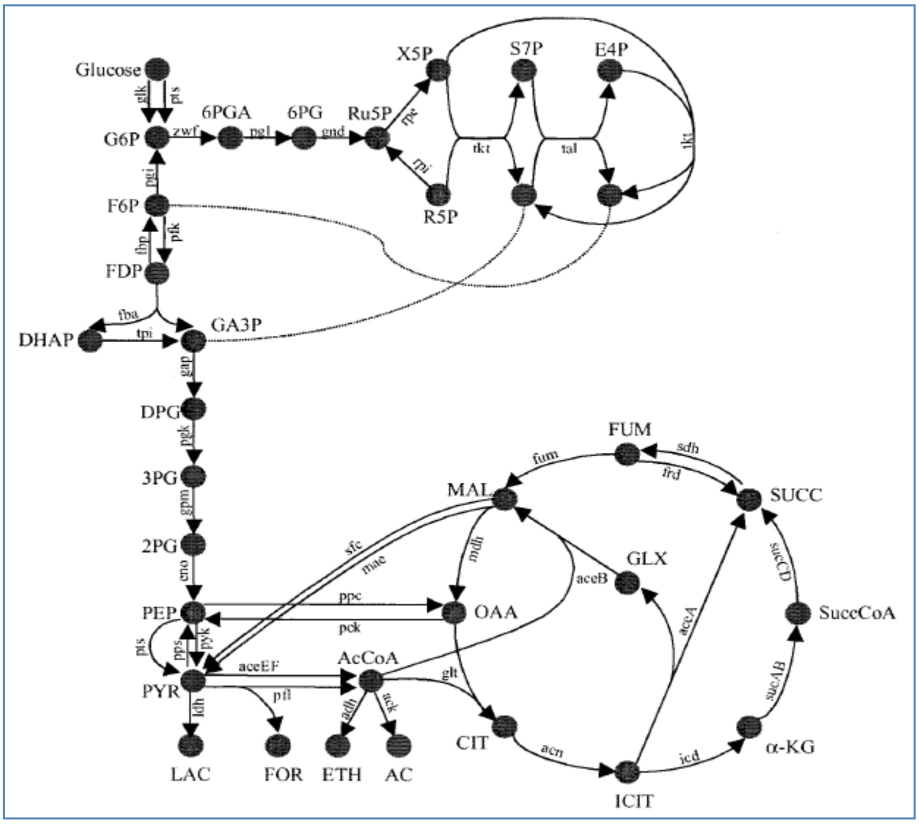


Figure 1.4: Central metabolic network pathways of *Escherichia coli* (Edwards *et al.*, 2000)**.** Graphical network illustration of the central cell metabolism, which serves energy maintenance and consists of three main steps: glycolysis, pentose phosphate pathway and citric acid cycle. Nodes symbolize metabolites and arrows chemical reactions. Abbreviations for metabolite names are typed in capital letters. Small letters are abbreviations for names of catalyzing enzymes.

An overview over network reconstruction and analysis is illustrated in Figure 1.5. The first step of network analysis is metabolic network reconstruction. Metabolic networks are reconstructed by various biological information resources with different reliability and availability. Genomic and comparative analysis data are easier to accomplish but less reliable than biochemical data. However, genome annotation and metabolic network analysis alike are frequently based on comparative analyses. *In silico* network reconstruction starts with the generation of a list of all involved metabolites and reactions as well as the corresponding catalyzing enzymes. The common gene set of prokaryotes leads to reliable bona fide data (compare Subchapter 1.2.2). The enzyme set of bona fide networks reconstructed by comparative analysis is subsequently supplemented and corresponding biochemical reactions are determined by experimental data. Some experiments are designed for particular investigations but a lot of experimental information is already stored in organism specific databases like the encyclopedia of *Escherichia coli* genes and metabolism (EcoCyc) or databases, which include more than one organism like the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Karp *et al.*, 1999; Ogata *et.al.*, 1999). Experimental data, as well as physiological data about an organisms metabolic abilities (e.g. to grow in media without specific nutrients or essential amino acids) help to complete the list of network components. The integrity (completeness) and correctness of the reconstructed network is crucial for its further analysis.

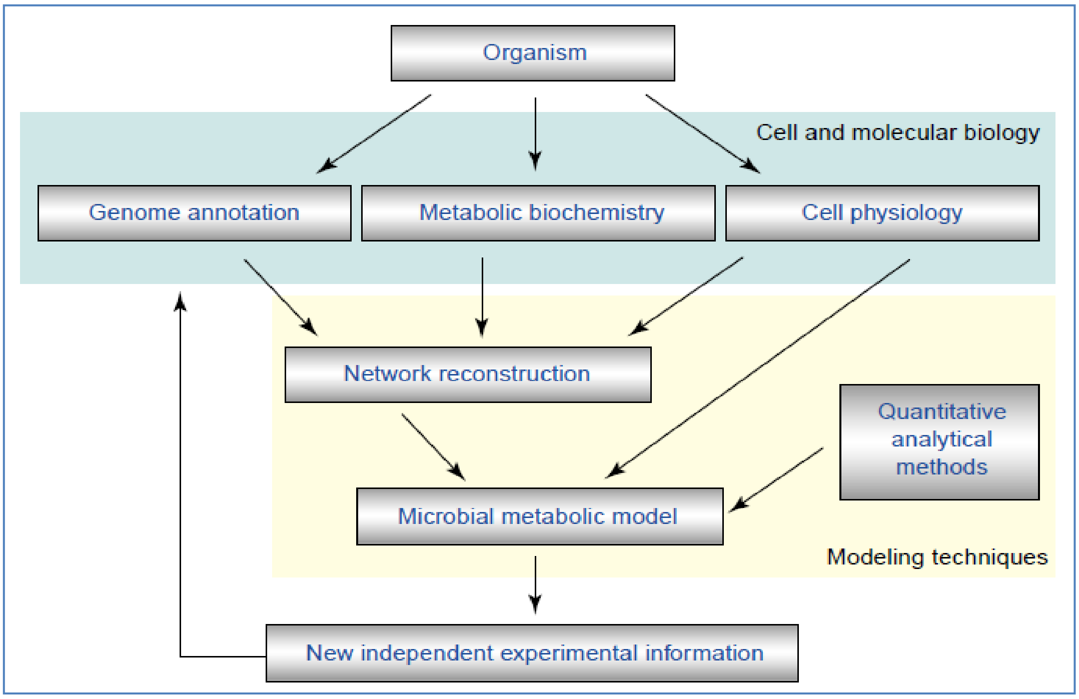

Organism
Cell and molecular biology
Genome annotation
Metabolic biochemistry
Cell physiology
Network reconstruction
Quantitative analytical methods
Microbial metabolic model
Modeling techniques
New independent experimental information

Figure 1.5: Metabolic network analysis (Covert *et al.*, 2001)**.** Metabolic network analysis starts with network reconstruction via genome annotation, metabolic biochemistry and cell physiology data. Subsequential computational network analyses are based on network reconstruction and mathematical models that describe the biological meaningful properties of the system, taking physiological data and data from quantitative experimental analyses into account. Reconstruction and modeling processes alternate repeatedly until the networks components and behavior are completely analyzed.

The resulting metabolic map, which includes the network components of a system and all connections, is used as input for computational model reconstruction and analysis of the system's behavior. Pathways through the network are directly established from the reconstruction, as well as different lengths of alternative pathways. Physiologically relevant pathways through the network can be obtained without knowledge about exact kinetic properties of each reaction, using stoichiometry and cell physiology information. Mathematical models for computational simulations of the cell's network behavior – also under specific environmental constraints – take physiological information like substrate uptake rates and biologically relevant objectives like maximization of biomass production into account.

Both, network reconstruction and simulation are iterative processes, in which *in silico* modeling and experimental verifications alternate (Covert *et al.*, 2001)

1.3.2 Evolution of metabolic networks

Not only the structure and behavior but also the **evolution of metabolic networks** is analyzed via comparative genomics. Since orthologous enzymes have very often the same function, it is possible to assign knowledge about catalyzed reactions of the reference enzyme to their orthologous counter parts, which builds a backbone for subsequent experimental proof.

For biologically meaningful comparative function prediction, well investigated genomes should be used as templates. Completely annotated genomes of well established model organisms are the best source for reliable data. In general, metabolic networks of monocellular organisms are easier to analyze than more complex systems like tissues or multi cellular organisms. The gram negative bacterium *Escherichia coli* is a fast growing molecular biological model organism with almost completely sequenced and annotated genome that lives in the lower intestine of warm-blooded organisms like humans. The metabolic network of *E. coli* substrain K12 MG1655 has been well investigated, and is therefore a useful reference for comparative studies of the evolution of prokaryotic metabolic networks (Palsson, 2001).

Any evolutionary analysis needs multiple comparisons for a meaningful reconstruction of evolutionary events (see Chapter 1.3.1). The same is true for evolutionary history reconstruction of metabolic networks. Contemplation of the whole network may not resolve biologically meaningful evolutionary trends. Instead of analyzing all biochemical reactions separately, the focus on the development of **functionally connected modules** leads to a better insight into the development of the whole system. The detection of differences in the composition of corresponding functional modules in different genomes may be used to trace back their origin. Information about the evolution of network modules can subsequently be associated to reconstruct the development of bigger parts of the network.

As mentioned before, the network consists of functional modules, which work together to decompose substrates or assemble molecules of different complexity. These modules consist of reaction pathways that aim a specific metabolic task. Reactions within a pathway depend on each other if there is no alternative pathway for the uptake of an involved intermediate metabolite. Reactions, with dependent reaction rates or fluxes are so called **coupled reactions**. Reactions can be directionally, partially or fully coupled (linearly dependent). A reaction v_j is directionally coupled to another one v_i, if it can only be processed, if there is a non-zero flux i through v_i. If this relation is true in both directions, the coupling is partial. If their reaction rates i and j are linearly dependent, v_i and v_j are fully coupled. Coupled reactions are e.g. detected via Flux Coupling Analysis (Burgard *et al.* 2004). Corresponding catalyzing enzymes are also considered to be coupled.

For the evolutionary analysis in this work, orthologous coupled enzymes are assigned, excluding paralogs (because they violate function prediction), followed by phylogenetic species tree reconstruction. Finally, their appearance in every branch of the phylogenetic tree is checked to trace back the existence of coupling in common ancestors (see Chapter 3).

1.4 Aim of work

This work deals with the evolutionary analysis of metabolic coupling based on comparative genomics, encompassing the development of a **orthology assignment tool phyloCOP** (phylogeny-based Clusters of Orthologous Proteins) for the detection of paralogy-free orthologous clusters, which are subsequently used to trace back the **evolutionary history** of fully coupled reactions in different *Escherichia coli* strains. *Escherichia coli* K12 MG1655, a model organism for metabolic network analysis, was used as reference for the detection of orthologous coupled subsets in the other genomes.

As described in previous sections, investigation of gene compositions of newly sequenced genomes as well as function predictions are often based on comparative genomics.

Contemplation of orthologous and paralogous relationships among genomes resolves the phylogenetic history of genes and genomes respectively. For reliable comparative function prediction, orthologs must be distinguished from paralogs. Various phylogeny-based methods exist for the correct resolution of orthologous and paralogous relationships, which are often computationally expensive. However, exclusion of paralogous relationships is no hindrance for successful resolution of phylogenetic history of prokaryotic genomes, since gene duplications occur with much lower frequency than in eukaryotes and thus have lower effects on genome evolution.

PhyloCOP assigns orthologous proteins into paralogy-free clusters. Although genomes are processed in phylogenetic order – which supports correct orthology assignment, the algorithm does not rely on gene tree calculations in each clustering step. Therefore, it can be assumed that phyloCOP is computationally less expensive than other phylogeny-based methods like for example SYNERGY (Subchapter 1.2.6.3). Because of user-defined **customizable** parameters phyloCOP can be adjusted to the requirements of different research aims.

Chapter 2 PhyloCOP: An orthology assignment tool with flexible features

Maybe the most intuitive way to study something new is by comparison. Differences and similarities to familiar things help to categorize and lead the way to further analyses. Advantages of applying comparative genomics on genome analysis and evolutionary studies have already been discussed in Chapter 1.

Discrimination between homology types may be complex and computationally intensive, depending on the diversity and kind of investigated genomes. Moreover, homology assignment results are created to serve as input for further biological analyses. Homology assignment algorithms therefore aim to maintain a balance between the production of reliable data and reduction of computational complexity.

While several methods for the identification of **orthologs** based on mutual sequence similarity are available, there is **no commonly accepted standard method** for orthology assignment across multiple genomes (Kuzinar *et al.*, 2008). Variable research challenges, the diversity of investigated species and their favorite evolutionary mechanisms require different assignment strategies (compare Chapter 1.2).

Generating complete lists of orthologous and paralogous genes for instance needs a more complex algorithm than selecting a basic set of orthologous genes for phylogenetic tree reconstruction or function prediction. In the latter cases, a paralogy filter with higher specificity but lower sensitivity for orthology is sufficient. This is also true for prokaryotic species, which - compared to eukaryotes - contain a much lower amount of paralogs. Additionally, many prokaryotic genes have an orthologous counterpart in other prokaryotic genomes. This allows predicting the function of many genes by pairwise or multiple sequence comparisons along with the exclusion of paralogous genes, using transitivity (compare Chapter 1.2). This strategy only works, if the overall fraction of paralogous genes is low, since all co-orthologous relations are excluded as well. If the aim is to capture all co-orthologous relations or if eukaryotic genomes are included in the analysis an additional differentiation of in-paralogy and out-paralogy is required (Chapter 1.2.2).

Additional differentiation of xenologs from other homologs may be needed to resolve bacterial and archaeal phylogenetic history (see Chapter 1.2.2). The chromosomal environment of orthologous genes is often syntenic. Laterally transferred genes, in the contrary, mostly have no similar chromosomal environment. However, the application of the synteny criterion is limited to closely related genomes, since the chromosomal environment of orthologous genes may differ among distantly related species.

A standard method for orthology assignment should be applicable to several of these scientific tasks and at the same time produce results fast and easy, which is hard to achieve by an algorithm with fixed features. An algorithm with flexible user-defined parameters on the other hand, can easily be adjusted to serve various research issues. Some orthology assignment parameters affect assignment stringency, like sequence similarity search quality or required transitivity (connectivity) between the members of an orthologous cluster. For the analysis of various genomes, these parameters should be user-defined, since assignment stringency depends on genome diversity.

Comparative research is also often based on a well investigated genome. Therefore a reference-based orthology assignment method with a user-defined reference is assumed to produce optimal results.

The basic research aim of this work was to create an orthology assignment method, whose results can be further used for function prediction in prokaryotes and subsequent reconstruction of metabolic network evolution. Instead of creating another algorithm that serves a specific research aim, I focused on the development of a tool with user-defined parameters, which are adjustable to other research issues.

2.1 Approach

The balance between fast and precise orthology prediction among prokaryotes is maintained best by a clustering algorithm. As already described in Chapter 1.2.2, the well-known COG database (Clusters of Orthologous Groups of proteins) includes orthologous clusters, generated based on transitivity via triangle linkage clustering. Co-orthologous/in-paralogous relationships are also included in the clusters (Tatusov *et al.*, 2000).

However, COG's triangle clustering of distantly related genes has a striking drawback: Recent pairs of orthologs, which exist only in two directly related genomes, cannot be detected. Based on COG's idea of taking only the best mutual sequence similarity hits between genomes into account, I developed a greedy phylogeny-based algorithm for orthology assignment, **phyloCOP (phylogeny-based Clusters of Orthologous Proteins)**, which combines transitivity with phylogenetic relatedness and detects both recent and ancient orthology. The phyloCOP clustering algorithm uses transitivity and phylogenetic distance for orthology assignment. PhyloCOP is designed for the determination of a **basic set** of orthologous protein clusters from a broad range of prokaryotic species; such sets are particularly useful for **function prediction** or **phylogenetic analyses**. Unlike COG, phyloCOP therefore excludes paralogs as well as co-orthologs. Because only one-to-one orthologs are included in the final orthologous clusters, correct function prediction is improved and transitivity is not violated (see Chapter 1.2).

Additionally, phyloCOP provides a set of **user-defined parameters**, with which the algorithm can be adjusted to diverse scientific needs. Based on the research question, the user selects a reference species used for cluster initialization. Proteins are added to clusters gradually by examining genomes in the order of increasing phylogenetic (or sequence similarity) distance to the reference genome. This phylogeny- and reference-based clustering process decreases computational complexity, run time and memory usage. The user specifies an e-value cut-off for the BLAST results (Altschul *et al.*, 1990). The degree of transitivity is set via a parameter α, which determines the minimal fraction of genes in a cluster that have to be best BLAST hits of any newly added gene (Figure 2.1). PhyloCOP's customizable features make it a useful tool for the investigation of a wide range of specific research questions in the study of prokaryotic genomes, and can be applied to datasets encompassing several hundred species.

2.2 Algorithm

The **phyloCOP** algorithm is implemented in **Perl** (http://www.perl.org/) and executed from the command line (Unix/MacOS, Windows). PhyloCOP is implemented as Perl scripts, which are distributed under GPL2 and can be freely downloaded from www.cs.uni-duesseldorf.de/AG/BI/Software/phyloCOP. The phyloCOP user's manual and Perl program together with example files for the usage can be found in folder Appendix_A in the attached CD. Running phyloCOP requires some **data preparation steps** – in particular pairwise sequence alignments of the proteins of all investigated genomes via BLAST, which are explained in Subchapters 2.2.2 and 2.3.

2.2.1 Greedy phylogeny-based clustering

PhyloCOP is developed as **greedy algorithm**, which is computationally less expensive than for example dynamic programming when applied to large datasets. Greedy algorithms solve mathematical problems step wise. In each step the local optimum is chosen in order to find the global optimal solution. This can be problematic, since the final solution may only be a local optimum (Jones and Pevzner, 2004).

To increase the chance to find the optimal final solution instead of a local optimum, greedy and **phylogeny-based** programming is combined in phyloCOP. The **general idea** of the algorithm is that a gene D is assigned into an orthologous cluster if it has best BLAST hits to more than the user-defined fraction of **current** cluster members, and at the same time more than the user-defined fraction of cluster members have a best BLAST hit to the same gene D. In each step genes of one genome are compared to already clustered genes and eventually added to a cluster if all clustering criteria are fulfilled in the current iteration (see Subchapter 2.2.4). Although this leads to the detection of optimal clusters in each step, it is possible that not all orthologous relationships are detected in this way, especially if phylogenetic distances among the genomes are not taken into account.

In the beginning of the clustering procedure, clusters include only a small number of members. Therefore, relations between proteins assigned first, determine the direction of the later clustering procedure and initial cluster members have the strongest impact on the final cluster content. Since orthologous genes from closely related genomes are less affected by mutational changes than distantly related genes, it can be assumed that their sequences are more similar to each other than to genes from other genomes. In contrast, sequence similarity of two orthologous proteins from distantly related genomes might be too low to detect, which may lead to assignment failure in later steps.

In order to avoid assignment failure due to low sequence similarity and to find all transitive orthologous relations, the reference-based phyloCOP algorithm processes genomes step-wise in increasing phylogenetic distance to the reference. The phylogenetic order must be obtained previously, for example via phylogenetic tree reconstruction by a Neighbor-Joining algorithm based on a pairwise sequence similarity scoring matrix (see Chapter 2.3).

2.2.2 Data preparation

First, pairwise mutual inter-genomic sequence similarities are obtained by running protein BLAST with options –m 8 for tabular alignment output and –b 1 for the best BLAST result only (http://blast.ncbi.nlm.nih.gov/Blast.cgi; Altschul *et al.*, 1990). The initial dataset for every genome must be distributed as FASTA file including all protein sequences. Those genome files are reciprocally aligned against each other via NCBI (National Center for Biotechnology Information) protein BLAST. A Perl script for the required multiple pairwise protein BLAST runs is distributed together with the phyloCOP program under GPL2 and can be freely downloaded from www.cs.uni-duesseldorf.de/AG/BI/Software/phyloCOP. In each pairwise comparison each genome is used once as BLAST database and once as query genome. Proteins – not nucleotide sequences – are aligned, since differences in synonymous substitution rates between genomes may lead to incomplete orthology assignments (compare Chapter 1.2.1). BLAST results are stored as a collection of folders – one for each genome - in which a folder includes the resulting files from pairwise alignments of one proteome with all others.

Required input data for a phyloCOP run are the pairwise BLAST sequence alignment results and one concatenated protein sequence FASTA file for each analyzed genome. An additional file with a list of all analyzed genomes in increasing phylogenetic distance to the reference genome must also be supplied by the user (see below and phyloCOP user's manual in Appendix_A, attached CD).

2.2.3. User-defined parameters and general overview

Once started, the user follows phyloCOP's command line instructions for setting the parameter values and provides all required data via standard input (in particular one protein sequences file per genome in FASTA format and the BLAST results). All required information, like the required user-defined parameters and paths to input and output files can also be provided as input command line options when calling phyloCOP. **Four user-defined parameters are set**:

1) An e-value cut-off for the BLAST results selection. The e-value (expectation value), shows how likely a BLAST alignment could be generated by chance. The default BLAST e-value cut-off is 10. The **quality of a BLAST alignment** with this value is low, since every BLAST run finds 10 matches with similar quality

by chance and not because of significant sequence similarity. Lower e-values imply BLAST alignment results of higher quality. But if the expectation value is too low, similarities of evolutionary distant genes may not be detected. Depending on the phylogenetic relatedness of the investigated dataset, an e-value cut-off is defined by the user. The analysis of distantly related genomes requires a higher e-value cut-off than orthology prediction for closely related genomes. BLAST alignments with e-values higher than the cut-off are discarded from orthology assignment (Altschul *et al.*, 1990).

