Protein Structure Modelling using Evolutionary Information and Cryo-EM Data

Inaugural-Dissertation

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

vorgelegt von

Tatjana Maria Braun
aus München

Jülich, März 2017

aus dem Institute of Complex Systems, Strukturbiochemie (ICS-6)
des Forschungszentrums Jülich

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

Referent:	Jun. Prof. Dr. Gunnar Schröder
Korreferent:	Prof. Dr. Georg Groth
Tag der mündlichen Prüfung:	06.07.2017

Erklärung

Ich versichere an Eides statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist.
Die Dissertation wurde in dieser oder ähnlicher Form noch bei keiner anderen Institution eingereicht. Ferner habe ich bislang keine erfolglosen oder erfolgreichen Promotionsversuche unternommen.

_____

Signature

_____

Date

Abstract

Proteins can be seen as the major building blocks of life - they are involved in nearly every task of the cell and mediate most of its essential form and structure. To fully understand the biological role and function of a protein, its three-dimensional structure needs to be known.

In this work, two new approaches for computational protein modelling and structure prediction are introduced. The methods can contribute to filling the ever increasing gap between known protein sequences and structures.

(1) Experimental structure determination can be time-consuming, expensive and is not possible for every type of protein. Computational protein prediction has therefore become an important addition to the experimental methods for obtaining information about a protein's structure. While substantial progress has been made over the years, so far, predicting a protein's structure from only its sequence, i.e. *ab initio/de novo* protein structure prediction, is only possible for small proteins. Incorporating additional information, e.g. in form of sparse experimental data, enables structure prediction for larger protein structures as well. Following the discovery that inferred residue-residue contacts from co-evolving residues can successfully be used for protein structure prediction, we have combined evolutionary information with an iterative sampling protocol of the Rosetta molecular modelling suite to obtain accurate atomic models.

(2) Most atomic-resolution structures known to date were either determined using X-ray crystallography or NMR spectroscopy. While structure determination using cryo-electron microscopy has been limited to low-resolution models and complexes for many years, recent advances make it now possible to obtain density maps with near-atomic or even atomic resolution. These high-resolution density maps allow to directly build full-atomic models. This task is however challenging and, because being quite recent, has not been extensively explored so far. To address this problem, we have developed a completely new *de novo* model-building approach for cryo-EM maps in the near-atomic resolution range that combines backbone tracing, sequence non-specific fragment assembly, and automated side-chain assignment.

Zusammenfassung

Proteine sind wesentliche Bausteine des Lebens - sie sind in so gut wie jede Aufgabe der Zelle verwickelt und bestimmen deren Struktur und Form. Um die biologische Rolle und die Funktionen eines Proteins komplett zu verstehen, muss man seine dreidimensionale Struktur kennen.

In dieser Arbeit werden zwei neue Herangehensweisen zur computergestützten Proteinmodellierung und -vorhersage vorgestellt. Beide Methoden können in Zukunft helfen, die immer größer werdende Lücke zwischen der Anzahl an bekannten Proteinsequenzen und -strukturen zu verkleinern.

(1) Experimentelle Strukturbestimmung ist sehr zeitintensiv, kostspielig und nicht unbedingt für alle Proteinarten möglich. Die computergestützte Proteinvorhersage ist deshalb eine wichtige Ergänzung zu den experimentellen Methoden geworden, um Strukturinformationen von Proteinen zu bekommen. Die Methoden haben sich in den letzten Jahren kontinuierlich verbessert und weiterentwickelt, allerdings funktioniert die Vorhersage der Struktur ausschließlich basierend auf der Sequenz, sogenannte *ab-initio*- oder *de-novo*-Vorhersagen, auch heute nur für kleine Proteine. Das Verwenden von zusätzlichen Informationen, z.B. spärlicher experimenteller Datenlage, ermöglicht die Vorhersage auch für größere Proteinstrukturen. Basierend auf der Entdeckung, dass korrelierte Mutationen in Proteinfamilien und die daraus abgeleiteten Distanzinformationen zur korrekten Strukturvorhersage beitragen können, haben wir diese mit dem iterativen Sampling-Protokoll der Rosetta Software-Suite kombiniert.