2) The reference genome for cluster initialization. Note, that the name of the reference genome should be based on the name used for genome FASTA and BLAST result files (without the file endings like ".faa" and ".blast"). Dots and spacing characters should be avoided. If the reference sequence file is for example called "Escherichia_coli_K12.faa", the name of the corresponding BLAST folder "Escherichia_coli_K12" is the reference species name.

3) The order in which genomes are processed based on their phylogenetic distance to the reference genome is supplied by the user as text file. Genomes are clustered based on their **increasing phylogenetic distance to the reference genome** (or in random order for α = 1.0). As previously explained for the reference genome, genome names in this file must be identical to FASTA and BLAST file names without file endings.

4) The reciprocal hit degree cut-off parameter α, which determines the minimal fraction of genes in a cluster that have to be best BLAST hits of any newly added gene, which corresponds to the degree of transitivity of orthology relations. Higher α values increase connectivity between cluster members, which means a higher degree of transitivity between clustered orthologous genes. PhyloCOP allows α-values between 0.5 and 1.0. Lower cut-off values are not allowed, since phyloCOP aims the detection of orthologous clusters. Clusters with lower transitivity would be more or less identical to pairs of orthologs based only on their mutual sequence similarity. If the reciprocal hit degree of a gene to a cluster is > α (or = α, for α = 1.0) the gene is assigned to the cluster, after the paralogy check of phyloCOP is finished, which is also connected to α (compare section 2.2.4). A phyloCOP run with α = 1.0 produces exclusively clusters in which all members are reciprocal best sequence

similarity hits to each other, so called orthologous cliques. Cliques can be obtained in random genome order, since all cluster members have to be connected to one another.

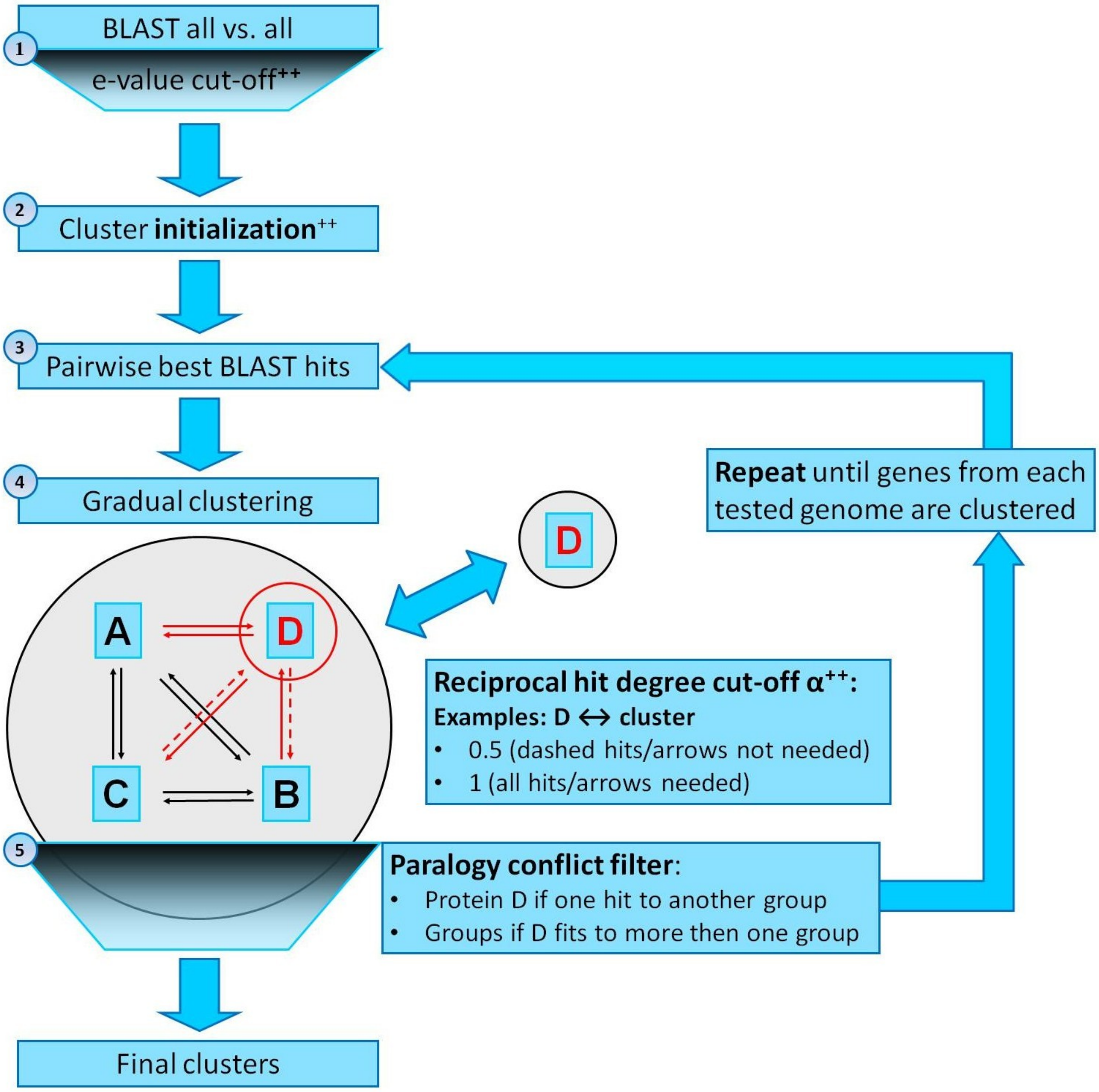

1
BLAST all vs. all
e-value cut-off++
2
Cluster initialization++
3
Pairwise best BLAST hits
4
Gradual clustering
A
D
C
B
D
Repeat until genes from each tested genome are clustered
Reciprocal hit degree cut-off α++:
Examples: D ↔ cluster
• 0.5 (dashed hits/arrows not needed)
• 1 (all hits/arrows needed)
5
Paralogy conflict filter:
• Protein D if one hit to another group
• Groups if D fits to more then one group
Final clusters

Figure 2.1: The phyloCOP algorithm consists of five steps.

1) selection of previously generated BLAST results by the user-defined e-value cut-off, **2)** initializing one cluster per reference gene, **3)** the recognition of pairwise best BLAST hits between all current species, **4)** the gradual clustering, in which a gene D is clustered into a group if its reciprocal hit degree to and from the group is above the user-defined level α, and **5)** a paralogy conflict filter, in which **a)** clusters are marked if they share the same genes and **b)** genes are removed from the analysis if they fit to only one cluster due to step 4 but have at least one additional reciprocal best BLAST hit to another cluster. Steps 4 and 5 are repeated for every gene-cluster combination. User-defined parameters and clustering steps are marked with ++.

Figure 2.1 shows a flow chart of the algorithm. **The algorithm consists of five steps: 1)** selection of BLAST results with e-values < user-defined e-value cut-off **2)** reference-based cluster initialization, **3)** obtaining pairwise best BLAST hits between already clustered genes and newly added genomes, **4)** the gradual clustering, in which a gene D is clustered into a group if its reciprocal hit degree to and from the group is above the user-defined level α, and **5)** a paralogy conflict filter, by which genes are removed if they are assigned to a cluster in step 4 but have a reciprocal best BLAST hit to another cluster, and clusters are marked if the same gene is assigned to them in step 4.

2.2.4 Preparative algorithmic steps

The algorithm starts with two preparative steps: **data selection and cluster initialization**. Pairwise genomic best BLAST hits are selected for further analysis if their e-value is smaller than the user-defined cut-off. Each gene in the reference genome initializes one cluster. The protein sequence of the gene is written in a file named by its gene identification number.

2.2.5 Iterative algorithmic steps

All following steps, starting from the third step – in which **pairwise best BLAST hits** between current cluster members and genes are determined – are **repeated** until **all species** have been processed. In each iterative step proteins of one genome are assigned to the clusters obtained in the previous iteration. This also means that the decision to assign a cluster in a previous step affects all following steps. If an orthologous relationship among the first two genomes was not detected, the algorithm might fail to detect orthologs from other genomes. Therefore, genomes are clustered in phylogenetic order with increasing phylogenetic distance to the reference genome (compare Chapter 2.2.1). The clusters grow after each step, which implies more proteins must be compared in the next iteration.

In the first iteration, clustering is simply based on the mutual best BLAST hits between the initial cluster proteins and the proteins of the closest relative of the reference genome. All selected reciprocal BLAST hits between the proteins of the reference genome and its closest relative are obtained. Proteins of the newly added genome with a reciprocal best BLAST hit to the initial cluster protein are assigned to the corresponding cluster. After the first iteration, clusters may include one or two proteins. Since only best BLAST hits are taken into account, no protein can be assigned to two clusters during the first iteration.

Paralogous relations among closely related proteins are automatically detected in later steps, when clusters are filled with more proteins.

Starting from the third genome, clusters may include more than one protein. The clustering procedure gets more complex (step 4) and the paralogy filtering step 5 takes effect. The flow chart in Figure 2.2 displays phyloCOP's assignment and paralogy filtering steps in detail.

After selected pairwise best BLAST hits between all cluster members and the proteins of the newly added genome are determined in the third step of each iteration, genes are **assigned to clusters** in the main fourth step of the phyloCOP algorithm if their **reciprocal hit degree** to and from the current cluster is higher than the user-defined level α (or equals α for α = 1.0) (see Figures 2.1 and 2.2). The hit degree f between gene and cluster is defined as the fraction of genes already in the cluster with a best BLAST hit to (and from) the examined gene: f = bbhs/N, where bbhs is defined as the number of best BLAST hits and N the number of genes in the current cluster. Note that, while at least a fraction α of hits is required in both directions (the tested gene D must be best blast hit in its genome for αN genes in the cluster, and αN genes in the cluster must be best blast hits in their genomes for D), the hits don't have to be reciprocal. Connectivity between genes in clusters created with α cut-offs < 0.5 is low for the detection of real transitive orthologous relations and may lead to more false orthology assignments. Thus, phyloCOP does not allow α < 0.5. Since cut-off values α ≥ 0.5 prevent the assignment of more than one protein per genome in the same orthologous group, each species can contribute at most one gene to a cluster.

The parameter α is also essential for the last selective step 5 of the clustering procedure – the **paralogy conflict filter**, which is directly connected to the protein assignment step 4. Two main actions are performed during step 5:

1) **Clusters are marked** as potentially involved in paralogy conflicts if they share some identical genes. This means in particular, that a protein fulfills the assignment criteria (see description for step 4) for more than one cluster and vice versa. If clusters include similar proteins they might be candidates for a fusion, since it can be assumed that they most probably resulted from ancestral gene duplications instead of being independent simple one-to-one orthologous clusters. If, for example, the same protein D is included in both clusters x and y and two different proteins have been assigned from all other genomes in each of the clusters, it is

likely, that the duplication event happened in the common ancestor of all species after the speciation event that led to D. More than one similar protein in clusters x and y may hint for more recent duplication events. For a detailed resolution of paralogous relations further steps would be necessary, but because of the main interest in function prediction and evolutionary studies in bacteria and archaea, these cases are simply excluded from subsequent analyses. PhyloCOP aims to detect simple one-to-one orthologous clusters and to separate them from paralogous clusters.

2) Some paralogous or co-orthologous relations might be hidden and only revealed, if additional genomes are included in the analysis. Therefore, **genes are removed** from the analysis if they fulfill the assignment criteria (step 4) for only one cluster but at the same time have at least one additional reciprocal hit to another cluster (Figure 2.2).

The paralogy conflict filter can also be seen as a combination of if-conditions with a web of marked clusters that serve as traps for other paralogy conflicts during the clustering process. The paralogy filter therefore takes also a retroactive effect on protein assignment that was done at the beginning of the clustering process.

Decisions in both steps 4 and 5 are based on the value of α. Higher α increases within-cluster connectivity but decreases the amount of clusters marked as problematic. The last two steps (steps 4 and 5) are repeated for each a cluster combination. The whole iterative clustering process (steps 3 to 5) is repeated in phylogenetic order, until all species are examined and final orthologous clusters are captured and separated from marked clusters. Final clusters – termed phyloCOPs – can be used for subsequent comparative functional analyses and further evolutionary studies.

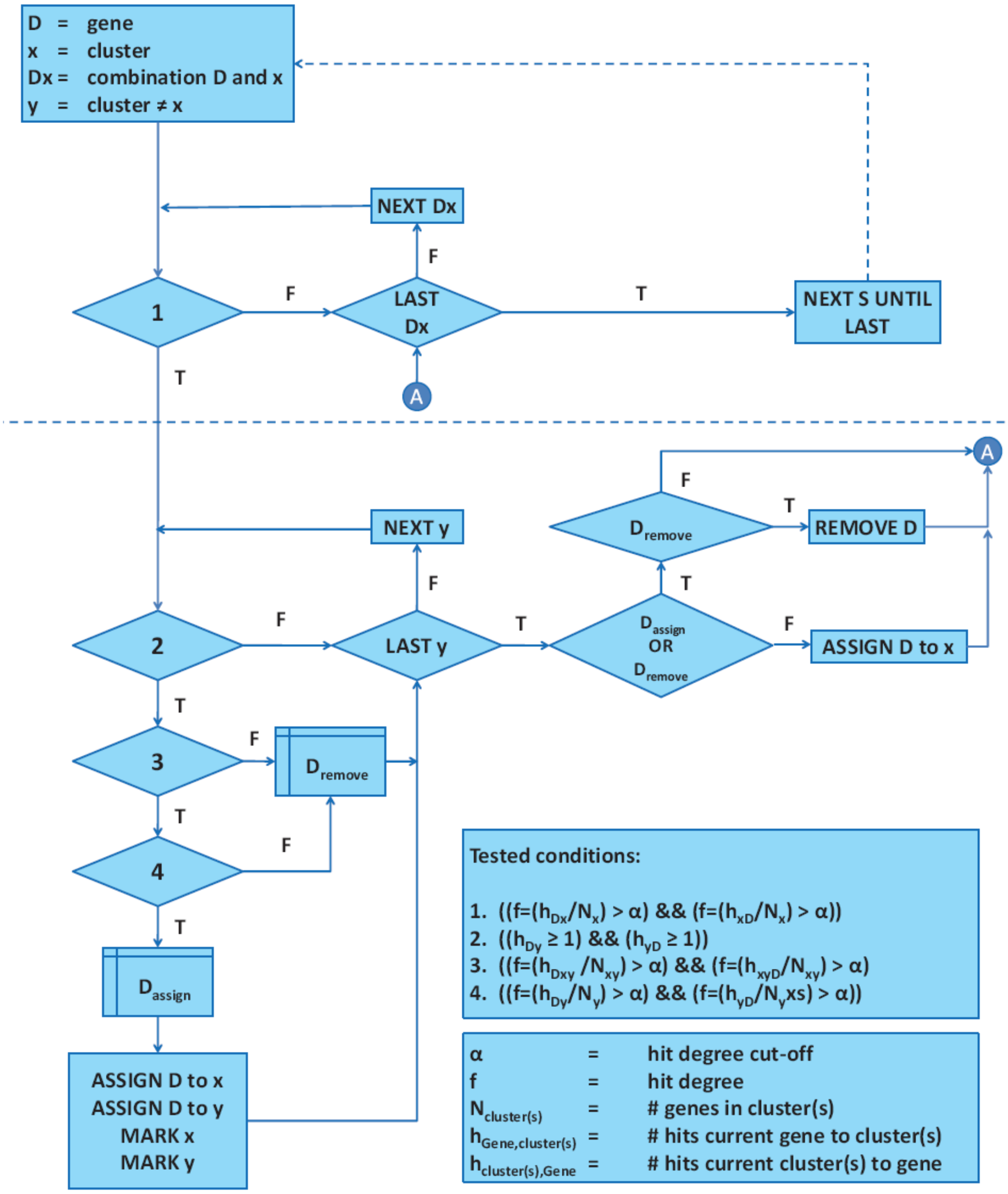

D = gene
x = cluster
Dx = combination D and x
y = cluster ≠ x
NEXT Dx
F
1
F
LAST Dx
T
NEXT S UNTIL LAST
T
A
A
F
D_{remove}
T
REMOVE D
NEXT y
F
T
2
F
LAST y
T
D_{assign} OR D_{remove}
F
ASSIGN D to x
T
3
F
D_{remove}
T
4
F
T
D_{assign}
ASSIGN D to x
ASSIGN D to y
MARK x
MARK y
Tested conditions:
1. $((f=(h_{Dx}/N_x) > \alpha)$ && $(f=(h_{xD}/N_x) > \alpha))$
2. $((h_{Dy} \geq 1)$ && $(h_{yD} \geq 1))$
3. $((f=(h_{Dxy} /N_{xy}) > \alpha)$ && $(f=(h_{xyD}/N_{xy}) > \alpha)$
4. $((f=(h_{Dy}/N_y) > \alpha)$ && $(f=(h_{yD}/N_yxs) > \alpha))$
α = hit degree cut-off
f = hit degree
$N_{cluster(s)}$ = # genes in cluster(s)
$h_{Gene,cluster(s)}$ = # hits current gene to cluster(s)
$h_{cluster(s),Gene}$ = # hits current cluster(s) to gene

Figure 2.2: Flow chart of phyloCOP's gene assignment and paralogy conflict filter step. This flow chart describes the connection between the gene assignment and paralogy filter step and provides a detailed overview of the filter mechanism. The paralogy filter is embedded into the gene assignment step, and separated from previous and subsequent steps by a dashed horizontal line. Gene assignment consists of a collection phase, in which the number of current cluster members is derived, and a selection phase including the paralogy filter. Three decisions are made: 1. A gene D is only processed by the filter if it fits to a current cluster x based on its reciprocal hit degree, which must exceed the user-defined level α (or must be equal α if α = 1.0). 2. If the first condition is true, phyloCOP checks whether D has at least one reciprocal hit with another cluster y. D is assigned to x if the second condition is false for every cluster y. 3. For a true second condition, the reciprocal hit degrees f between D and the union of x and y are compared to α. D is removed from the analysis if the reciprocal hit degrees

f between D and the union of x and y are < α. 4. If the third condition is true, and f of D from and to y > α, D is assigned to x and y and both clusters are marked for including paralogs when fused; in this case, x and y are likely the result of a duplication after the speciation event which led to D. After all genes from the currently examined species have been compared to all clusters, a cluster is marked for paralogy conflict if it includes several genes from the current species. Dashed arrows mark the transition to the next iteration, in which all steps are repeated with the next species. For α = 1.0, conditions 1, 3 and 4 are true, if the reciprocal hit degree = α.

After all genes from the currently examined species have been compared to all clusters, a cluster is marked as problematic if it includes several genes from the current species. Dashed arrows mark the transition to the next iteration, in which all steps are repeated with the next species. For α = 1.0, conditions 1, 3 and 4 are true, if the reciprocal hit degree is = α.

2.2.6. Output

In the final step, protein sequences of assigned orthologous genes are copied into previously initialized corresponding cluster specific FASTA files named by the corresponding reference gene IDs; these can be easily selected for further processing (e.g., alignment and phylogeny reconstruction). FASTA files of groups with paralogy conflicts are stored in a separate folder. Gene identification numbers of genes excluded by the paralogy conflict filter are saved in a separate file. The main output of phyloCOP is two tables. One of them is a list of all final orthologous clusters with the gene identification numbers of all clustered genes – excluding clusters marked for paralogy conflict. The other one is a separated list of all clusters with paralogy identification marks, indicating also paralogous connections between groups. Both tables include complete information about the gene composition of each orthologous cluster (see phyloCOP user’s manual in Appendix_A, attached CD). Note that the total number of phyloCOPs (including marked clusters with paralogy conflicts) always corresponds to the number of genes in the reference genome, since initialized clusters that only include the reference gene are not removed.

2.3 Dataset and data preparation

Different datasets have been prepared to evaluate the results of phyloCOP’s orthology assignment. Required data preparation steps are explained in this subchapter, regarding also the differences for both data evaluations discussed in this work.

2.3.1 Initial datasets

The protein sequences of all analyzed prokaryotic genomes were downloaded from GenBank via the NCBI ftp site (ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria). All genome sets are listed as tables in folder Appendix_B in the attached CD. Protein FASTA files were concatenated for each genome respectively and subsequently used as input for pairwise reciprocal protein BLAST runs until all genomes have been processed (see Chapter 2.2.2).

2.3.2 BLAST

The Basic Local Alignment Search Tool is a collection of different programs for sequence analysis. BLAST compares sequences to each other. One sequence is used as template to find similar parts of another sequence, which can be a genome or a database of genomes. The located overlaps are called local alignments. The score of an alignment indicates its significance. The alignment with the highest score is assumed to be the best local hit for the template sequence. The total score of an alignment is calculated with the help of a substitution matrix, which specifies a weighted identity score for each amino acid match and mismatch at every position of the sequence. Some aligned positions of one protein sequence have no amino acid counterpart in the other aligned protein. Such gaps are scored by a gap scoring system, which charges initial gaps with a higher penalty than subsequently connected gap positions, since one mutation may insert or delete multiple positions at once. The basic idea of the algorithm is that alignments with many matches should include short pieces with high similarity. Thus, the BLAST algorithm starts to search short pieces with fixed length and significant total score, which is the sum of the scores for each of matched or mismatched position minus the gap penalties. Neighboring pieces with significant total scores above a specific parameter are then connected to each other to get a coherent alignment in both directions. The best BLAST hit is the alignment with the highest total score (Altschul *et al.*, 1990). The command line version of the Basic Local Alignment Search Tool *blastall* can be downloaded from the NCBI website (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/).

Blastall needs two input formats for pairwise alignment. One sequence serves as query, the template for which a match is searched in a database. Each concatenated FASTA files is once used as query. *Blastall* compares each protein sequence in the query FASTA file to a database, which includes all protein sequences of another genome. For the pairwise alignments in this work, each genome served once as query and once as database. No genome was compared to itself. The database must have a specific format. Preformatted database files can be downloaded from ftp://ftp.ncbi.nih.gov/blast/db. The output of every pairwise *blastall* run is a table (option -m 8), in which the main information about each aligned protein pair is listed, like gene IDs, borders of the alignment, identity, score and e-value.

2.3.3 Phylogeny estimation

As previously described, phyloCOP assigns orthologous genes step-wise in the order of increasing phylogenetic distance to the reference genome. Phylogenetic distances among genomes can be obtained via different algorithmic approaches based on mutual sequence similarities of their protein sequences.

2.3.3.1 Multiple alignment and PhyML

PhyloCOP's phylogenetic run order for all comparisons based on the smaller dataset of 14 *E. coli* strains was previously obtained based on orthologous clusters determined by the SOG (Synteny-based Orthologous Genes) algorithm (Chapter 2.4.1.1; Esser, 2010). Protein sequences of clusters determined by SOG (here called SOGs), which include one gene from all 14 genomes, were aligned by **MUSCLE** (**multiple sequence comparison by log-expectation**) (Edgar, 2004). In general, multiple sequence alignment algorithms compare similarities of multiple sequences at the same time. This is a very complex, computationally intensive task, which is usually done via heuristics using a progressive greedy algorithm. MUSCLE was the method of choice because of its fast and accurate performance.

Subsequently, not well aligned, divergent parts of the multiple alignment were removed by **Gblocks** (http://molevol.cmima.csic.es/castresana/Gblocks.html) (Castresana, 2000). Alignments were concatenated and then used as input for PhyML, which estimates phylogenies by Maximum Likelihood (Guindon *et al.*, 2003). The general idea of phylogenetic tree reconstruction via Maximum Likelihood is to find the tree that maximizes the likelihood of the observed sequences under a chosen evolutionary substitution model. The likelihoods for all possible evolutionary scenarios are calculated, which means the probabilities of all ancestral amino acid compositions based on the evolutionary model used. The resulting tree has the most likely evolutionary history. In order to support the phylogeny estimation, I generated 100 bootstrap trees during the PhyML run. I used the JTT substitution model (Jones *et al*., 1992). Phylogenetic relatedness of each genome to the reference was finally obtained based on the topology of the resulting phylogenetic tree. Although the phylogenetic distance based on tree topology it is only an approximation of the real phylogenetic relations between prokaryotes (because of LGT), it is sufficient as guideline for transitive step-wise orthology assignment in the order of decreasing sequence similarity. A similar strategy was applied to determine the phylogenetic order for the phyloCOP run for the evolutionary analysis described in Chapter 3.

2.3.3.2 Pairwise BLAST alignment and PHYLIP Neighbor

PhyML could not obtain phylogenetic relations for the bigger datasets, because of a memory overflow. Thus, a computationally less expensive Neighbor-Joining method was used to determine the phylogenetic distances of all other genome sets. A distance matrix was generated from the mean score values of the 200 best mutual BLAST sequence similarity hits using the following formula:

$$d = 1 - \left(\frac{best\ BLAST\ \text{score mean}}{best\ \text{BLAST score mean max}} \right)$$

Distances d were calculated for all pairs of genomes. Values for the distances lie between 0 and 1. The smallest distance is 0 for the pair of genomes with the highest mean score (best BLAST score mean max). Subsequently, a phylogenetic tree was computed from the distance matrix with the PHYLIP Neighbor program, which is based on a Neighbor-Joining algorithm (http://evolution.genetics.washington.edu/phylip.html). Neighbor-Joining algorithms create relative phylogenetic trees from pairwise sequence similarities based on the idea of minimal evolution, normalized by the computation of mean distances for each genome to all others in every step of the algorithm (Saitou *et al.*, 1987). After that, the phylogenetic order was estimated from the branch distances. Genomes with shorter distance to the reference genome are closer related to it and were clustered first.

2.4 Evaluation and results

PhyloCOP's accuracy has been tested by **comparative evaluation**. At first phyloCOP was compared to similar methods by checking the mutual consistency of results. In a second evaluation, orthologous groups from a huge dataset of 539 bacterial and archaeal genomes were generated and the consistency of previously defined universally distributed COGs (Clusters of Orthologous Genes) was checked (Ciccarelli *et al.*, 2006). The model prokaryote *Escherichia coli* K12 was used as the reference genome in all tests.

2.4.1 Comparison of sequence similarity based orthology assignment methods

Two phyloCOP variants with the lowest and highest possible reciprocal hit degree cut-off (α = 0.5 and α = 1.0) were compared to other clustering concepts (see Table 2.1). The second phyloCOP variant, here named phyloCliquesOP, corresponds to searching fully connected cliques of genes; with α = 1.0, all members of a cluster have to be reciprocal best hits of each other. A simple **reference-based** algorithm is represented by starCOG (star wired Clusters of Orthologous Genes). SOG (Synteny-based Orthologous Genes) is reference-based like starCOG but refines orthology assignment via **synteny**, while orthology prediction with a **graph clustering algorithm** is represented by an application of MCL (the Markov Cluster Algorithm) (van Dongen, 2000). Flow charts of starCOG, SOG and MCL are presented in Figure 2.3. The e-value cut-off for the BLAST results was 10^{-10} in all tests except for SOG (e-value = 10^{-1}), were only BLAST results with > 70% of identical positions were used for clustering.

2.4.1.1 SOG (Synteny-based Orthologous Genes)

Unlike phyloCOP, SOG takes **synteny** into account, which avoids the clustering of most xenologous proteins (compare Chapter 1.2; Figure 2.3). Here, genes with similar sequences are only clustered if they are flanked by similar genes. As the number of genomic rearrangements increases with increasing evolutionary distance, synteny-based methods work best for closely related genomes. Like phyloCOP clusters, SOGs are initialized by reference genes. In contrast to phyloCOP, each genome is used as reference genome in turn. Genes with reciprocal sequence similarities (which must **not** necessarily be **best** BLAST hits) to genes of the reference genome and located in a similar chromosomal environment are added to clusters. The last step of SOG consists in building a consensus of the results with the different reference genomes; final clusters must be identical for more than half of the reference genomes (Esser, 2010). Comparisons between phyloCOP and SOG are interesting, because they reveal the role of synteny consideration.

2.4.1.2 StarCOG (star wired Clusters of Orthologous Genes)

StarCOG (star wired Clusters of Orthologous Genes) is a simplified variant of SOG that does not take synteny into account and avoids assignments to more than one cluster, as **only best** BLAST hits are clustered (Figure 2.3). Like phyloCOPs, starCOGs are initialized by **reference** genes. In contrast to phyloCOP, each genome is used as the reference genome in turn. Genes with reciprocal sequence similarities to the reference genome are added to clusters. The members of these intermediate clusters are only connected by their reciprocal sequence similarity to the reference protein and sequence similarities among all other cluster members are not taken into account. The last step of starCOG consists in building a consensus of the results with the different reference genomes; final clusters must be identical for more than half of the reference genomes. Members of final starCOGs are connected via mutual sequence similarities, like the proteins in phyloCOPs. Comparisons between phyloCOP and starCOG are interesting, because they reveal the effects of phyloCOP's reference centred algorithm and phylogenetic distance consideration.

In order to get a better impression of these effects, we also tested a simplified variant of starCOG: starCOG K12, which – like phyloCOP – generates clusters, based on just one reference species (here *E. coli* K12). Such a simple reference-based algorithm is often used, e.g., in the context of genome sequencing projects. In contrast to phyloCOP, genes in starCOG K12 must only fit to the reference gene. Therefore, usage is limited to genomes which are closely related to the reference.

2.4.1.3 Application of MCL (Markov Cluster Algorithm)

MCL obtains orthologous clusters from a graph, in which genes are vertices and edges represent sequence similarities between them. The BLAST output table includes all required information for initial graph assembly, and serves as standard input for orthology assignment with MCL (van Dongen, 2000; Enright *et al.*, 2002). In order to avoid assignment artefacts caused by shared domains, global alignment identities of reciprocal best BLAST hits are optionally used as input in a non-standard application of MCL (Dagan *et al.*, 2007) (Figure 2.3). The Markov Clustering Algorithm is based on **graph transition probability**

estimates. In the initial graph, pairwise alignment scores or global alignment identities are represented by edges with different weights.

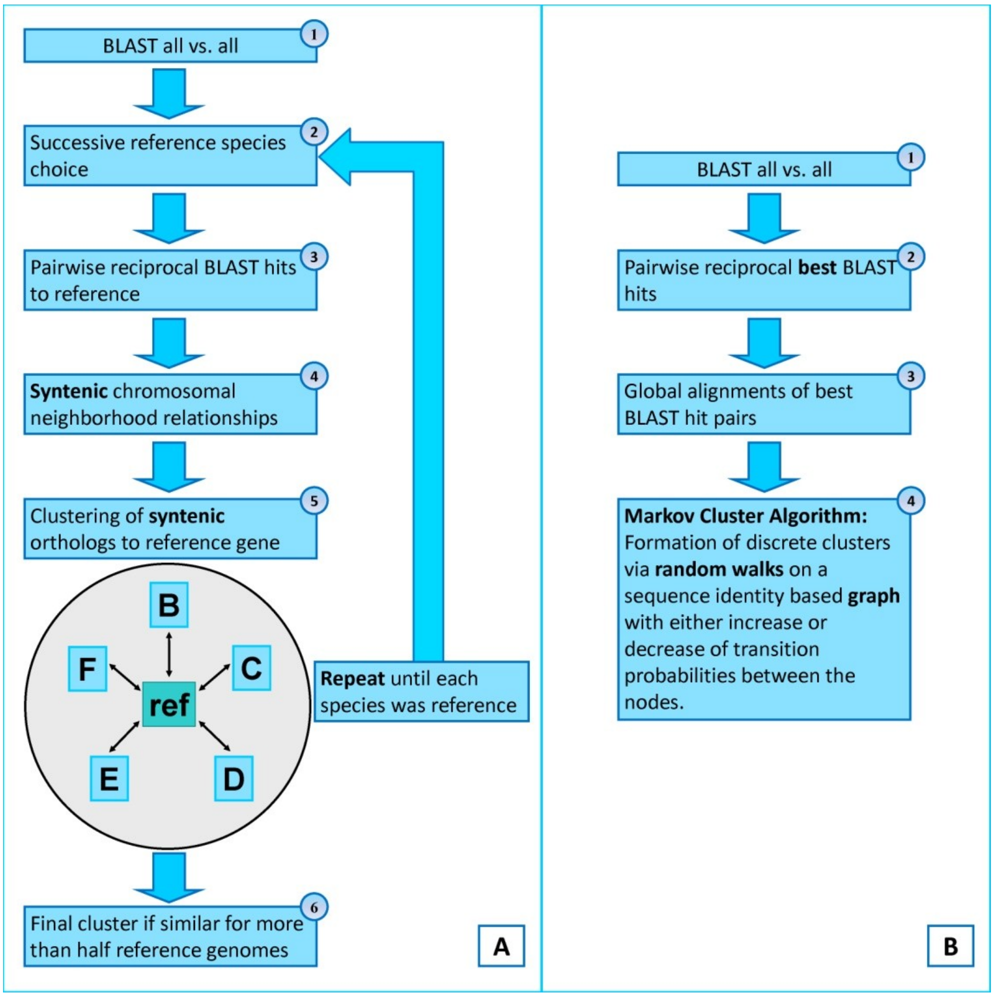

BLAST all vs. all
1
Successive reference species choice
2
Pairwise reciprocal BLAST hits to reference
3
Syntenic chromosomal neighborhood relationships
4
Clustering of syntenic orthologs to reference gene
5
B
F
C
ref
E
D
Repeat until each species was reference
Final cluster if similar for more than half reference genomes
6
A
BLAST all vs. all
1
Pairwise reciprocal best BLAST hits
2
Global alignments of best BLAST hit pairs
3
Markov Cluster Algorithm: Formation of discrete clusters via random walks on a sequence identity based graph with either increase or decrease of transition probabilities between the nodes.
4
B

Figure 2.3: Orthology clustering algorithms used for phyloCOP evaluation.

(A) SOG consists of 6 steps and is synteny based, while the starCOG algorithm works similar without taking synteny into account. The fourth step is skipped in starCOG, and genes are clustered due to their reciprocal **best** BLAST hits to reference genes. The workflow is repeated iteratively until each species has been used as the reference genome. In the last step, a consensus is built, in which clusters that are identical in more than half of the iterations (reference genomes) are assigned as final clusters.

(B) Application of MCL (van Dongen, 2000) based on reciprocal best BLAST hits. The third step of global alignment is optional. Depending on the dataset, either e-values of the reciprocal best BLAST hits or identities of global alignment results are used as MCL input.

Genes that belong to the same orthologous cluster are expected to be more similar to each other than to members of other clusters. The sequence similarity scores are used to generate a stochastic Markov Matrix that represents transition probabilities between proteins in the graph. This process, called expansion, is based on the expectation that much more edges with low similarity exist than pairs of genes, which makes the **probability** of a flow through **a single weak edge** very **low**. Two genes that are connected with many other genes – which are mostly genes that belong to one cluster – will have a low transition probability. Expansion corresponds to the simulation of flow through the graph.

The next step is called inflation, in which the Hadamard product (entry-wise product) of the matrix is calculated. If the transition probability in the Markov Matrix was low, the new matrix entry becomes high after inflation, while higher transition probabilities lead to a lower entry (Enright *et al.*, 2002). By alternating expansion and inflation steps, walks between proteins that belong to different orthologous cluster are step by step weakened and walks between proteins of the same orthologous cluster promoted until the graph consists of separated parts, which correspond to distinct orthologous clusters. Gene assignment is independent from genome affiliation, causing paralogy or co-orthology relationships within clusters; however, genes are never assigned to more than one cluster.

2.4.1.4 Comparison of algorithm complexity

The complexity of an algorithm is estimated by the **big-O notation**, which describes the worst case usage of computational resources (Bachmann, 1894; Ottmann und Widmayer, 2002). It is determined by the most time and memory consuming step of an algorithm, while multiplicative constants are discarded. The big-O notation is therefore especially useful for runtime estimation for algorithms with large input.

BLAST results are the main input for all tested methods. BLAST consist of three basic steps: 1) The generation of a list of words that have a higher alignment score than a threshold T with any words of the query sequence (default T = 11), 2) Scanning the database for matches to the words in the list, and 3) Extension of the word hits to get aligned sequence pairs with scores > cut-off S. The runtime for protein BLAST with one query sequence in big-O notation is:

$$O(aW + bN + cNW / 20^{w})$$

The runtime for the compilation of the word list depends on the number of generated words W. The speed of the database scan is correlated with the number of residues in the database N. The complexity of the final determination of BLAST pairs via hit extension is

NW divided by the maximum possible number of 20^w words with length w (default w = 3). Thus, BLAST runtime depends on the query sequence length and database size and can be modulated by word length and threshold T (Altschul *et al.*, 1990).

Let's define **M** as the **number of genomes** and **G** the average **number of genes per genome** (which corresponds to the number of query sequences for each BLAST input) in our analysis. If W=N, W<=O(N). If **L** is the average gene **length**, W=N=GL. The runtime for pairwise reciprocal BLAST runs of all versus all genomes is then:

$$O(M^2(2GL+(GL)^2/20^w))$$

For a maximal sequence size, W becomes constant:

$$O(M^2(GL+GL/20^w))=O(M^2GL+M^2GL/20^w)$$

The main term for the complexity for all versus all BLAST is:

$$O(M^2GL)$$

If only one genome is reciprocally aligned against all others (like for starCOG K12), the complexity of BLAST will reduce to:

$$O(MGL)$$

The **cluster initialization** step has the same complexity for both reference-based algorithms starCOG K12 and phyloCOP $O(G)$. It is $O(MG)$ for starCOG, since each genome serves once as reference.

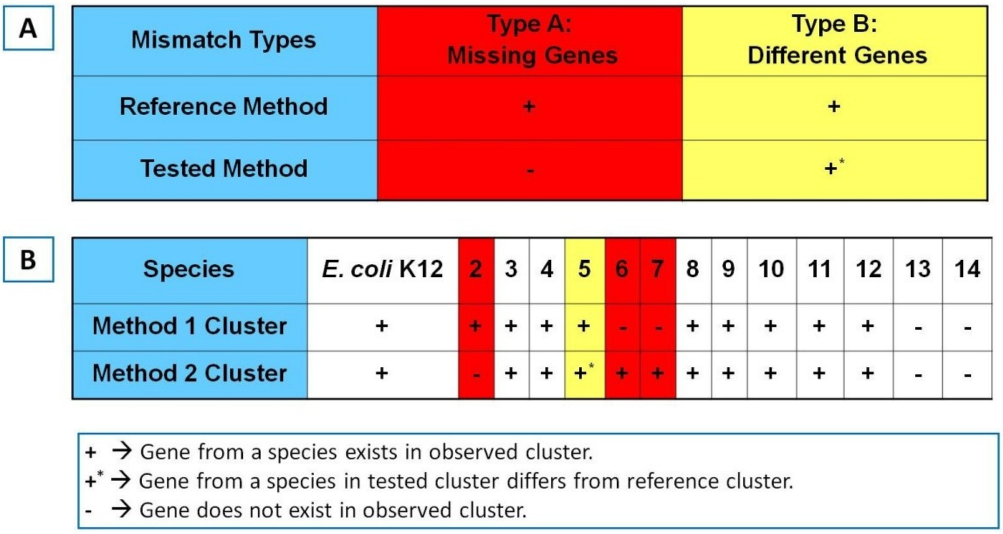
StarCOG K12 has the **lowest complexity** of all algorithms - $O(MG)$, followed by starCOG $O(M^2G)$ - which is basically a multiple run of starCOG K12 and subsequent choice of common clusters - and SOG $O(M^3G)$ (Esser, 2011 personal communication). In contrast to starCOG, starCOG K12 and phyloCOP algorithms that read the BLAST results separately for each pairwise comparison of two genomes, all pairwise BLAST results are read at the beginning of the SOG algorithm. Therefore, it is expected that SOG needs more memory, while the other methods require a higher number of hard disk accesses.

The complexity of MCL is $O(N^3)$, where N is the number of nodes in the graph (van Dongen, 2000). N corresponds to the number of genomes M multiplied by the number of genes per genome G. This leads to a complexity of $O(M^3G^3)$ for MCL, which is relatively high compared to the other methods, but reasonable because of the initial graph assembly.

PhyloCOP is as complex as MCL $O(M^3G^3)$. The most memory-consuming phyloCOP step is the storage of the reciprocal best BLAST hits between all proteins of the currently clustered genome and the clusters until the complete protein set of one genome is processed. However, phyloCOP assigns orthologous clusters from 539 genes over night with a memory usage below 2 GB. Despite its relatively low complexity in big-O notation, BLAST turned out to be the most time consuming step for the type of data analyzed in this thesis. Calculation of pairwise best BLAST hits between 539 genomes took one week using 10 CPUs. This is not surprising, since the runtime of BLAST depends on the sequence length, which is not the case for any of the tested orthology assignment methods.

2.4.1.5 Mismatch types

Methods were compared pairwise by a Perl script that uses the MySQL database for faster data processing (Levar, 2009). Only clusters which include the same *E. coli* K12 gene were compared. Clusters which include paralogs (potentially after a forced fusion with other clusters) were ignored. Two different mismatch types are specified. **Type A** mismatches are genes that are present in a cluster when built with one method, but missing when built with the alternative method. **Type B** mismatches are different genes from the same genome in the same cluster built with the alternative methods (see Figure 2.4). Summed mismatches are plotted in Figure 2.5.



Mismatch Types	Type A: Missing Genes	Type B: Different Genes
Reference Method	+	+
Tested Method	-	+*

Species	*E. coli* K12	2	3	4	5	6	7	8	9	10	11	12	13	14
Method 1 Cluster	+	+	+	+	+	-	-	+	+	+	+	+	-	-
Method 2 Cluster	+	-	+	+	+*	+	+	+	+	+	+	+	-	-

+ → Gene from a species exists in observed cluster.
+* → Gene from a species in tested cluster differs from reference cluster.
- → Gene does not exist in observed cluster.

Figure 2.4: Pairwise cluster method comparison.

(A) Mismatch types in pairwise cluster comparisons. Each method serves once as reference to which all other methods are compared.