(2) Die meisten der bis heute experimentell bestimmten Proteinstrukturen wurden mit Hilfe von Röntgenkristallographie oder Kernspinresonanzspektroskopie gelöst. Strukturaufklärung durch Kryo-Elektronenmikroskopie war lange Zeit auf niedrige Auflösungen von großen Komplexen beschränkt. Heutzutage können jedoch dank neuer Entwicklungen auch Dichtekarten mit sehr hoher Auflösung erstellt werden. Diese hochaufgelösten Dichtekarten machen es möglich, direkt atomare Modelle zu bauen. Der Prozess ist allerdings sehr fordernd und die Anzahl der Methoden für diese neue Aufgabenstellung noch sehr gering. Wir haben deshalb ein neues Protokoll zur *de-novo* Modellierung von Proteinstrukturen in hochauflösende Dichtekarten entwickelt: Das Protokoll baut in einem ersten Schritt die Cα-Positionen des Proteinrückgrats, vervollständigt diese mit Hilfe sequenz-unspezifischer Proteinfragmente und weist im Anschluss vollautomatisch

die Seitenketten zu.

Acknowledgements

I am incredibly grateful for all the help and support I got from the people that have accompanied me, all or part of the way, on my PhD journey.

First of all, I would like to express my deepest appreciation and thanks to my supervisor Gunnar Schröder for taking me under his wing after the first year of my PhD. During the past years, he gave me the freedom to follow my own ideas while constantly supporting me with his knowledge and experience.

In addition, I would like to express my sincere gratitude to Oliver Lange for advising the first year of my PhD and his continuos support thereafter.

I would also like to thank all the coauthors of my publications, with special thanks to Julia Koehler Leman for her constant help while writing my first complete manuscript.

This thesis wouldn't have been the same without all the people that I've shared my office with, had scientific discussions, went for lunch and coffee or simply enjoyed some break from work. I therefore want to thank my group members Amudha, Zhe W., Dennis, Carla and Michaela; Dusan, Martin and all the others from the computational biochemistry group; and finally Zhe Z., Zaiyong, Justin and my other colleagues from the former Lange group.

Finally, I would like to thank my family and friends for accompanying and supporting me before and during this work. I am deeply indebted to my parents and my partner Florian for their continuous support and encouragement.

List of Publications

This thesis includes peer-reviewed articles and submitted manuscripts about the work undertaken in the course of my doctoral project.

Published:

- **Braun T**, Koehler Leman J, Lange OF. *Combining Evolutionary Information and an Iterative Sampling Strategy for Accurate Protein Structure Prediction.* PLoS Comput Biol. 2015 Dec 29;11(12):e1004661.
- **Braun T**, Vos MR, Kalisman N, Sherman NE, Rachel R, Wirth R, Schröder GF, Egelman EH. *Archaeal flagellin combines a bacterial type IV pilin domain with an Ig-like domain.* Proc Natl Acad Sci U S A. 2016 Sep 13;113(37):10352-7.

Submitted:

- **Braun T**, Wang Z, Schröder GF. *Automatic Protein Structure Modelling into cryo-EM density maps using EMfasa.* Currently under review[1].

Other contributions carried out throughout the course of my PhD studies, not included in this thesis:

- Reichel K, Fisette O, **Braun T**, Lange OF, Hummer G, Schäfer LV. *Systematic evaluation of CS-Rosetta for membrane protein structure prediction with sparse NOE restraints.* Proteins. 2016 Dec 9. doi: 10.1002/prot.25224.
- Schöppner P, Csaba G, **Braun T**, Daake M, Richter B, Lange OF, Zacharias M, Zimmer R, Haslbeck M. *Regulatory Implications of Non-Trivial Splicing: Isoform 3 of Rab1A Shows Enhanced Basal Activity and Is Not Controlled by Accessory Proteins.* J Mol Biol. 2016 Apr 24;428(8):1544-57.
- **Braun T**, Orlova A, Valegård K, Lindås AC, Schröder GF, Egelman EH. *Archaeal actin from a hyperthermophile forms a single-stranded filament.* Proc Natl Acad Sci U S A. 2015 Jul 28;112(30):9340-5.

[1] as of February 2017

Contents

List of Figures

List of Tables

List of Abbreviations

3DEM	3-Dimensional Electron Microscopy
Å	Ångström
CASP	Critical Assessment of protein Structure Prediction
cryo-EM	Cryo-Electron Microscopy
DEN	Deformable Elastic Network
EM	Electron Microscopy
FSC	Fourier Shell Correlation
MCSA	Monte Carlo Simulated Annealing
NMR	Nuclear Magnetic Resonance
RMSD	Root-Mean-Square Deviation
SEM	Scanning Electron Microscopy
SNR	Signal-to-Noise Ratio
TEM	Transmission Electron Microscopy

Part I

Motivation and Background Theory

1. Motivation and Outline

Proteins are large and complex molecules that account for about half of a cell's total dry mass and mediate most of its essential structure and function [1]. Due to their large versatility and the resulting multitude of different functionalities that are fundamental to life, proteins are a focus of attention in biological research.
The function of a protein is directly related to its three-dimensional structure [2], making protein structure determination vital for understanding critical biological processes. The biological activity of a protein depends on atomic detail, and even slight changes in the molecular structure can significantly affect the protein's activity, potentially resulting in diseases. A detailed understanding of these changes is therefore necessary for designing a therapy.