(B) Cluster comparison example: Two clusters are compared to each other if their *E. coli* K12 gene is identical. Red and yellow colored positions mark mismatches. In this example three Type A and one Type B mismatches can be found. A similarity score for each compared cluster is derived from the fraction of matched positions. Finally, the number of all mismatches between two methods is calculated for each mismatch type separately.

2.4.1.6 Results for a dataset of 14 genomes

A dataset of 14 *E. coli* strains was used as input for method comparison (Appendix_B, attached CD). Because of comparisons to synteny-based SOG, only chromosomal genes have been taken into account. The total number of paralogy-free clusters is nearly the same for all examined methods (Table 2.1). The match score of most compared clusters is one or close to one, indicating a high identity between clusters produced by different methods. The fraction of assigned genes which have been differently classified into comparable clusters by the compared methods is shown in Table 2.2. All methods produced consistent results, with at most 2.5 % of differently assigned genes among phyloCliquesOP and SOG (Table 2.2).

Table 2.1: Compared variations of orthology assignment methods.

Two phyloCOP variants that differ in α, three SOG variants and a non-standard application of MCL were compared. The table displays the total number of identified (clusters total) and the number of **paralogy free** clusters, which include a gene from the reference genome ***E. coli* K12** (**clusters compared**). Note that the total number of clusters (including paralogy conflicts) derived from all phyloCOP applications always corresponds to the number of proteins in the reference genome, even if some of the clusters include only the reference gene and are otherwise empty. The same is true for the simple reference-based starCOG K12.

Assignment method	Variation	Clusters total	Clusters compared
phyloCOP	hit degree cut-off α = 0.5	4132	4126
phyloCliquesOP	hit degree cut-off α = 1.0	4132	4132
SOG	synteny	6497	4005
starCOG	no synteny	4047	4041
starCOG K12	no synteny, ref. *E. coli* K12	4132	4132
MCL	reciprocal best BLAST hits	6872	3995

Table 2.2: Fractions of mismatched genes comparing two methods respectively.

Type A mismatches for both methods as reference are **added** together. The fraction of genes that are missing in comparable clusters created by one of the methods (mismatch Type A) is low, with a maximum of **2.4 %** (*phyloCliques+SOG). The percentage of mismatch **Type B** genes, which have been assigned into different clusters, is close to 0 for all comparisons.

Compared methods	% Mismatch Type A	% Mismatch Type B
phyloCOP+phyloCliquesOP	1.3	0.01
phyloCOP+SOG	1.9	0.1
phyloCOP+starCOG	0.02	0
phyloCOP+starCOG K12	0.2	0.01
phyloCOP+MCL	0.4	0.002
phyloCliquesOP+SOG	**2.4***	0.1
phyloCliquesOP+starCOG	0.7	0
phyloCliquesOP+starCOG K12	1.3	0
phyloCliques+MCL	1.1	0.004
SOG+starCOG	1.5	0.1
SOG+starCOG K12	2.0	0.1
SOG+MCL	1.1	0.1
starCOG+starCOG K12	0.1	0.002
starCOG+MCL	0.1	0.002
starCOG K12+MCL	0.4	0.01

The tested methods mainly differ in their inclusiveness. **Gene absence** in one of two compared clusters (mismatch **Type A**) occurs more frequently than **gene differences** (mismatch **Type B**) (Tables 2.3 and 2.4; Figures 2.5 and 2.6). The highest number of mismatched proteins is found between **phyloCliquesOP** and **SOG**. At the same time, both are more **exclusive** than all other methods (Tables 2.2 and 2.3; Figure 2.5). StarCOG K12 and phyloCOP are the most inclusive applications, with the lowest number of 'missing' proteins. Results of both starCOG variants and phyloCOP are very similar, suggesting that a method based on one reference only is sufficient for closely related genomes.

However SOG's results differ from the others, since it clusters some proteins that are not assigned by any of the other methods except MCL. SOG assigns a small number of genes into different groups (Table 2.4; Figure 2.6). The reason might be its consideration of synteny (Esser, 2010). MCL's graph clustering algorithm and the computationally less expensive starCOG give very similar results.

In general, results of all tested methods are similar, indicating that phyloCOP works well for the tested dataset of closely related genomes. However, the intended application of phyloCOP is the analysis of large, phylogenetically diverse datasets. PhyloCOP's reference centered algorithm and phylogenetic distance consideration are expected to facilitate the analysis of more distantly related genomes.

Table 2.3: Pairwise method comparison - Mismatch type A.
This table shows the number of genes clustered by one method (referred to as reference method), missing in the compared clusters generated by the other method (referred to as tested method). **Reference methods are written in columns and tested methods in rows** (plotted in Figure 2.5)**.**

		Genes identified with:					
		phyloCOP	**phyloCliquesOP**	**SOG**	**starCOG**	**starCOG K12**	**MCL**
Genes missing in:	**phyloCOP**	-	50	247	7	106	50
	phyloCliquesOP	601	-	572	340	672	50
	SOG	725	642	-	530	793	488
	starCOG	5	10	224	-	26	21
	starCOG K12	0	0	226	0	-	28
	MCL	132	82	62	51	188	-

Table 2.4: Pairwise method comparison - Mismatch type B.

This table shows the number of differently assigned genes in compared clusters generated by two different methods (plotted in Figure 2.6).

	phyloCOP	phyloCliquesOP	SOG	starCOG	starCOG K12	MCL
phyloCOP	-	3	56	0	5	1
phyloCliquesOP	3	-	46	0	0	2
SOG	56	46	-	48	55	46
starCOG	0	0	48	-	1	1
starCOG K12	5	0	55	1	-	3
MCL	1	2	46	1	3	-

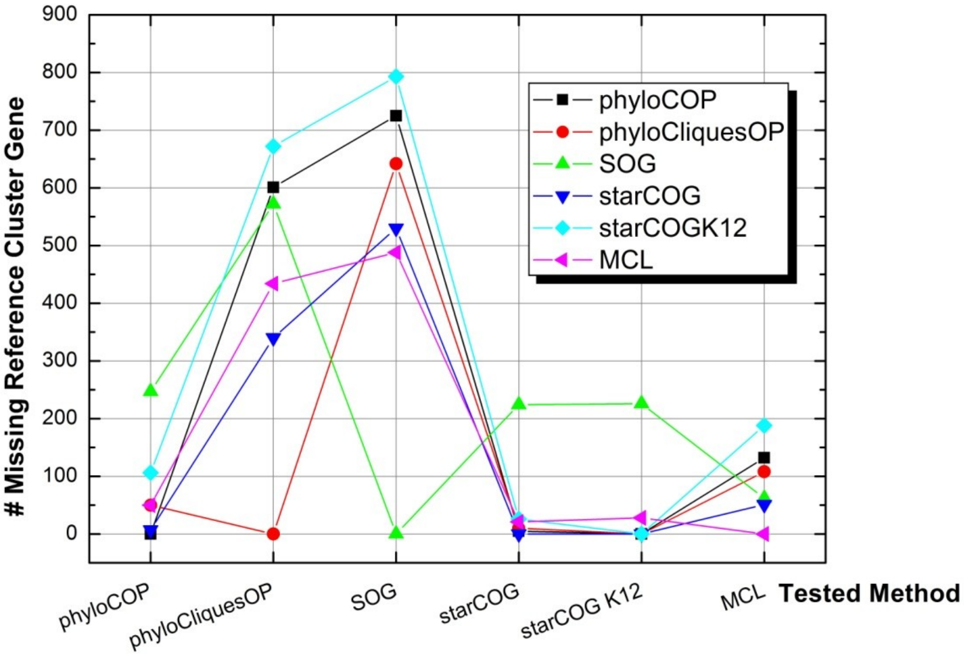


Figure 2.5. Results of pairwise cluster method comparison.

Distribution of mismatch **Type A**. Number of genes included in the clusters generated by the reference method which are missing in clusters determined by the tested method. Connections between symbols serve as marker for the behavior of each reference in all comparisons. SOG is the application with the highest number of missing reference cluster proteins. StarCOG K12 reveals the lowest number of missing reference cluster proteins compared to the other methods.

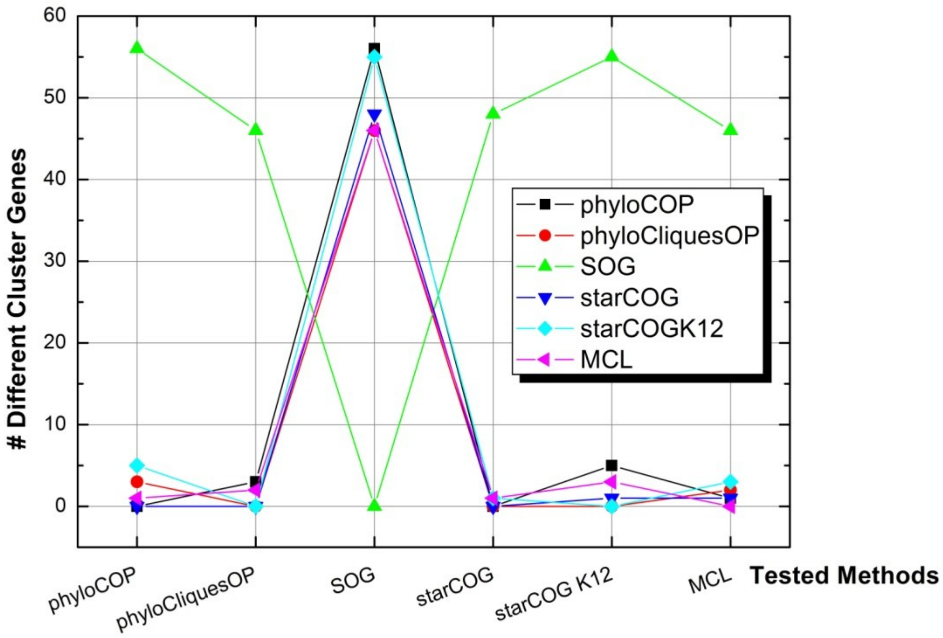


Figure 2.6: Results of pairwise cluster method comparison.
Distribution of mismatch **Type B**. Number of all gene differences between corresponding clusters of two compared methods. Connections between symbols serve as marker for the behavior of each reference in all comparisons. SOG is the only method that differs significantly from the others. The occurrence of gene differences is low compared to the absence of genes (see Figure 2.5).

2.4.1.7 Results for a dataset of 539 distantly related genomes

To test phyloCOP's performance on datasets of distantly related genomes, we applied phyloCOP, starCOGK12 and MCL to 539 bacterial and archaeal genomes of broad phylogenetic diversity. Since the e-value cut-off showed a significant impact on the integrity, defined here as completeness, of orthologous clusters that correspond to 7 universally distributed COGs, it was chosen based on the distribution of best BLAST e-values of the corresponding 7 *E. coli* K12 proteins to all others (compare Chapter 2.4.2, Figure 2.10). Only best BLAST hits with an e-value > 10^{-3} were used for orthology assignment. For MCL we did not perform the optional global alignment but directly used the reciprocal best BLAST hits as input.

Histograms of the cluster size distributions of clusters created by phyloCOP, starCOG K12 and MCL look similar (Figures 2.7, 2.8 and 2.9). Because of the phylogenetic divergence between the genomes most clusters include only a small number of genes. PhyloCOP generated 3167 paralogy-free clusters. StarCOG K12 always detects as many clusters as reference proteins, which are 4131 for *E. coli* K12. 3738 MCL clusters from the huge dataset include *E. coli* K12 genes, but only 1376 are paralogy-free and therefore comparable.

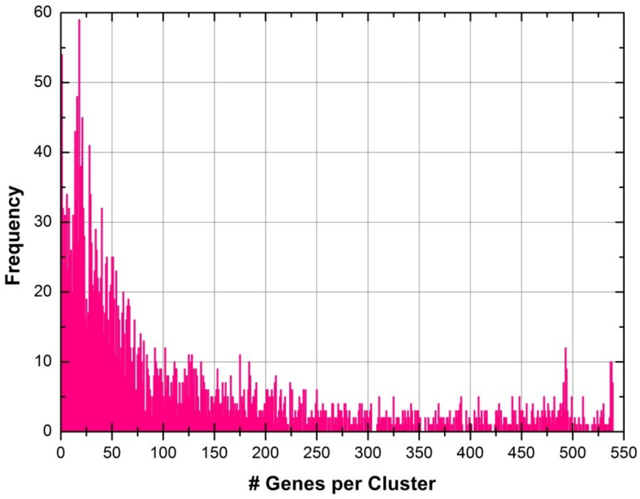


Figure 2.7: Histograms of phyloCOP cluster sizes. Distribution of orthologous clusters generated with phyloCOP: The histogram shows frequencies of protein cluster sizes for the 3167 clusters generated with phyloCOP from 539 completely sequenced prokaryotic genomes with *E. coli* K12 as reference species, an e-value cut-off of 10^{-3} and a reciprocal hit degree α = 0.5. 964 clusters were marked by the paralogy conflict filter and thus excluded from comparative evaluation.

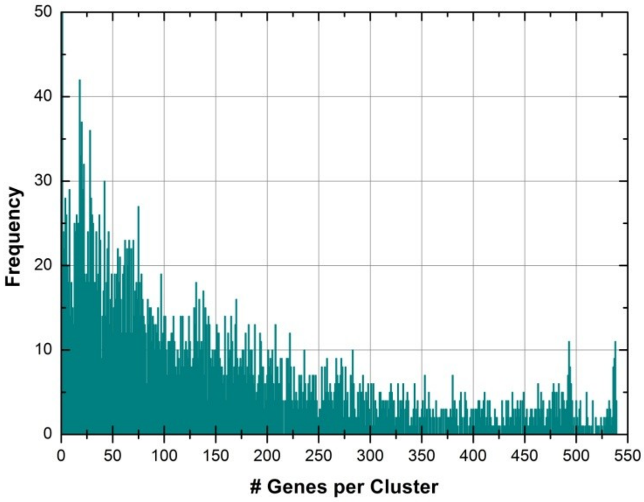


Figure 2.8: Histogram of starCOG K12 cluster sizes generated from 539 genomes.

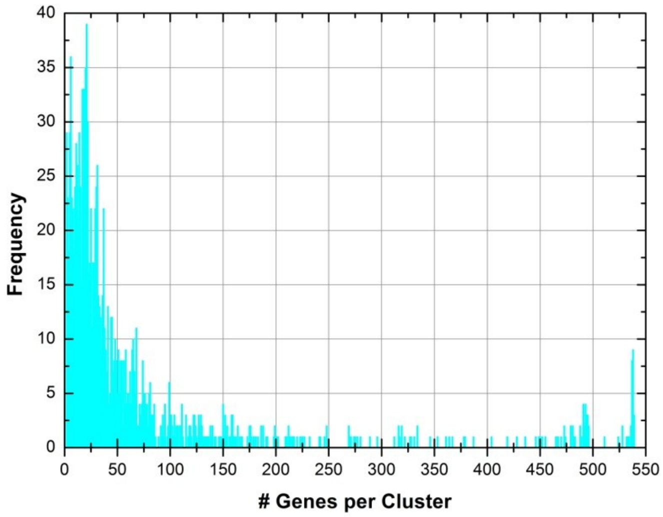


Figure 2.9: Histograms of MCL cluster sizes. Orthologous clusters generated by MCL from 539 genomes: Out of 95054 MCL clusters, 1376 paralogy-free *E. coli* K12 gene including clusters were detected. The MCL cluster size distribution is similar to that one of phyloCOP and starCOG K12.

In contrast to the small-scale *E. coli* analysis, MCL assigned only half the number of *E. coli* K12 genes to paralogy-free clusters: 257 of these MCL clusters cannot be found in the phyloCOP results, because they were excluded for paralogy there. Therefore, only 1117 clusters are comparable between the methods. As in the analysis of the smaller dataset, the appearance of mismatch Type B is very low (49 mismatches, which are about 0.06 % of all genes assigned in comparable clusters). The frequency of genes, which were **assigned by only one** of the methods, is **much higher**: 10.8 %, where MCL contributes > 2/3 of the missing genes and phyloCOP < 1/3. The number of comparable clusters between starCOG K12 and MCL was only slightly higher (1375).

Table 2.5: Number of type A mismatches between orthologous clusters created with three orthology assignment methods from 539 distantly related genomes.
This table shows the number of genes clustered by one method (referred to as reference method), missing in the compared clusters generated by the other method (referred to as tested method). **Reference methods are written in columns and tested methods in rows.**

		Genes identified with:		
		phyloCOP	**starCOG K12**	**MCL**
Genes missing in:	**phyloCOP**	-	103983	7219
	starCOG K12	3720	-	3521
	MCL	2047	15137	-

In contrast, 3167 starCOG K12 and phyloCOP clusters were comparable. Tables 2.5 and 2.6 give an overview of the number of mismatched proteins between all clusters. StarCOG K12 is the most inclusive application with much higher number of proteins that cannot be found in corresponding MCL clusters or phyloCOPs (Table 2.5). Since paralogy is not recognized by starCOG K12, this is maybe due to false positive orthology relations. In addition, the number of type B mismatches between phyloCOP and starCOG is higher than in any other comparison.

Table 2.6: Number of type B mismatches between orthologous clusters created with three orthology assignment methods from 539 distantly related genomes.
This table shows the number of differently assigned genes in compared clusters generated by two different methods.

	phyloCOP	starCOG K12	MCL
phyloCOP	-	2696	49
starCOG K12	2696	-	171
MCL	49	171	-

2.4.2 Comparison to universally distributed Clusters of Orthologous Genes

It has been hypothesized that a small set of universally distributed genes exists; such universal clusters should include at least one gene from almost all species, regardless of their phylogenetic diversity. Such ubiquitous genes (e.g., encoding ribosomal proteins) are mostly primordial with a long phylogenetic history (Ciccarelli *et al.*, 2006; Chapter 1.3).

PhyloCOP clusters created from 539 completely sequenced prokaryotic genomes are compared to 30 previously defined universally distributed COGs (Clusters of Orthologous Genes), which include genes from 168 prokaryotic and 23 eukaryotic species (Ciccarelli *et al.*, 2006; Appendix_B). Each universal COG includes one *E. coli* K12 gene. Corresponding gene IDs have been retrieved from the STRING database and NCBI (http://string-db.org/newstring_cgi/show_input_page.pl; http://www.ncbi.nlm.nih.gov/sites/entrez).

Most orthologous clusters of genes from diverse species are sparse, since many genes appear to be adaptations to specific environmental conditions, or appear only in selected phylogenetic branches. This is also true for phyloCOP, starCOG K12 and MCL clusters (Figure 2.6).

Table 2.7: Universally distributed COGs and corresponding *E. coli* K12 gene IDs.**
#Marked COGs include paralogs.

Universal COGs*	*Escherichia coli* K12**	Universal COGs*	*Escherichia coli* K12**
COG0012	16129166#	COG0099	16131177
COG0016	16129670	COG0100	16131176#
COG0048	16131221	COG0102	16131121
COG0049	16131220#	COG0103	16131120
COG0052	16128162	COG0172	16128860#
COG0080	16131813#	COG0184	16131057
COG0081	16131814	COG0186	16131190#
COG0087	16131199	COG0197	16131192
COG0091	16131194	COG0200	16131180
COG0092	16131193	COG0201	16131179#
COG0093	16131189	COG0202	16131174#
COG0094	16131187#	COG0256	16131183
COG0096	16131185	COG0495	16128625#
COG0097	16131184	COG0522	16131175#
COG0098	16131182	COG0533	16130960

*Ciccarelli *et al.*,2006

**http://string-db.org/newstring_cgi/show_input_page.pl; http://www.ncbi.nlm.nih.gov/sites/entrez

In order to evaluate phyloCOP's assignment results, we checked the **integrity** of phyloCOP clusters that correspond to the 30 universally distributed clusters based on included *E. coli* K12 genes (see Table 2.7); this was compared to the integrity of the corresponding starCOG K12 and MCL clusters. A cluster has the highest integrity if it includes a gene from each analyzed genome.

A less stringent e-value cut-off was required for the large scale analysis because of the lower sequence similarities of orthologous proteins from distantly related species. In order to find a suitable e-value cut-off that is sufficient for the detection of universally distributed orthologous proteins in distantly related species, phyloCOP and MCL were tested with e-value cut-offs 1 and 10^{-10}. The integrity of 7 of 30 universal clusters was much better for the e-value cut-off = 1 than for the e-value cut-off = 10^{-10}. The histogram of best BLAST e-value distributions for 7 *E. coli* K12 proteins that correspond to these clusters shows that an e-value of 10^{-3} is sufficient to include most proteins with a best BLAST hit with one of the 7 universally distributed proteins in the analysis and to exclude a peak of non-significant BLAST hits above. The e-value cut-off = 10^{-3} was used in all comparative analyses with the large dataset.

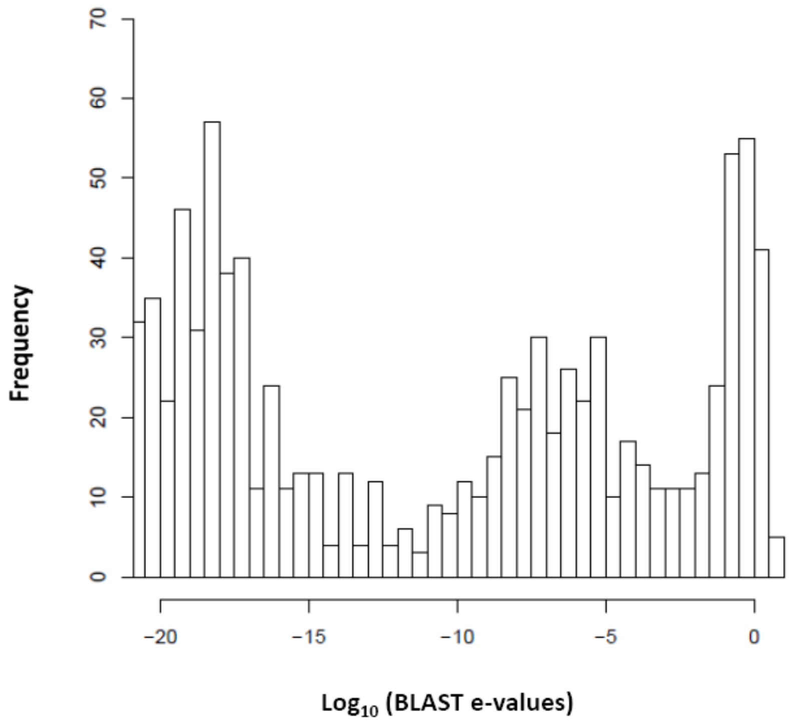


Figure 2.10: Histogram of the log-scale distribution of e-values of best BLAST hits with 7 universally distributed *E. coli* K12 proteins.

None significant best BLAST hits can be excluded using the e-value cut-off 10^{-3} in all analyses at only a low cost of not detected universally distributed orthologs. Most best BLAST hits with the 7 observed proteins have an e-value of 10^{-50} (not shown in the graph).

Figure 2.11 shows how many genomes are included in phyloCOP's, starCOG K12's and MCL's counterparts of the universal COGs. The integrity of all 30 examined phyloCOPs is above 490 (out of 539) species. 22 clusters include genes from more than 530 genomes. All universally distributed COGs have been found as well by the application of starCOG K12 and MCL, with comparable integrity to the corresponding phyloCOPs. Only 5 universally distributed phyloCOPs, 1 starCOG K12 cluster and 4 MCL clusters include less than 500 genomes (Figure 2.11). The slightly higher integrity of the starCOG K12 clusters is most probably a reflection of non-significant orthology assignment, since many proteins assigned by starCOG K12 are missing in comparable phyloCOP and MCL clusters (mismatch Type A, Table 2.5), and a relatively high number of proteins is differently assigned by starCOG K12 (mismatch type B, Table 2.6). 11 universally distributed COGs include a low number of paralogs, which were not detected in the corresponding phyloCOPs (Table 2.7; Ciccarelli *et al.*, 2006). 12 universal MCL clusters include paralogs, of which 10 are paralogous in the universal COGs (compare Tables 2.7 and 2.8), while paralogs cannot be detected by starCOG K12. Increasing the sensitivity of phyloCOP's paralogy detection by changing the paralogy filter criteria may be required to find of the low number prokaryotic paralogs in the universal COGs. However, it is not described, which genomes contribute more than one gene in a universal COG (Ciccarelli *et al.*, 2006).

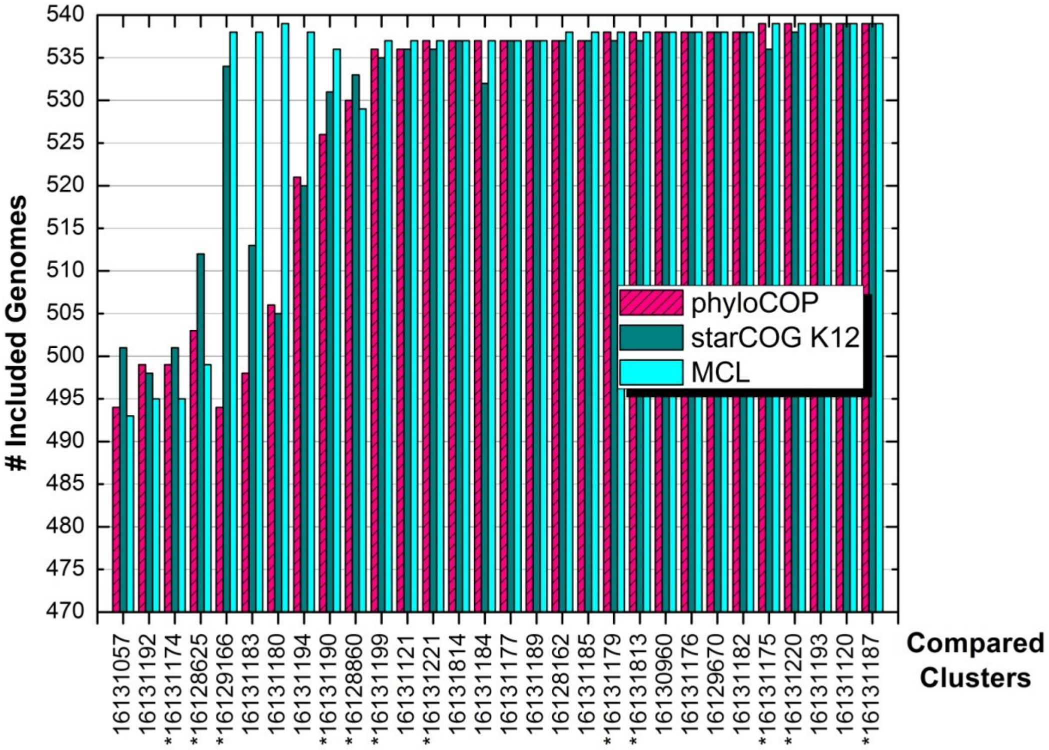


Figure 2.11: Integrity of universally distributed clusters generated by phyloCOP, starCOG K12 and MCL. Corresponding phyloCOP, starCOG K12 and MCL clusters identified by the *E. coli* K12 gene IDs (all clusters include only one *E. coli* K12 gene). The maximal number of included genomes is 539. *MCL clusters that include more than one gene per genome.

Table 2.8: 12 universally distributed MCL clusters include paralogs.

Twelve universally distributed COGs found by MCL include more than one gene from some genomes. For two of this twelve MCL clusters no paralogous proteins were detected in the corresponding universal COGs (compare Ciccarelli *et al.*, 2006).

***E. coli* K12 gene ID**	**# genomes**	**# paralogs**
16131190	491	1
16131221	493	1
16131199	494	2
16131174	495	8
16131175	495	15
16128625	499	2
16128860	529	7
16131179	538	8
16129166	538	6
16131220	539	1
16131813	539	1
16131187	539	1

Most of the missing archaeal and bacterial species in the universal phyloCOPs are very distantly related to *E. coli* K12, which turned out to be the reason for exclusion. After using phyloCOP for orthology detection among 43 archaeal genomes with *Methanosarcina acetivorans* as a reference, nearly all universally distributed clusters were completely filled (compare Table 2.9 and Figure 2.12). A list of the 43 archaea is supplied in Appendix_B in the attached CD.

Table 2.9: Universally distributed COGs and corresponding reference genome gene IDs.** *Methanosarcina acetivorans* was used as reference genome for archaeal orthology assignment analysis. For one universal COG (COG0202) no gene could be found in *Methanosarcina acetivorans*.

Universal COGs*	***Methanosarcina acetivorans*****	**Universal COGs***	***Methanosarcina acetivorans*****
COG0012	20093188	COG0099	20089977
COG0016	20089069	COG0100	20089979
COG0048	20090123	COG0102	20089485
COG0049	20090122	COG0103	20089486
COG0052	20089489	COG0172	20092841
COG0080	20093063	COG0184	20089827
COG0081	20093064	COG0186	20089951
COG0087	20089942	COG0197	20089081
COG0091	20089947	COG0200	20089964
COG0092	20089948	COG0201	20089965
COG0093	20089952	COG0202	NA
COG0094	20089955	COG0256	20089961
COG0096	20089957	COG0495	20090469
COG0097	20089958	COG0522	20089978
COG0098	20089962	COG0533	20092505

* Ciccarelli *et al.*,2006

** http://string-db.org/newstring_cgi/show_input_page.pl; http://www.ncbi.nlm.nih.gov/sites/entrez

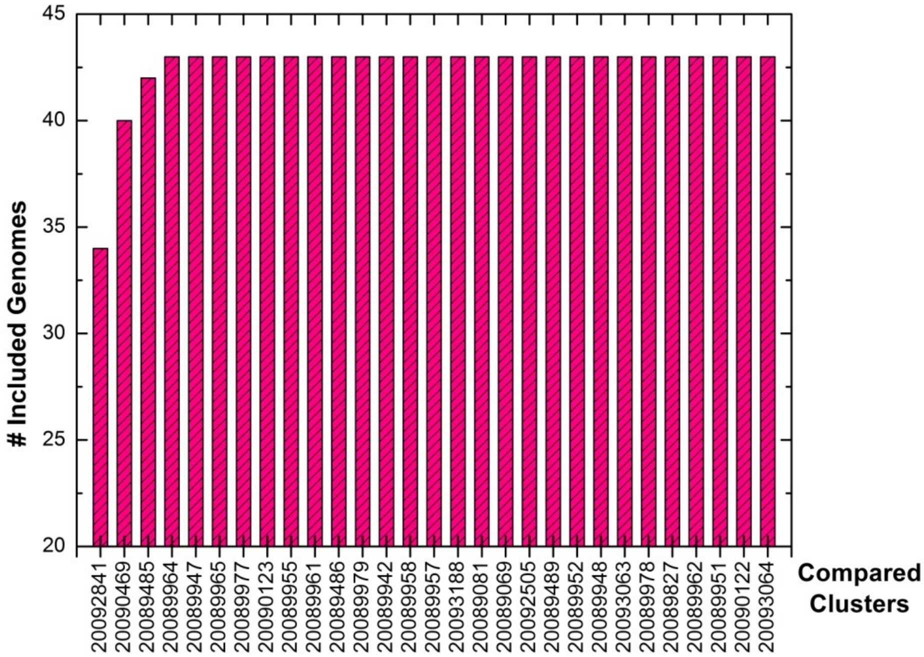


Figure 2.12: Integrity of archaeal genomes in corresponding universally distributed phyloCOPs. Corresponding phyloCOPs identified by *Methanosarcina acetivorans* gene IDs (all clusters include only one *Methanosarcina acetivorans* gene).

2.5. Discussion

The phylogeny-based orthology assignment algorithm phyloCOP has been developed for comparative analyses of diverse prokaryotic genomes. Unlike other algorithms, which assign orthologous genes in a pairwise manner, phyloCOP uses transitivity of orthology to build clusters. This leads to the detection of distant orthologous relationships, which cannot be achieved by simple pairwise sequence similarity consideration. Due to user-defined parameters, phyloCOP can be tailored to various **research interests** and adapted to datasets of different phylogenetic diversity.

Application of phyloCOP to a dataset of 14 closely related genomes revealed high consistency with the results of other sequence similarity- and transitivity-based algorithms. As expected, **synteny** consideration affects the results, which slightly differ from the others, because xenologs that do not share the same chromosomal environment are filtered (Dagan *et al.*, 2007). Note that synteny consideration is limited to orthology detection in closely related genomes because synteny among orthologous genes gets lost during evolutionary time (compare Chapter 1.2.2). The results of complex **graph clustering** with MCL do not significantly differ from phyloCOP or even a **simple reference-based** algorithm like starCOG K12. It appears that closely related orthologous proteins are clustered well by all tested algorithms. Depending on the chosen reciprocal hit degree cut-off α, phyloCOP can be adjusted for being more inclusive or exclusive in orthology assignment and paralogy detection. PhyloCOP with the reciprocal cut-off value α = 0.5 is one of the most inclusive methods for the small dataset analysis together with starCOG K12.

For further evaluation, phyloCOP was used to assign orthologous genes from a **large and diverse dataset** of 539 genomes. Resulting clusters were compared to starCOG K12 MCL clusters from the same dataset. The inclusiveness of starCOG K12 is much higher compared to phyloCOP and MCL for the large dataset. Also a relatively high number of proteins was assigned differently compared to phyloCOP. Most probably starCOG K12 assigned a large number of false positive orthologs, since paralogy is ignored and proteins in starCOG K12 clusters must only have a reciprocal best BLAST hit to the reference proteins, ignoring transitive orthology relations to other proteins in the cluster. Therefore, it can be assumed that a restriction of the phyloCOP α cut-off values between 0.5 and 1 is reasonable.

Like in the small-scale comparison, most of the genes clustered by phyloCOP and MCL were assigned into corresponding clusters, while the fraction of genes, clustered by phyloCOP but not by MCL and vice versa was much higher than for the small dataset. In contrast to phyloCOP the application of MCL is not reference-based. Thus, the majority of MCL clusters (about 96 %) were excluded from comparisons, because they included no *E. coli* K12 gene. About 60 % (2362 out of 3738) of the remaining MCL clusters were excluded because of paralogy, while only about 20 % (965 out of 4132) of the phyloCOPs were excluded for the same reason. This is a big difference and the question is if the larger number of paralogs detected with MCL reflects biological reality or not, since gene duplication in prokaryotes is not as frequent as in eukaryotes. Concerning the larger dataset, **paralogy** assignment seems to be very **different** in both methods. Among the small number of comparable clusters, MCL clusters include a large number of proteins that were not detected in corresponding phyloCOPs. Unlike phyloCOP, MCL does not consider phylogenetic distance during the clustering process, which may lead to false orthology assignments (see Chapter 2.2). In general, successful application of MCL to orthology assignment relies strongly on diverse preparative steps that are not part of the standard application of MCL (van Dongen, 2000, Enright *et al.*, 2002). In addition, paralogous clusters are not automatically removed by MCL and the output is only ordered by cluster size. Thus, subsequent usage of MCL clusters for further analyses requires additional preparative steps. PhyloCOP clusters may be the better choice for later **function prediction**, based on a reference genome.

The **inclusiveness** of phyloCOP is similar to starCOG K12 and MCL considering universally distributed COGs. 12 of the corresponding universal MCL clusters include more than one protein from one genome. Ciccarelli *et al.* detected paralogous proteins of prokaryotic origin in 10 corresponding universal COGs. However, it is not mentioned which proteins are paralogous in the universal COGs.

PhyloCOP detects most of the universal orthologs among diverse genomes. Most absent proteins belong to distantly related bacterial and archaeal genomes that may not be captured by transitive orthology relations to the other cluster members. When applied to only archaeal genomes, the integrity of phyloCOPs that correspond to universal COGs was almost complete (Figure 2.12).

All results of the comparative evaluations demonstrate phyloCOP's ability to detect transitive connections between closely and most of the distantly related orthologs. PhyloCOP assigns paralogy-free orthologous clusters, which can be used for function prediction and subsequent analysis of the evolutionary history of cell function.

Chapter 3 Tracing back the evolution of metabolic networks

Metabolic network components, which are basically metabolites and reactions, have been characterized for multiple genomes by genome annotations, biochemical experiments and cell physiology experiments. Complex systems like genome-scale metabolic networks are often described mathematically, based on the list of all involved metabolites and reactions (compare Chapter 1.3). Reaction uptake rates, so called fluxes through reactions, measure the production of one metabolite from another precursor metabolite. Cellular functions like growth and adaptation to environmental changes are investigated and simulated computationally.

As mentioned before, genome annotation is one of the first steps of comprehensive metabolic network investigation. The basic idea for investigating the evolution of metabolic networks is similar to evolutionary genome analysis. Similar reactions in various species are often catalyzed by orthologous enzymes. Knowledge about functional components (reactions) of a well investigated metabolic network is therefore transferable to metabolic networks of other species. Many components for metabolic network reconstruction are found by comparison and a backbone of the new metabolic network can be created.

It is a very complex task to trace back the evolution of the complete genome-wide metabolic network of a taxonomic group. A better approach for understanding the evolution of the whole network is to investigate the evolution of functional network modules first, which consist of reactions that are connected in a common pathway through the network. The evolutionary analysis of functionally connected reactions in this work is a first step towards understanding the development of genome-wide metabolic networks. Coupled reactions subsets are functionally connected reactions that can be detected with lower computational effort than extreme pathways (see below).

Coupled reactions are often embedded in pathways that lead to a specific product without existence of alternative reaction pathways. Such couplings are biologically meaningful, since they play a critical role for the functionality of the whole metabolic network. As already explained in Chapter 1, the fluxes of reactions which belong to coupled reaction subsets depend on each other. This dependency of reaction rates can be Boolean, which means zero flux of one reaction leads to no reaction rate in the other, or linear (Chapter 1.3). In order to define basic components for the production of specific metabolites, it is of interest to detect the first appearance of coupled reactions during the course of evolution. Since ancestral networks cannot be directly observed, comparisons between components of different metabolic networks lead the way for evolutionary network reconstruction. This comparative genomics based approach is facilitated by the fact that large databases of metabolic cellular reactions exist for specific species, and a large number of genomes are fully annotated.

Application of a comparative approach and concentration on a specific subset of reactions simplify the evolutionary analysis of metabolic networks, since whole-network reconstructions of only a few networks are required. It is convenient to compare the enzyme set of one well investigated metabolic network to that of other genomes. Because of their lower complexity, most well investigated whole-genome metabolic networks are from monocellular species - e.g., *Saccharomyces cerevisiae* for eukaryotes and *E. coli* K12 for prokaryotes (Reed *et al.*, 2003).

In this work, I focused on the **evolution of coupled reaction pairs** in the metabolic network of *E. coli*. Orthologous clusters of 14 *E. coli* strains were created by the previously developed phylogeny-based reference centered orthology assignment algorithm phyloCOP, which filters ambiguous paralogous relations (list of strains in Appendix_B in the attached CD). The phylogenetic history of coupled reaction pairs was then investigated, based on the appearance of orthologs. In particular, I was interested if orthologous genes associated with both, one or none of the coupled orthologous reactions existed at ancestral time points, relative moments in evolutionary time that correspond to common ancestors of each group of strains with the reference *E. coli* K12. In order to get a general overview about the development of coupled reactions over time, mean values from all coupled reaction pairs were calculated for each of the three cases at all ancestral time points.

3.1 Coupled subsets and extreme pathways

Genome-scale metabolic networks describe the cellular metabolism of a life form and are often illustrated as a graph, in which metabolites are depicted as nodes connected by vertices corresponding to reactions that convert one metabolite into the other (Figure 3.1 a). This graphical representation of the metabolic network is simplified, since regulatory molecules are often not included. However, it is possible to describe the information inside the graph mathematically as a stoichiometric matrix, an *in silico* representation of the metabolic network that can be directly used as input for computational simulation of all metabolite fluxes through a cell. General cell behavior and different biological scenarios, like adaptive processes caused by environmental changes or the effects of specific gene mutations can be simulated as well (Figure 3.1 b).

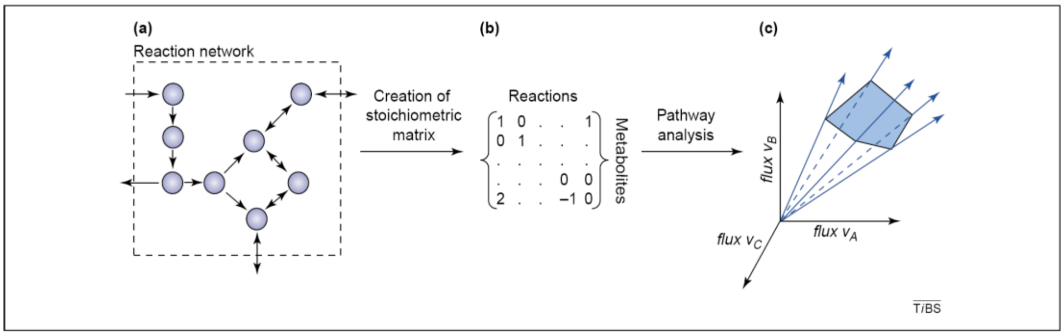


Figure 3.1: Mathematical characterization of metabolic network functions (Papin *et al.*, 2003).

(a) Graphical representation of a metabolic network. Nodes symbolize metabolites and arrows reversible or irreversible reactions. (b) Rows in the stoichiometric matrix correspond to metabolites and columns to reactions. Numbers refer to reaction fluxes from and to a metabolite and are based on reaction stoichiometry. 0 means that a reaction does not produce or consume a metabolite, while positive numbers refer to the production of a metabolite by the flux through a reaction and negative numbers to metabolite consumption. (c) High-dimensional flux space: axes correspond to a flux through reactions A, B and C. Network-based pathways like, e.g., extreme pathways curtail the set of possible pathways through the network in steady state condition. Environmental constraints are simulated by inequalities that curtail the possible values for specific fluxes through the network. The flux cone includes all possible flux distributions of the simulated network.

Linear algebra describes the metabolic network at steady state by the mass balance equation Sv = 0, where S is the stoichiometric matrix as described above and v the flux vector that corresponds to the fluxes through all reactions. FBA (Flux Balance Analysis) is a method that simulates different cellular aims based on linear programming. Assuming that the cells metabolism is in steady-state and applying thermodynamic rules, FBA calculates a possible solution for the flux distribution of all reactions that allows the cell to achieve a (biologically meaningful) aim, like maximal energy or biomass production, which is mathematically described as an objective function. Environmental constraints are simulated via inequalities that allow only specific ranges of flux values through some reactions. A valid set of metabolic fluxes at steady state for a specific objective function is retrieved from the analysis of the stoichiometric matrix. This set of fluxes can be understood as vectors in a more dimensional space whose axes are defined by the flux levels of the individual reactions (Figure 3.1 c).