Experimental structure determination is both expensive and time consuming, resulting in a large discrepancy between the number of known protein sequences and the number of experimentally solved three-dimensional structures, the so called "protein structure gap" [3]. In the past years, computational structure prediction has therefore become an important player to fill the ever increasing gap between known protein sequences and experimentally determined protein structures. While being an ongoing challenge for several decades, predicting a protein structure from its sequence alone remains a major problem. Current *ab initio* methods are able to accurately predict the structures of small proteins (100 residues) with increasing protein size, however, additional data, e.g. in form of sparse experimental data or sequence-based information, is needed to restrict the large conformational search space in order to obtain accurate models [4].

While most experimentally derived protein structures to date have been solved using X-ray Crystallography and NMR Spectroscopy, cryo-electron microscopy has recently started to become an increasingly important tool for structural biology. Thanks to recent advances, density maps obtained by cryo-EM can nowadays reach near-atomic and even atomic resolution, allowing to directly build atomic models. The process is, however, challenging and the number of methods tailored to this type of problem is still low.

In this thesis, two different approaches for computational protein structure calculation

and prediction are introduced: 1) evolutionary information, derived from correlated mutations in multiple sequence alignments of homologous protein sequences, was used in combination with an iterative sampling protocol of the Rosetta molecular software suite for accurate protein structure prediction, and 2) a new *de novo* modelling approach based on near-atomic resolution density-maps obtained by cryo-electron microscopy was developed and tested on several experimental density maps.

The thesis is organised in four parts. Part I shortly introduces the aim of this thesis and briefly summarises the fundamentals of protein structure, outlines standard methods for experimental structure determination and lays out common approaches in computational structure prediction to provide the reader with all the background information necessary to follow the findings in this thesis. Part II, starting on page 15, shows how evolutionary information coupled with an iterative sampling strategy can be used for accurate protein structure prediction. Part III, page 25 and ongoing, introduces a newly developed method, called EMfasa, for *de novo* structure modelling into near-atomic cryo-EM density maps. Both parts include detailed background information about their specific topics and include peer-reviewed articles presenting the respective methods and results. Part IV summarises the findings in this thesis and gives an outlook about the tools' applicabilities in the future.

2. Background Theory

This chapter briefly summarises the fundamentals of protein structure, the primary experimental methods used for structure determination in structural biology, and the main ideas in computational structure prediction. More detailed information and background theory tailored to the two different projects discussed in this thesis can be found in Part II and III, respectively.

2.1 Fundamentals of Protein Structure

The major building blocks of a protein are the so called amino acids - organic compounds containing an amino group (-NH2), a carboxyl group (-COOH), a hydrogen atom, and a specific side-chain (R group) that are all linked to a central carbon (called Cα). There are 22 proteinogenic amino acids, i.e. amino acids that are incorporated into proteins during translation, whereof 20 are in the standard genetic code and two can be added by special translational mechanisms [5]. The side-chain is specific to each amino acid and accounts for its specific chemical properties that can be loosely grouped into classes, as shown in Figure 2.1.

Amino acids can be linked together by forming a so called peptide bond through a reaction of their respective carboxyl and amino groups, see Figure 2.2. While two or more amino acids connected via a peptide bond are referred to as a *peptide*, a long, continuos,

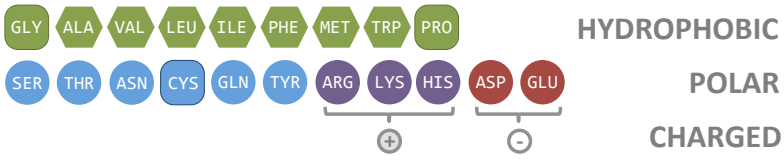

Figure 2.1 The 20 standard amino acids and their properties. The standard amino acids are commonly grouped into three classes: hydrophobic, polar, and charged. These classes can furthermore be subdivided into several subclasses, e.g. aromatic, large, small, and so forth. Some amino acids stand out for their unique properties: Proline forms a bond with its own amino group and therefore has only limited flexibility. It is also referred to as helix-breaker. The side chain of Glycin consists only of a single hydrogen atom, giving it the unique property of being achiral. Cystein can form disulfid bonds via oxidation of its sulfhydryl group and therefore greatly increases the stability of a protein's structure.