The whole metabolic network can be characterized by the edges of the flux cone, the **extreme pathways**, which define the borders of possible fluxes through the network under steady state conditions. Extreme pathways form the linear basis of the stoichiometric matrix which means that linear combinations of them are sufficient to characterize all possible ways through the network.

Biologically, extreme pathways describe all anabolic and catabolic paths of a system, in which a substrate is assimilated and changed into a product that is either excreted or used for biomass production (Schilling *et al.*, 2000; Papin *et al.*, 2003). Extreme pathway analysis has been successfully applied to small-scale metabolic networks. However, the much higher number of reactions in a genome-scale network leads to an extremely high number of possible combinations and identified pathways. The identification of all extreme pathways in genome-scale networks is therefore computationally expensive and very slow.

The local structure in genome-scale metabolic networks can be described faster and easier via **Flux Coupling Analysis (FCA)**, which aims to detect **coupled reaction subsets** (Burgard *et al.*, 2004). Like extreme pathways, coupled reaction subsets are detected via linear programming based on the stoichiometric matrix under steady state conditions. Instead of maximizing biomass, the Flux Coupling Finder (FCF) algorithm - here referred to as LP-FCF (linear programming FCF) - determines the minimum and maximum flux ratios for every pair of non-blocked reactions (with positive flux). In each simulation one of the fluxes is fixed and the other one is either maximized or minimized. Ratios are calculated and compared. Based on their values it is decided whether a reaction pair is coupled or not. Three different coupling types are distinguished. Two reactions are **directionally coupled (** $v_1 \rightarrow v_2$ **)** if a zero flux for one reaction leads to a zero flux for the other but not vice versa, **partially coupled (** $v_1 \leftrightarrow v_2$ **)** if the condition for the directional coupling is true in both directions, and **fully coupled (** $v_1 \Leftrightarrow v_2$ **)** if the fluxes of both reactions depend linearly on each other, which means that a non-zero flux through one reaction implies a non-zero flux through the other. The FCF algorithm performs a series of linear programming steps. Not each pair of reactions has to be tested, since coupled relations are transitive. Thus, complete sets of coupled reactions are detected, which reduces the number of tested combinations. Combinatorial explosion is avoided and computation accelerated (Burgard *et al.*, 2004).

However, a lot of linear optimization problems must be solved by the FCF algorithm. Larhlimi and Bockmayr introduced a faster method for the detection of coupled reaction subsets at steady state conditions, in which coupling types are detected based on the **reaction reversibility type** (Larhlimi and Bockmayr, 2006). Like for the calculation of extreme pathways, the whole solution space in steady state is described by the flux cone, which is calculated from the stoichiometric matrix. The cone is based on a characteristic set of irreversible reactions. Exhaustive calculation of all extreme pathways is avoided. Since reversible reactions are not split into two irreversible ones, possible combinations are further reduced. In the following, this method is called MMB-FCF (minimal metabolic behavior FCF), to distinguish it from the LP-FCF algorithm. It is assumed that reaction couplings can only exist between similar reaction types. After removing blocked reactions, three types of reactions are distinguished: irreversible, pseudo-irreversible (reversible by network model definition but restricted in directionality by adjacent irreversible reactions), and fully reversible reactions.

Three coupled reaction types were classified by Larhlimi and Bockmayr, 2006 that correspond to the three types introduced by Burgard *et al.*, 2004. In this work, mathematical definitions of coupled reaction types are based on Larhlimi and Bockmayr, while corresponding names are taken from Burgard *et al.*:

$i \xrightarrow{=0} j:\ \mathrm{v}_i = 0 \Rightarrow v_j = 0$ **directionally coupled** **(1)**

$i \xleftrightarrow{=0} j:\ \mathrm{v}_i = 0 \Leftrightarrow v_j = 0$ **partially coupled** **(2)**

$i \sim \quad \mapsto \quad \mathbb{R} \quad {}_i$ **fully coupled** **(3)**

(1) Reactions *i* and *j* are directionally coupled if zero flux v_i through reaction *i* causes zero flux v_j through reaction j. (2) They are partially coupled, if the condition for the directional coupling is true in both directions. (3) Two reactions are fully coupled if conditions (1) and (2) are true, and either their reaction rates are similar (λ=1) or linear dependent. Because of its faster run time for the *E. coli* K12 model used here, compared to LP-FCF, MMB-FCF was used to detect the coupled reaction subsets that were used for the evolutionary analysis in this work (Larhlimi and Bockmayr, 2006).

3.2 Dataset and data preparation

MMB-FCF was applied to detect coupled reaction subsets in the *in silico* model of *E. coli* K12 MG1655 (Fritzemeier, 2010). A set of fully coupled reaction pairs was extracted from the detected fully coupled reaction subsets. PhyloCOP was used to generate 4126 paralogy-free orthologous clusters from 14 *E. coli* strains (compare Chapter 2). Pairs of orthologous clusters that correspond to the set of fully coupled reaction pairs were subsequently used for evolutionary analysis.

3.2.1 Selection of fully coupled reaction subsets

The expanded genome-scale *in silico* model of ***E. coli*** **K12 (*i*JR904)** that was created for constraint-based metabolic network modeling includes 931 reactions and 625 metabolites. This model is suitable for further comparative genome analyses, since GPR (Gene to Protein to Reaction) associations are included (Reed *et al.*, 2003). Coupled reaction subsets were determined by running the MMB-FCF algorithm implemented in *Gnu R* (http://www.r-project.org/) for *SyBil* (*Systems Biology Library*), a recent project of the Institute of Bioinformatics at the Heinrich-Heine-University Düsseldorf (Fritzemeier, 2010). The 142 identified fully coupled reaction sets were used for further analysis. A paralogy-free phyloCOP cluster was detected for each of the 322 proteins associated with the 142 coupled reaction sets.

Only reactions that are catalyzed by enzymes (and transport via transport proteins) are useful for comparative gene analysis. Thus, reactions that are not catalyzed by an enzyme were excluded. Based on the GPR associations of the *in silico E. coli* K12 model, a coupled reaction pair may be associated to two genes or more. Isoenzymes with similar function catalyze the same reaction independently from each other. Some enzymes consist of peptide subunits, which are encoded by different genes. In this case a gene that catalyzes one reaction is coupled to many genes simultaneously, which is similar for enzyme complexes. Only couplings with at most **three GPR associations** were taken into account. Two coupled reactions $(a \sim$ may correspond to one of the following gene compositions: **1)** $(A \sim$, **2)** $(A_1 \vee A_2 \sim$, where A_1 and A_2 encode isoenzymes, which both catalyze reaction a independently, or **3)** $(A_1 \wedge A_2 \sim$, where genes *A₁* and *A₂* encode parts of an enzyme complex.

Simple and **complex** coupling conditions were differentiated. Corresponding coupled reaction pairs are subsequently called simple and complex. A coupling in a genome is existant if both reactions can be catalyzed. Based on this assumption, it was decided whether two reactions in a genome are coupled or not. The coupling condition for reaction pairs with GPR association $(A \sim$ is **simple**. Reactions a and b are assumed to be coupled in a genome if orthologs were detected for both genes A and B. Coupling conditions for the two other GPR associations are **complex**. Two reactions with GPR association $(A_1 \vee A_2 \sim$ are assumed to be coupled if orthologs for B and A_1 **or** A_2 are detected, since one isoenzyme is sufficient to maintain functionality of reaction a. This is not the case if genes A_1 and A_2 correspond to parts of an enzyme complex. Therefore, coupling of reaction pairs with GPR association $(A_1 \wedge A_2 \sim$ is assumed to exist only if orthologs for both genes A_1 **and** A_2 as well as B were detected in a genome. An exception to this rule might occur if a mutation led to a transfer of the whole catalyzing function to one of the genes A_1 or A_2. In the following chapters, reactions associated with detected orthologous genes are called orthologous reactions and coupled reaction pairs in the *in silico E. coli* K12 model are called (coupled) reference reaction pairs. 219 coupled reaction pairs were extracted for further analysis from 130 coupled reaction subsets.

3.2.2 Selection of corresponding orthologous clusters

Paralogy free orthologous clusters of 14 *E. coli* strains and *E. coli* K12 as reference species were created with phyloCOP (α = 0.5) as explained in Chapter 2. In contrast to the comparative evaluation of the performance of phyloCOP, complete genomic protein sets were used including plasmid genes. The phyloCOP run order was newly determined, based on a previous run of phyloCliquesOP (reciprocal hit degree cut-off α = 1.0). Orthologous cliques can be assigned in random genome order, since all members of a clique are connected by mutual best sequence similarity hits. Cliques that include a gene from each genome were used to create a phylogenetic species tree with PhyML from concatenated multiple alignments (similar to the phylogeny estimation explained in Chapter 2.3.3.1). *Salmonella typhi* and *Salmonella typhimurium* LT2 served as outgroup to root the phylogenetic species tree (Figure 3.2). Distances of other genomes to *E. coli* K12 were obtained from the phylogenetic tree with the BioPerl CPAN module Bio::TreeIO (http://search.cpan.org/~cjfields/BioPerl-1.6.1/Bio/TreeIO.pm). The resulting phyloCOP run order is listed in Table 3.1.

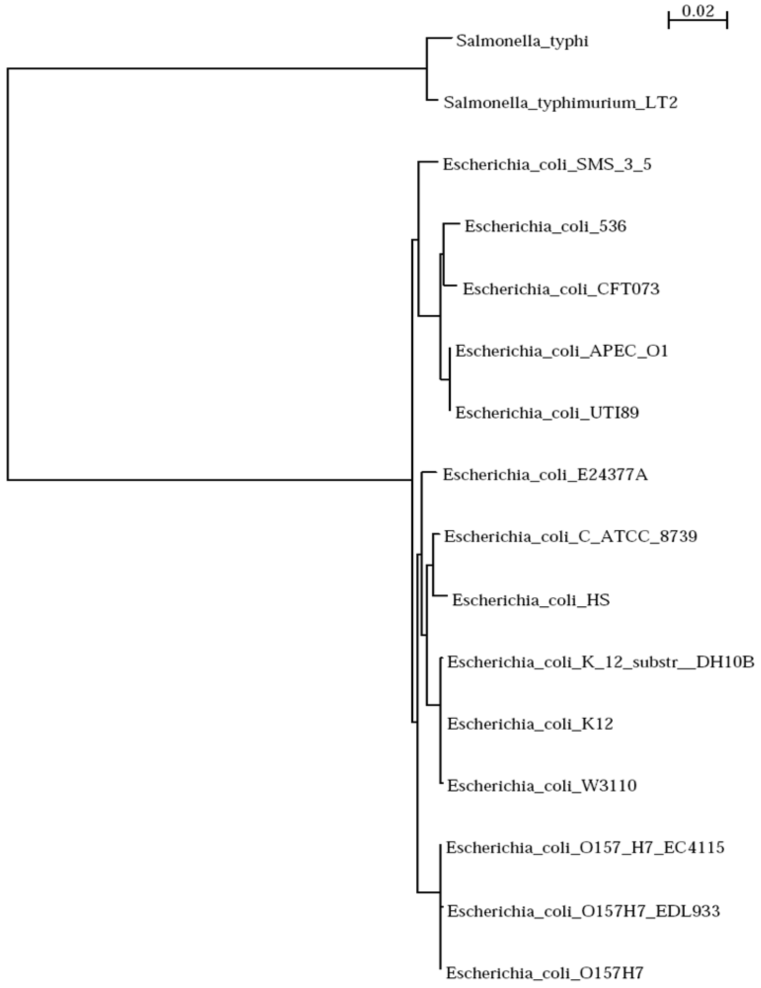


Figure 3.2: Phylogenetic species tree obtained from phyloCliques

The *Salmonella* outgroups marks the root of the tree. All branches have bootstrap support > 96 %.

Table 3.1: PhyloCOP genome run order in increasing phylogenetic distances to *E. coli* K12.

Escherichia_coli_K12
Escherichia_coli_W3110
Escherichia_coli_K_12_substr__DH10B
Escherichia_coli_C_ATCC_8739
Escherichia_coli_E24377A
Escherichia_coli_HS
Escherichia_coli_O157_H7_EC4115
Escherichia_coli_O157H7
Escherichia_coli_O157H7_EDL933
Escherichia_coli_SMS_3_5
Escherichia_coli_APEC_O1
Escherichia_coli_UTI89
Escherichia_coli_CFT073
Escherichia_coli_536

4126 paralogy-free phyloCOPs were detected and used for further analysis. Every gene in the *E. coli* K12 in *silico* model has a locus tag number, while phyloCOPs are named by the gi-number of the corresponding reference protein. Corresponding orthologous clusters were chosen for further analysis by translating the locus tag numbers of genes associated with coupled reference reaction pairs into the corresponding NCBI gi-numbers of the proteins. Since no corresponding gi-number was found for some locus tag numbers five pairs were excluded from further analysis. The remaining 214 reaction pairs consist of the following GPR types: 158 pairs $(A \sim$, 37 pairs $(A_1 \vee A_2 \sim$ and 19 pairs $(A_1 \wedge A_2 \sim$ (Tables in Appendix_C, attached CD).

3.3 Evolutionary analysis

I inspected the existence of coupled reaction pairs from an *in silico* model of *E. coli* K12 in 13 other *E. coli* strains, to get insights into the metabolic network evolution of prokaryotes. A relative evolutionary time scale was introduced that allows tracing back the existence of coupled reference reactions in all common ancestors of *E. coli* K12 and the other genomes. Results for all coupled reactions at each relative ancestral time point were united and visualized in a graph.

3.3.1 Determination of relative ancestral times

Ancestral points were classified relative to the reference genome *E. coli* K12. The phylogenetic species tree was reconstructed from complete phyloCOPs by multiple alignment and Maximum Likelihood tree reconstruction similar to the phylogeny estimation from orthologous cliques (Chapter 2.3.3.1). The rooting position for the phyloCOP species tree could be adopted from the phyloCliquesOP species tree because of their equal topology (Figures 3.2 and 3.3). Common ancestors between the reference genome and all other genomes that refer to internal nodes that connect a group of genomes with the reference, mark relative ancestral times starting with A for the reference itself until G for the root (Figure 3.4). Genomes connected to *E. coli* K12 by the same common ancestor are assigned to the corresponding relative time point. Some ancestral points are related to only one genome, while others are related to many. In order to evaluate the degree of coupling of a reaction pair at a time point, results are normalized by the number of genomes that share the same ancestor with the reference (Table 3.2).

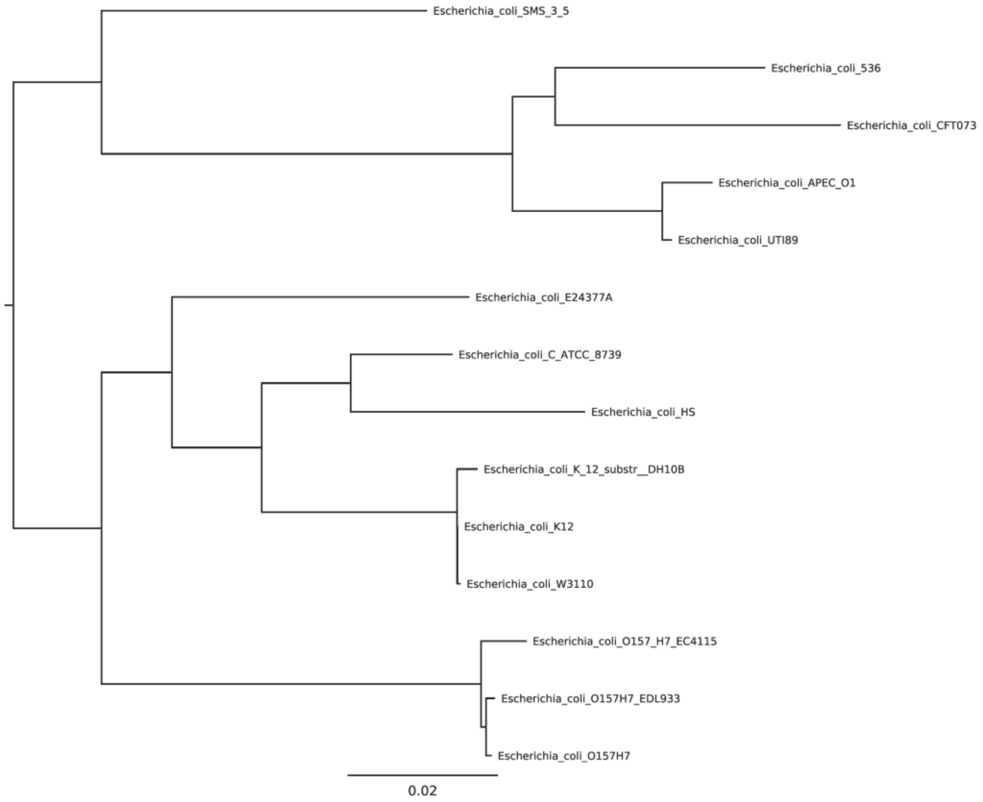


Figure 3.3: Phylogenetic tree reconstructed by PhyML.

The position of the root arrow was adopted from the phylogenetic tree in Figure 3.2.

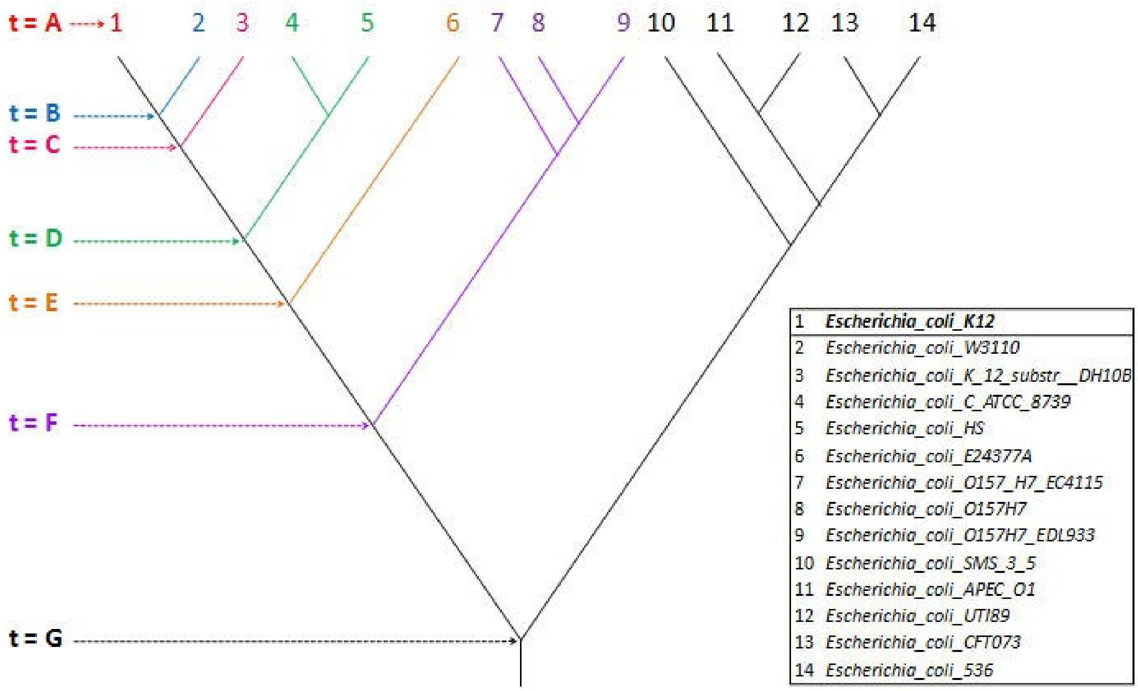

t = A
1
2
3
4
5
6
7
8
9
10
11
12
13
14
t = B
t = C
t = D
t = E
t = F
t = G
1 Escherichia_coli_K12
2 Escherichia_coli_W3110
3 Escherichia_coli_K_12_substr__DH10B
4 Escherichia_coli_C_ATCC_8739
5 Escherichia_coli_HS
6 Escherichia_coli_E24377A
7 Escherichia_coli_O157_H7_EC4115
8 Escherichia_coli_O157H7
9 Escherichia_coli_O157H7_EDL933
10 Escherichia_coli_SMS_3_5
11 Escherichia_coli_APEC_O1
12 Escherichia_coli_UTI89
13 Escherichia_coli_CFT073
14 Escherichia_coli_536

Figure 3.4: Non-scaled representation of the phylogenetic tree rooted by *Salmonella*. Genomes belong to similar colored relative ancestral time points (A - G).

Table 3.2: Genomes and corresponding ancestral time points

Genomes that share the same common ancestor with the reference genome *E. coli* K12 were assigned to the same ancestral time point (compare Figure 3.4). The number of assigned genomes differs for each time point. To evaluate the degree of ancestral coupling at a time point, the number of detected couplings is normalized based on the number of genomes that share the same ancestor with the reference.

Genome	Ancestral time point	Norm. factor w
Escherichia_coli_K12	A	1
Escherichia_coli_W3110	B	1
Escherichia_coli_K_12_substr__DH10B	C	1
Escherichia_coli_C_ATCC_8739	D	½
Escherichia_coli_HS	D	½
Escherichia_coli_E24377A	E	1
Escherichia_coli_O157_H7_EC4115	F	1/3
Escherichia_coli_O157H7	F	1/3
Escherichia_coli_O157H7_EDL933	F	1/3
Escherichia_coli_SMS_3_5	G	1/5
Escherichia_coli_APEC_O1	G	1/5
Escherichia_coli_UTI89	G	1/5
Escherichia_coli_CFT073	G	1/5
Escherichia_coli_536	G	1/5

3.3.2 Detection of ancestral coupled reaction pairs

In order to trace back the development of the coupled reference reaction pairs, their existence was determined at ancestral time points. At first, each coupled reaction pair was examined separately. Orthologous genes associated to a coupled reaction pair were scanned in each species. Based on this, it was decided whether **both, one** or **none** of the reactions appear in the metabolic network of the species. It is considered that the coupling of a reaction pair got lost in a genome if a necessary orthologous gene for one of the reactions was not detected. On the other hand, a coupling is assumed to exist in a genome if all necessary orthologous genes for both reactions were detected. The lack of necessary orthologous genes for both reactions hints that they got lost simultaneously. In this case the coupling may have persisted until the gene loss event. Since more than one genome may be assigned to an ancestral time point, results for each appearance (both, one and none) must be normalized by the corresponding factor w (Table 3.2). Subsequently, mean values of all normalized fractions of both, one and none detected orthologous reactions $\overline{m}$ were calculated from all coupled reference reaction pairs at each ancestral time point respectively.

$$w = 1/N_g \tag{1}$$

$$n = xw \tag{2}$$

$$\overline{m} = \sum_{i=1}^{N_c} n_i \Big/ N_c \tag{3}$$

(1) The normalization factor w depends on the number of genomes N_g that are associated with a relative ancestral time. (2) Numbers of both, one and none orthologous reactions (x) detected in genomes that belong to a relative ancestral time point are normalized by the normalization factor. (3) The mean over the results for all coupled reference reaction pairs was calculated. Corresponding mean values $\overline{m}$ for all coupled reference reaction pairs were computed for each relative ancestral time, by summing all normalized values n of both, one and none detected orthologous reactions respectively, divided by the total number of coupled reference reaction pairs (N_c).

In addition it was assumed that, if both reactions appear in at least one associated genome, the coupling was present in the common ancestor (ancestral time point). This simplified assumption was made to determine the **last common ancestor** between the reference and the other genomes, in which the reactions were **not yet coupled**.

3.4 Results

The averaged normalized results for the total number of 214 coupled reference reaction pairs are listed in Table 3.3. Fractions of all 214 coupled reference reaction pairs for which both, one or none orthologous reactions were detected are displayed in Figures 3.5 and 3.6. Most coupled reference reaction pairs appear coupled at all ancestral time points. As expected, the fraction of reference reaction pairs, for which no orthologous reaction was found, increases with higher phylogenetic distance to the reference genome. Interestingly, time point C that is closely related to the reference genome is an exception to this trend. About 10 % of the reference couplings do not exist or both reactions are deleted in *E. coli* DH10B, the only genome associated to C. Figure 3.6 shows that this is mostly due to the absence of one reaction of a pair only (68 % of the fractions one and none together). In addition, a closer look at Table 3.3 reveals that reference reaction pairs with GPR association $(A_1 \vee A_2 \sim$ have the highest fraction of uncoupled reaction pairs at time point C, where always one of the reactions was detected. In general, fractions of one detected orthologous reaction were higher than none, except at time point G.

Table 3.3: Normalized fractions of orthologous coupled reaction pairs

Normalized fractions of both, one and none reactions that are orthologous to all coupled reference reaction pairs at each ancestral time point were calculated for the three different GPR association types respectively as well as for all 214 clusters (total). The highest number of uncoupled reference reaction pairs was detected at the ancestral time point C, which corresponds to the substrain *E. coli* K12 DH10B (marked pink).

$(A \sim B)$	**Both**	**One**	**None**
A	1	0	0
B	1	0	0
C	0.911	**0.050**	0.038
D	0.981	0.016	0.003
E	0.994	0	0.006
F	0.962	0.025	0.013
G	0.929	0.030	0.041
$(A_1 \vee A_2 \sim$	**Both**	**One**	**None**
A	1	0	0
B	1	0	0
C	0.865	0.135	0
D	0.986	0.014	0
E	1	0	0
F	1	0	0
G	0.995	0.005	0
$(A_1 \wedge A_2 \sim$	**Both**	**One**	**None**
A	1	0	0
B	1	0	0
C	1	0	0
D	1	0	0
E	0.895	0.105	0
F	0.895	0.053	0.053
G	0.958	0.042	0
Total	**Both**	**One**	**None**
A	1	0	0
B	1	0	0
C	0.911	**0.060**	0.028
D	0.984	0.014	0.002
E	0.986	0.009	0.005
F	0.963	0.023	0.014
G	0.943	0.027	0.030

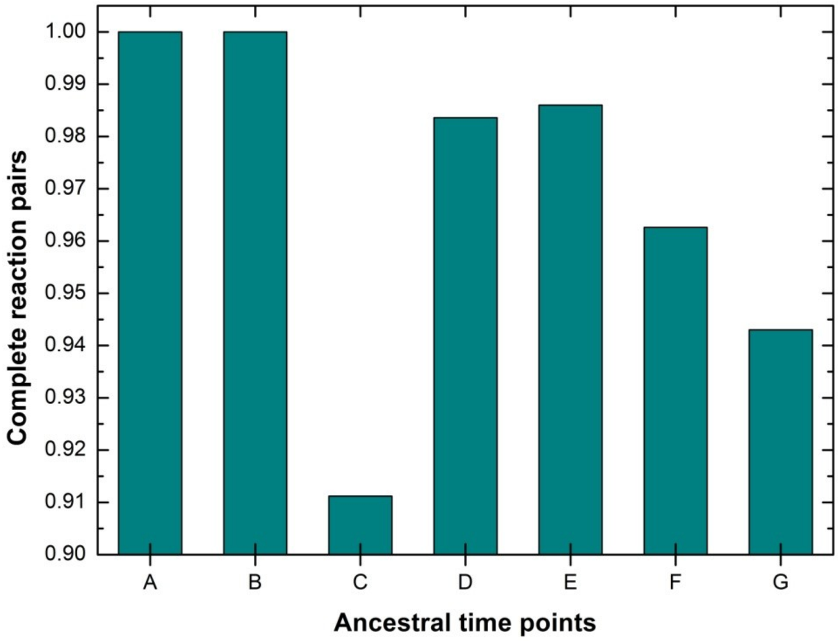


Figure 3.5: Fractions of complete orthologous coupled reaction pairs.

This graph shows the fraction of coupled reference reaction pairs for that all orthologous genes that are needed for the function of both reactions were detected.

Time point A refers to the reference genome *E. coli* K12, from which coupled reaction pairs were obtained. About 9 % of the coupled reference reaction pairs are not coupled or deleted at time point C, which is associated with *E. coli* K12 DH10B.

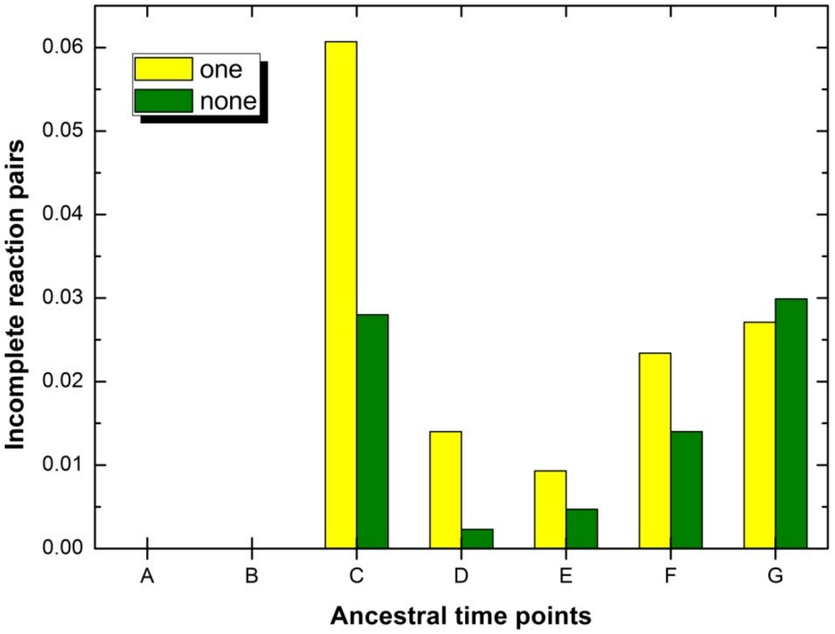


Figure 3.6: Fractions of coupled reference reaction pairs for which not all orthologs were detected.

The overall fraction of incomplete reaction pairs is low. If only one orthologous reaction was detected, the coupling is assumed not to be existent in a genome. In B (*E. coli* W3110) all reaction pairs remain coupled, while the highest amount of uncoupled reactions was found at time point C (*E. coli* K12 DH10B).

In the following, I refer to the lack of orthologous reference genes in a genome as gene loss. If two reactions are coupled, the probabilities that one or both corresponding genes are missing should differ from the gene loss probabilities of uncoupled reactions.

Gene loss probabilities β for all combinations of two genes that are associated with GPR-type $(A\sim$ coupled reaction pairs were calculated from the fraction of missing genes p at each ancestral time point:

$$\beta_o = 2p(1-p)$$ **one gene is missing** **(1)**

$$\beta_n = p^2$$ **both genes are missing** **(2)**

Fractions of missing genes p were derived by dividing the number of missing orthologs by the total number of possible orthologous genes at each ancestral time point respectively. Fractions p are displayed in Figure 3.7 together with the fractions of $(A\sim$ coupled reaction pairs for which only one or none orthologous gene was detected.

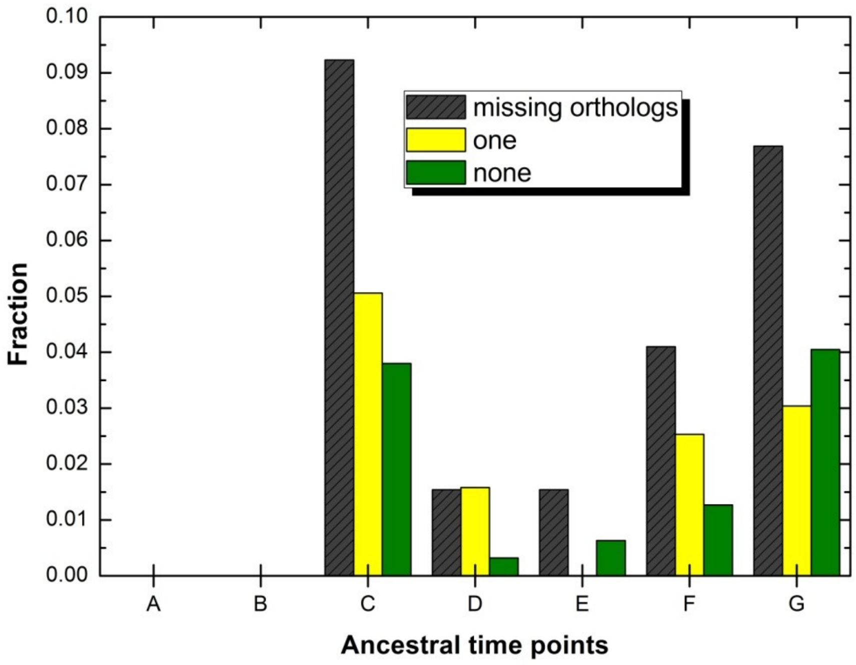


Figure 3.7: Fractions of GPR type $(A\sim$ incomplete coupled reaction pairs and missing orthologs p.

Most missing orthologs were detected at time point C (*E. coli* DH10B).

The fraction of coupled reactions for which only one orthologous gene was detected is much lower and the fraction of coupled reactions with two missing genes are significantly higher than the predicted probabilities of lost genes for independent pairs of genes (Figures 3.8 and 3.9). Gene loss is obviously correlated with reaction coupling at each ancestral time point. Genes that catalyze coupled reactions are more frequently lost together than independent genes.

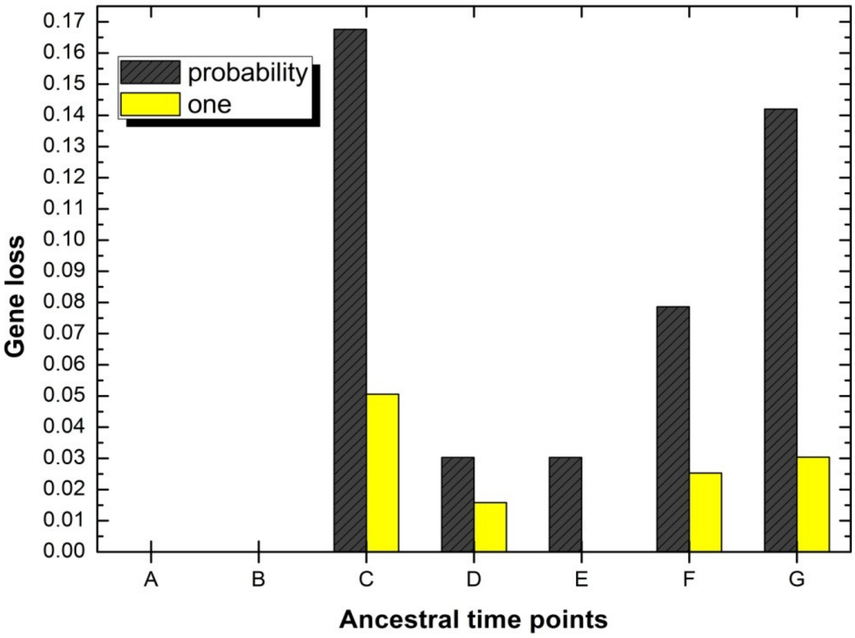


Figure 3.8: Probabilities of gene loss β_o and corresponding fractions of GPR type $(A\sim$ coupled reaction pairs for which only one orthologous gene was detected.

The fractions of coupled reactions with one missing gene are much lower than the predicted probabilities of lost genes for independent pairs of genes.

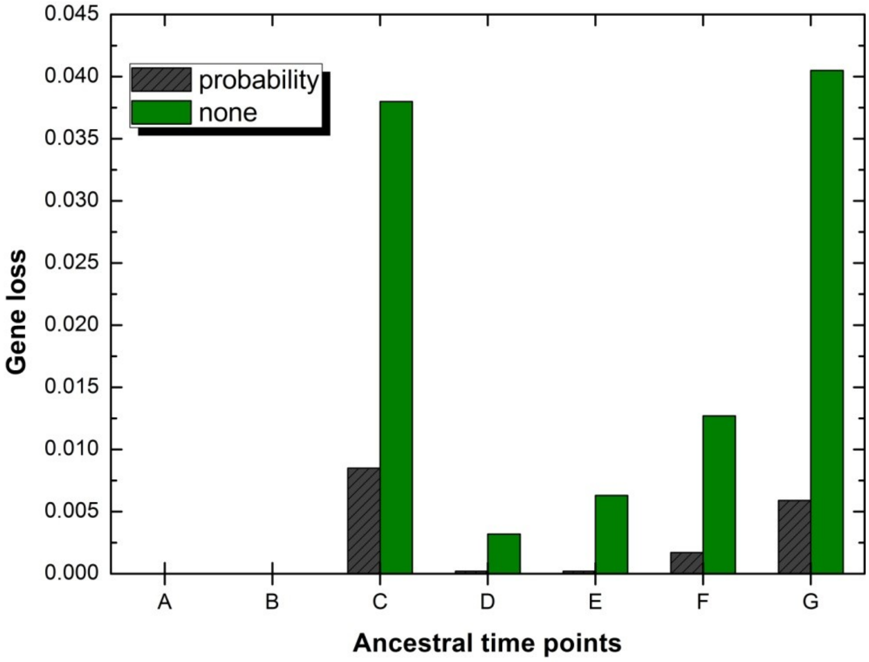


Figure 3.9: Probabilities of gene loss β_n and corresponding fractions of GPR type $(A\sim$ coupled reaction pairs for which no orthologous genes were detected.

The fractions of coupled reactions with two missing genes are significantly higher than the predicted probabilities of lost genes for independent pairs of genes.

In a simplified assumption a coupling is present at an ancestral time point if it exists in one associated genome. Besides, the coupling is not considered cancelled if all genes for both reaction pairs are deleted. Results for all 214 examined coupled reaction pairs mostly reflect the former described fractional results, especially for time points B and C (compare Table 3.3 with 3.4 and Figures 3.5 and 3.6 with 3.10). In contrast, all reference reaction pairs appear coupled at D and G, which means that for each of the coupled reference reaction pairs at least one associated genome includes all required orthologous genes.

Table 3.4: Numbers of detected coupled, deleted and not coupled reaction pairs at each ancestral time point

All 214 coupled reaction pairs were included in this analysis. Results based on the simplified assumption that two reactions are coupled in a common ancestor if orthologs of all associated genes are detected in at least one genome that belongs to an ancestral time point. At time point F only one reaction was detected for 4 coupled reaction pairs in at least one of three genomes, while 3 of them were completely deleted in at least one genome (only one reaction pair is definitely not coupled). In correspondence to the more detailed fractional analysis, a significant number of reference reaction pairs appear not coupled at time point C.

Times	Coupled	Deleted	Not coupled
A	214	0	0
B	214	0	0
C	195	6	13
D	214	0	0
E	211	1	2
F	210	3	1
G	214	0	0

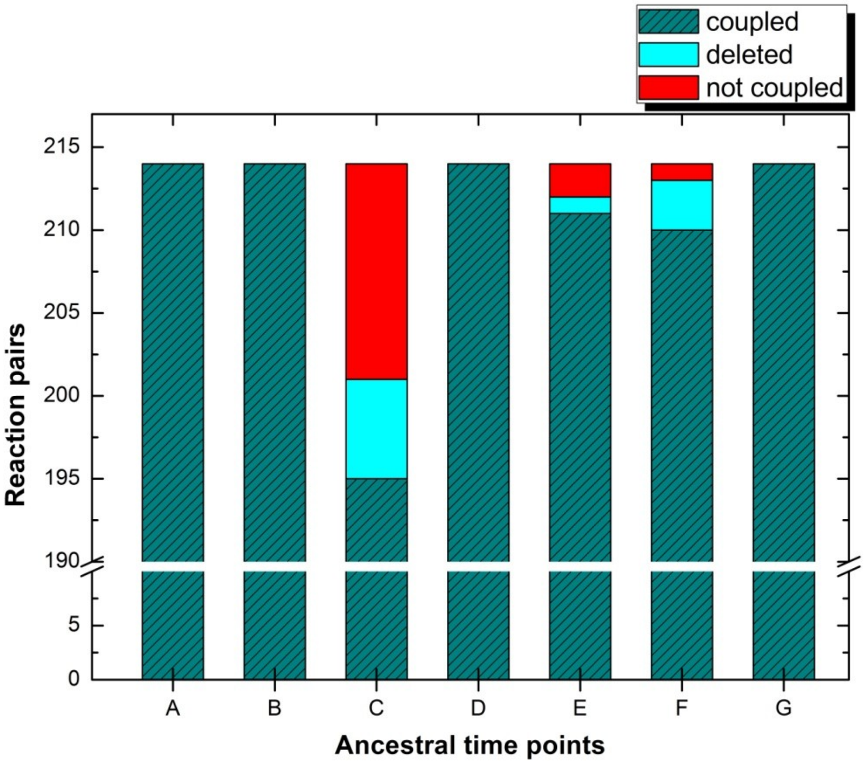


Figure 3.10: Comparison between the fractions of coupled, deleted and not coupled reference reaction pairs.

The histogram shows total numbers of coupled, deleted and not coupled reaction pairs, instead of fractions. The coupling is present if all orthologous genes for both reactions are found in at least one genome that belongs to an ancestral time point. Most reaction pairs remain coupled. Time point C shows the highest number of uncoupled pairs.

The coupling of all investigated reaction pairs seems to be initialized at a former ancestral time point than G, which could not be traced back by the closely related genomes investigated in this study. The reason for the higher number of uncoupled reactions in C is most probably gene loss in *E. coli* K12 DH10B (see next Subchapter).

3.5 Discussion

The evolutionary development of functional components in prokaryotic metabolic networks was investigated based on comparative genomics. Coupled reaction subsets were detected by applying the MMB-FCF algorithm on an expanded genome-scale *in silico* model of *E. coli* K12 MG1655 (*i*JR904). In order to trace back the development of selected coupled MG1655 reaction pairs over evolutionary time, the appearance of previously obtained corresponding orthologous genes associated with orthologous reaction pairs in 13 other *E.*

coli strains was checked. Genomes were assigned to relative ancestral time points that correspond to their common ancestors with the *E. coli* K12 MG1655 reference genome.

A general overview about the distribution of orthologous reactions, given by the percentage of orthologous reactions that correspond to both, one and none reaction of all coupled reaction pairs, shows that the number of uncoupled reference reaction pairs is low at all relative ancestral times. This was expected, since closely related genomes – strains of the same species – were investigated. The number of cases with one detected orthologous reaction was always higher than the number of cases in which no orthologous reaction was found except at time point G. The percentage of missing orthologs at each ancestral time point reflects this trend (Figure 3.8). However, each examined reference reaction pair remains coupled in at least one of the five genomes assigned to time point G, which is evidence that the last common ancestors of all coupled MG1655 reaction pairs are more ancient than G. It can be assumed that a repetition of the presented analysis with multiple genomes that are more distantly related to MG1655 may result in detecting the last common ancestry of several coupled reaction pairs.

It is likely that xenologs are included in the phyloCOP clusters, because I did neither differentiate xenologous proteins from orthologs nor exclude them during the assignment. If the ancestral time point of the initial appearance of a reaction coupling is detected, it is theoretically possible to detect the occurance of laterally transferred genes associated with coupled reaction pairs at more ancient ancestral time points. The detection of lateral gene transfer would then also require a data set of more distanty related genomes than used in this work.

The probability of a combined loss of two genes that are associated with a coupled reaction pair is higher than for two independent genes. It is also less likely that only one reaction partner is deleted. The probability of gene deletion of only one of two independent genes corresponds to the percentage of simultaneous gene deletion of both coupled genes. This shows that genes associated with coupled reaction pairs are often deleted together as a unit. The same can be assumed for laterally transferred genes.

The highest percentage of uncoupled reaction pairs was found at time point C, the second closest ancestral time point to the reference. The only genome associated with time point C is *E. coli* DH10B, whose phylogenetic relatedness to the reference *E. coli* K12 MG1655 is similar to the phylogenetically closest genome *E. coli* W3110 at time point B (Figure 3.3). In contrast to C, all coupled reference reaction sets were detected in B.

The progenitor of the substrains MG1655, W3110 and DH10B was the ancestral *E. coli* K12 wild-type. There are remarkable differences in the construction procedure of W3110 and DH10B. While W3110 and MG1655 directly diverged from wild-type K12 and only slightly differ due to the stronger galactose-fermenting property of W3110, DH10B was constructed by a series of recombination steps. Therefore MG1655 and W3110 are closer to the ancestral K12 wild-type. During the classical genetic strain construction of DH10B several unspecified DNA fragments were transferred and deleted in addition to targeted modifications (Hayashi *et al.*, 2006; Durfee *et al.*, 2008). In accordance to this, the highest percentage of missing orthologous genes was found at time point C.

The results of the evolutionary analysis in this work reflect both, the reported extensive sequence conservation similarity between MG1655 and W3110, and the effect of three large-scale deletions in DH10B: 1) the *lac* operon (Δ*lacX74*), 2) Δ(*mrr-hsdRMS-mcrBC*) that encodes six restriction enzymes, and 3) the *leuLABCD* operon Δ(*ara leu*)7697 (needs leucin to grow). All coupled reference reaction pairs remained coupled in W3110 because of its strong similarity with MG1655. 7 reaction pairs appear uncoupled at time point C because corresponding genes are affected by one of the deleted alleles (Table 3.5). In addition, 6 *E. coli* K12 coupled reaction pairs were lost because of these gene deletions.

Table 3.5: Gene composition of uncoupled and deleted reaction pairs in DH10B

Genes that belong to alleles deleted by targeted recombination in DH10B are marked **bold**. Deleted alleles are the reason for the abolishment of 13 coupled reaction pairs in DH10B. Among the 13 pairs, the number of orthologous reactions found in DH10B matches to the number of deleted orthologous genes (Durfee *et al.*, 2008).

Coupled reference reaction pairs	Orthologous reaction	Deleted allele
b0331~b1276∨b0118	One	Δ*lacX74*
b0333~b1276∨b0118	One	Δ*lacX74*
b0334~b1276∨b0118	One	Δ*lacX74*
b4322~b3093∨b3909	One	Δ(*mrr-hsdRMS-mcrBC*)
b4323~b3093∨b3909	One	Δ(*mrr-hsdRMS-mcrBC*)
b4323~b3092	One	Δ(*mrr-hsdRMS-mcrBC*)
b4323~b3093	One	Δ(*mrr-hsdRMS-mcrBC*)
b0331~b0333	None	Δ*lacX74*
b0334~b0333	None	Δ*lacX74*
b0334~b0350	None	Δ*lacX74*
b0352~b0350	None	Δ*lacX74*
b4322~b4323	None	Δ(*mrr-hsdRMS-mcrBC*)
b0062~b0063	None	Δ(*ara leu*)7697

DH10B is not able to use lactose as a nutrient (Hayashi *et al.*, 2006; Durfee *et al.*, 2008). 7 coupled reaction pairs that belong to metabolic pathways involved in the transport and conversion of lactose do not exist or are no longer coupled because of the reported gene deletion. In addition DH10B lacks a set of restriction enzymes. Thus, reactions that belong to the corresponding pathways of DNA catabolism are no longer catalyzed and coupling to other reactions abolished. Orthologous genes that correspond to a deleted allele in DH10B were not detected in this analysis.

As mentioned before, phyloCOP does not differentiate between xenologs and orthologs. Laterally transferred genes are therefore included in the phyloCOP clusters. However, correct function prediction is not violated by xenologs, because their functions are often conserved. Xenologs should not be excluded from the evolutionary analysis, since LGT is a main evolutionary process in prokaryotes and reaction coupling may be horizontally transferred as well (Dagan and Martin, 2009). For the scientific task in this work, which is the evolutionary analysis of metabolic coupling, it is not necessary to differentiate between xenology and orthology during the clustering procedure, which is computationally expensive (see Chapter 2). This can be done afterwards by comparisons between gene trees and phylogenetic species tree.

In general, results reflect some gene deletions of catalyzing enzymes in anabolic and catabolic pathways of *E. coli* DH10B, which is another evidence for the reliability of orthologous clusters created by phyloCOP. However, inclusion of more distantly related genomes is required for a higher resolution of the development of coupled reaction pairs over time. Information about the first occurrence of coupled reaction subsets and lateral gene transfer can subsequently be used for further evolutionary studies of complete coupled reaction subsets – functional pathways through the cell, which is one of the first steps towards comprehending the evolution of prokaryotic metabolic networks.

Chapter 4 Conclusion and outlook

Systems biology, a new interdisciplinary scientific field of biology, alternates molecular biological analyses with computational simulation steps for integrative research. Experimental, physiological and genomic data serve as input for mathematical models that explain and simulate cellular behavior. The results of the simulations give new ideas for new experimental research.

Over the last decades numerous molecular biological high-throughput sequencing techniques have been developed, which led to an immense increase of sequenced genomes. On the other hand, comprehensive knowledge about cellular functions is only available for a few species. Therefore, comparative genomics is frequently used for protein function prediction of newly sequenced genomes (Chapter 1). Although various sequence similarity based methods exist for orthology assignment, there is no gold-standard so far. A reason for this is the different frequency of evolutionary mechanisms like gene duplication and lateral gene transfer in prokaryotes and eukaryotes. Furthermore, different research aims require either the exclusion of paralogy or xenology, or a complete resolution of orthology, paralogy and xenology. Therefore, the existing orthology assignment methods are designed to serve specific research aims. Most previously designed algorithms that aim to meet the requirements of various research aims and aim to resolve all homologous relations among diverse taxonomic clades are computationally expensive. The development of an orthology assignment algorithm with user-defined parameters that make it adjustable to various research aims is maybe the best solution for a better cost benefit trade-off (Chapter 2). Metabolic networks can also be analyzed by comparative genomics, since many proteins are functional components of the cellular metabolism. It is possible to form the backbone of the metabolic network of a species based on the reactions associated with functionally annotated proteins. Functional relations in a well investigated network can be transferred to the newly investigated networks and used for evolutionary studies (Chapter 3).

Evolutionary analysis deals with the investigation of the development of life over time. Ancestral states of biological systems, like metabolic networks, cannot directly be observed, but they can be reconstructed via comparisons between phylogenetically related organisms.

This work deals with comparative genomics applied to the analysis of prokaryotic metabolic network evolution. Orthologous genes are genes in different genomes with common phylogenetic origin. Since orthologs often share similar functions, their identification is used for function prediction in genomes that are not well investigated. Functional relations in the network of a reference species are then transferred to metabolic networks of other analyzed species.

A new **orthology assignment method** - **phyloCOP** - was developed in this work. The generated orthologous clusters served as input-data for further analyses, in particular protein function prediction, phylogeny reconstruction and subsequent **evolutionary analysis of metabolic coupling**. The development of phyloCOP was inspired by other sequence similarity-based clustering algorithms, like, e.g. COG, that exploit **transitive** orthologous relations between species. PhyloCOP is a **greedy phylogeny- and reference-based** algorithm that combines the better assignment accuracy of other phylogeny-based orthology assignment algorithms, like e.g. SYNERGY, with the higher computational speed and lower complexity of simple reference-based methods, like starCOG. PhyloCOP provides a set of **user-defined parameters** that make it adjustable to various research aims. The user chooses the reference genome, an e-value cut-off and the degree of transitivity to adjust the algorithm to the phylogenetic distance of the investigated genomes.

PhyloCOP’s performance was tested on a smaller dataset of closely related species and a larger dataset including more distantly related species. PhyloCOPs created from both datasets were compared to alternative orthology assignment methods. Comparisons based on the smaller dataset show high similarity between all methods, where phyloCOP together with reference-based starCOG K12 is the most inclusive method. In addition, high integrity of orthologous clusters that correspond to universal COGs was confirmed for the larger dataset of 539 species. Applied to this phylogenetically diverse dataset, simple reference-based starCOG K12 clustered a significantly higher number of genes than MCL or phyloCOP, which includes most probably a high number of false positive assignments. Unlike for the dataset of closely related genomes, the inclusiveness of phyloCOP and starCOG is not similar for distantly related genomes. This indicates that simple reference-based clustering is not sufficient for distantly related species. All three tested algorithms performed well in the assignment of orthologous universally distributed proteins. Clusters that correspond to universally distributed COGs included proteins from most of the tested genomes. Only for a few distant bacterial and archaeal genomes no orthologs were found. The good performance of starCOG K12 is surprising. Sequence similarity seems to be very high between orthologous universally distributed proteins even between relatively distantly related bacterial species. However, it has to be tested, if starCOG K12, phyloCOP and MCL clustered the same proteins. Since starCOG does not exclude paralogs false positive assignments are more likely. MCL, on the other hand, assigns a large number of proteins to clusters that include paralogs. The number of paralogous MCL clusters is significantly higher than the number of paralogous clusters excluded by phyloCOP. In order to test if MCL clustered paralogs correctly, paralogous proteins in the MCL clusters have to be observed. If they are e.g. isoenzymes in *E. coli* K12, correct paralogy assignment can be assumed. In this case the parameters of phyloCOP's paralogy filter must be changed in a way that more paralogous clusters are detected and filtered. For example, instead of assigning a protein to two clusters if it fits to each of them alone (based on the reciprocal hit degree > α, with each of the two clusters), it could already be assigned to both clusters if the reciprocal hit degree of the protein with the union of both clusters is > α (Chapter 2).

For the **evolutionary analysis of metabolic coupling**, protein functions of the well investigated metabolic network of *E. coli* K12 MG1665 were assigned to their orthologous counterparts in other *E. coli* strains. Relative ancestral time points were determined based on phylogenetic genome relations that were reconstructed from the phyloCOPs. Occurrence of orthologous proteins that correspond to selected coupled reaction pairs was checked at each ancestral time point (ancestor between the reference and the other genomes) in order to trace back the evolution of coupled reaction pairs over time.

No last common ancestor was detected at which any of the investigated coupled reaction pairs appeared for the first time, because all investigated genomes are closely related *E. coli* strains. At time point B all reference reaction pairs remained coupled, which reflects the high genome similarity between the reference genome and *E. coli* W3110. In contrast, a significant number of coupled reference reaction pairs are either deleted or uncoupled in the second closest ancestor to *E. coli* K12, *E. coli* DH10B (ancestral time point C). Gene deletion seems to be the reason for the non-existence of some couplings at C since, 1) many gene deletions are confirmed in DH10B and, 2) all corresponding orthologous coupled reaction pairs exist at ancestral time points that are more distant to the reference genome (Chapter 3). Since xenologs are not excluded from the phyloCOP clusters, it is possible to detect them after the clustering procedure. However, xenologs of genes that are deleted in *E. coli* DH10B were not detected.

As mentioned in Chapter 3, a differentiation of xenology from orthology via gene tree reconstructions during the orthology assignment is computational intensive and orthologs are only syntenic in closely related genomes. I wanted to create a method that assigns orthologs to clusters that can be easily processed for further functional or evolutionary analysis. Correct function prediction is not violated by xenologs, because their functions are often conserved, since LGT provides selective advantages. It is also interesting and important to trace back the lateral inheritage of coupled reaction pairs because LGT is a main evolutionary driving force in prokaryotes. LGT can be detected after the homology assignment by gene tree comparisons with the assumed species phylogeny. For all these reasons it is useful not to exclude xenologs during the clustering procedure.

Comparisons between gene loss probabilities for two independent gene pairs and the frequencies of deleted coupled reaction pairs reveal a correlation between reaction coupling and combined gene loss. Two genes that are associated with a coupled reaction pair are more often lost together and less often lost alone compared to independent genes. It is likely that two coupled genes are also more often laterally transferred together than alone. It can be assumed that reaction couplings last for a long time in metabolic network evolution. It is also interesting to investigate, why a coupled reaction pair is not conserved in a taxonomic clade. Reasons might be functional changes due to environmental selection forces. Many proteins have more than one function or have the potential to develop new functions. One of the proteins might have another (hidden) function that is useful to survive in a specific environment. Since we do not focus on the reactions themselves but on the proteins, we cannot see if the function associated with the coupled reaction pair got lost. If the protein that remains has another function and one protein that is associated with the other reaction of the coupled reaction pair got lost, it is likely that both coupled reactions got lost together. Experimental analysis might help to determine multiple functions and interactions of proteins.

Last common ancestors in which reference reaction pairs are not yet coupled plus LGT events between more distantly related species can most probably be detected by repeating the evolutionary analysis in this work with genome sets that include more distantly related prokaryotic taxonomic clades. Based on this, further evolutionary studies may focus on the development of complete reaction pathways through the network, which is an important step towards understanding the evolution of cell anabolism and catabolism.

References:

Altschul, S. F., Gish W., Miller W., Myers E. W. and Lipman D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**:403-410.

Bachmann, P. (1894). *Die Analytische Zahlentheorie. Zahlentheorie*. pt. 2 Leipzig: B. G. Teubner.

Bapteste, E., O'Malley, M. A., Beiko, R. G., Ereshefsky, M., Gogarten, J. P., Franklin-Hall, L., Lapointe, F. J., Dupré, J., Dagan, T., Boucher, Y. and Martin, W. (2009). Prokaryotic evolution and the tree of life are two different things. *Biol Direct* **4**:34-53.

Berg, J., Tymoczko J. and Stryer, L. (2002). *Biochemistry.* W.H. Freeman and Company. 5th edition.

Burgard, A. P., Nikolaev, E.V., Schilling, C. H. and Maranas, C. D. (2004). Flux coupling analysis of genome-scale metabolic network reconstructions. Genome Res **14**:301-312.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**:540-552.

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B., and Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**:1283-1287.

Covert M. W., Schilling, C. H., Famili, I., Edwards, J. S., Goryanin, I. I., Selkov, E. and Palsson, B. O. (2001). Metabolic modeling of microbial strains in silico. *Trends Biochem Sci* **26**:179-186.

Dagan, T. and Martin, W. (2007). Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci USA* **104**:870-875.

Dagan, T. and Martin, W. (2009). Getting a better picture of microbial evolution en route to a network of genomes. *Philos Trans R Soc Lond B Biol Sci* **364**:2187-2196.

Davison, J. (1999). Genetic exchange between bacteria in the environment. *Plasmid* **42**:73-91.

van Dongen, S. (2000) *Graph clustering by flow simulation.* Dissertation, Universiteit Utrecht.

Durfee, T., Nelson, R., Baldwin, S., Plunkett, G. 3d, Burland, V., Mau, B., Petrosino, J. F., Qin, X., Muzny, D. M., Ayele, M., Gibbs, R. A., Csörgo B., Pósfai, G., Weinstock, G. M. and Blattner, F. R. (2008). The complete genome sequence of Escherichia coli DH10B: insights into the biology of a laboratory workhorse. *J Bacteriol* **190**:2597-2606.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5:**113-131.

Edwards, J. S. and Palsson, B. O. (2000). Robustness analysis of the *Escherichia coli* metabolic network. *Biotechnol Prog* **16**:927-939.

Enright, A. J., Van Dongen, S. and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**:1575-1584.

Ermolaeva, M. D. (2001). Synonymous codon usage in bacteria. *Curr Issues Mol Biol* **3**:91-97.

Esser, C. (2010). *Der Einfluß lateralen Gentransfers auf die Evolution prokaryotischer Genome am Beispiel von Alpha-Proteobakterien und Stämmen von Escherichia coli.* Dissertation, Heinrich-Heine-Universität Düsseldorf.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst Zool* **19**:99-113.

Fritzemeier, C. J. (2010). *Flusskopplung in biologischen Netzwerken.* Bachelor thesis, Heinrich-Heine-Universität Düsseldorf.

Ghosh, A. and Bansal, M. (2003). A glossary of DNA structures from A to Z. *Acta Crystallogr D Biol Crystallogr* **59**:620-626.

Grosjean, H. and Fiers, W. (1982). Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene* **18**:199-209.

Guindon, S. and Gacuel, O. (2003). A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**:696-704.

Hayashi, K., Morooka, N., Yamamoto, Y., Fujita, K., Isono, K., Choi, S., Ohtsubo, E., Baba, T., Wanner, B. L., Mori, H. and Horiuchi, T. (2006). Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110. *Mol Syst Biol* **2**:2006.0007.

Jensen, R. A. (2001). Orthologs and paralogs - we need to get it right. *Genome Biol* **2**:INTERACTIONS1002.

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**:275-282.

Jones, N. C. and Pevzner, P. A. (2004). *An introduction to bioinformatics algorithms*. MIT Press, Cambridge.

Karp, P. D., Riley, M., Paley, S. M., Pellegrini-Toole, A. and Krummenacker, M. (1999). Eco Cyc: encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res* **27**:55-58.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B. and Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**:241-254.

Koonin, E. V., Mushegian, A. R. and Rudd, K. E. (1996). Sequencing and analysis of bacterial genomes. *Curr Biol* **6**:404-416.

Koonin, E. V. and Wolf, Y. I. (2009). Is evolution Darwinian or/and Lamarckian? *Biol Direct* **4**:42+.

Koski, L. B. and Golding, G. B. (2001). The closest BLAST hit is often not the nearest neighbor. *J Mol Evol* **52**:540-542.

Krishna, S. S. and Grishin, N. V. (2004). The finger domain of the human deubiquitinating enzyme HAUSP is a zinc ribbon. *Cell Cycle* **3**:1046-1049.

Kuzniar, A., van Ham, R. C., Pongor, S. and Leunissen, J. A. (2008). The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet* **24**:539-551.

Larhlimi, A. and Bockmayr, A. (2006). A New Approach to Flux Coupling Analysis of Metabolic Networks. *CompLife* **4216/2006**:205–215.

Levar, N. (2009). *Benchmarking of Methods for the Identification of Orthologs.* Bachelor thesis, Heinrich-Heine-Universität Düsseldorf.

Li, L., Stoeckert, C. J. Jr. and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**:2178-2189.

Muller, J., Szklarczyk, D., Julien, P., Letunic, I., Roth, A., Kuhn, M,. Powell, S., von Mering, C., Doerks, T., Jensen, L. J. and Bork, P. (2010). eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res* **38**:D190-195.

O'Brien, K. P., Remm, M. and Sonnhammer, E. L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* **33**:D476-580.

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **27**:29-34.

Ottmann, T. and Widmayer, P. (2002). *Algorithmen und Datenstrukturen*. Spektrum Akademischer Verlag, 4. Auflage.

Palsson, B. (2006). *Systems Biology: Properties of Reconstructed Networks.* Cambridge University Press, 1st Edition.

Papin, J. A., Price, N. D., Wiback, S. J., Fell, D. A. and Palsson, B.O. (2003). Metabolic pathways in the post-genome era. *Trends Biochem Sci* **28**:250-258.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison*. Proc Natl Acad Sci USA* **85**:2444-2448.

Rastogi, S., Liberles, D. A. (2005). Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol Biol.* **5**:28+.

Remm, M., Storm, C. E. and Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* **314**:1041-1052.

Reed, J. L., Vo, T. D., Schilling, C. H. and Palsson, B. O. (2003). An expanded genome-scale model of *Escherichia coli* K-12 (**iJR904 GSM/GPR**). *Genome Biol* **4**:R54+.

Roth AC, Gonnet GH, Dessimoz C. (2008). Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* **9**:518+.

Tatusov, R. L., Galperin, M. Y., Natale, D. A. and Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**:33-36.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**:406-425.

Schilling, C. H., Letscher, D. and Palsson, B. O. (2000). Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathwayoriented perspective. *J Theor Biol* **203**:229-248.

Sonnhammer, E. L. (2004). Genome informatics: taming the avalanche of genomic data.*Genome Biol.* **6**:301+.

Wall, D. P., Fraser, H. B. and Hirsh, A. E. (2003). Detecting putative orthologs. *Bioinformatics* **19**:1710-1711.

Wapinski, I., Pfeffer, A., Friedman, N. and Regev, A. (2007). Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics* **23**:i549-558. (a)

Wapinski, I., Pfeffer, A., Friedman, N. and Regev, A. (2007). Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449**:54-61. (b)