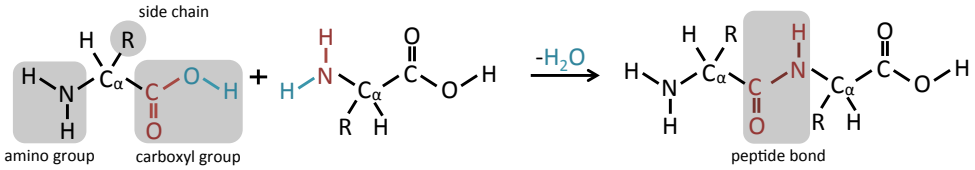

Figure 2.2 The peptide bond. The peptide bond is synthesised through a reaction of the respective carboxyl and amino groups of two amino acids during which a molecule of water is released. The process is therefore called a dehydration synthesis or condensation reaction. The two linked amino acids are called a dipeptide. The peptide bond is planar and rather rigid, resulting in important implications for the three-dimensional structure of a protein.

and unbranched peptide chain is called a *polypeptide*. Proteins consist of one or more long polypeptides that are joined together, cf. Section 2.1.1.4.

2.1.1 Four Levels of Protein Structure

Most proteins fold into unique three-dimensional structures. This structure has been organised into a four-tiered hierarchy as described in [6]. The four levels are briefly summarised in the subsections below.

2.1.1.1 Primary Structure

The primary structure of a protein describes the linear sequence of amino acids in the polypeptide chain. It is generally listed from its amino terminus (N-terminus) to its carboxyl terminus (C-terminus).

2.1.1.2 Secondary Structure

The secondary structure of the protein describes regular and characteristic local conformations of the polypeptide. These local conformations are formed by hydrogen bonds between the amine hydrogen and carbonyl oxygen atoms of the participating residues. The most common secondary structures, suggested as early as 1951 [7], are alpha (α) helices and beta (β) sheets. These structures are highly regular: the specific dihedral angle combinations are approximately repeated throughout the entire secondary structure element. Other forms of helices with energetically favourable hydrogen-bonding patterns, such as the 3_{10} and π helices, are only rarely observed in natural proteins.
The regular secondary structures described above are usually connected by irregular secondary structural elements, so called loops.

2.1.1.3 Tertiary Structure

The tertiary structure of a protein, also called fold, refers to the global three-dimensional structure. It is hold together through interactions between the side chains of the amino acids rather than backbone interactions which are primarily responsible for generating the secondary structure.

2.1.1.4 Quaternary Structure

Many proteins do not function as a monomer, but as a noncovalent association of two or more independent polypeptides. The quaternary structure describes the arrangement and number of the these subunits in the resulting multimeric or multisubunit complexes.

2.2 Experimental Structure Determination

Several experimental methods can be used to determine the three-dimensional structure of a protein. The most common methods used to study the structure of a protein are X-ray crystallography, NMR spectroscopy, and electron microscopy: the great majority of protein structures deposited in the RCSB Protein Data Bank (PDB) archive [8] has been solved by using at least one of these methods, see Table 2.1. All three methods, as well as their advantages and disadvantages, will be briefly summarised in the subsequent subsections.

2.2.1 X-ray Crystallography

Most of the experimentally derived structures to date have been determined using X-ray crystallography, c.f. Table 2.1.
For X-ray crystallography, as described in [10], a protein gets purified and crystallised and the crystal is subsequently subjected to a beam of X-rays. The proteins in the crystal diffract the X-ray beam into characteristic patterns that can be analysed to determine the

Table 2.1 PDB holdings breakdown

Exp. Method	Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
X-Ray	97349	1709	4926	4	103988
NMR	9917	1138	231	8	11294
EM	699	29	245	0	973
Other	255	7	8	14	284

Holdings breakdown of the RCSB Protein Data Bank [8] as of March 2016. Data taken from [9].

distribution of electrons from which an electron density map can be calculated. This map can be used to determine the location of each atom.

Each reflection on the diffraction pattern collected during X-ray crystallography is characterised by its amplitude and phase which are both represented in a mathematical description called "structure factor". The intensity of the reflection directly provides information about the amplitude, the information on the phase of the diffracted radiation is however lost. This is referred to as the "phase problem". To determine the electron density distribution in the crystal, both phase and amplitude need to be known. The most common methods used for phase determination are isomorphous replacement and molecular replacement [10].
The isomorphous replacement method [11] is usually used if no closely related structures are available. The method compares the data obtained from a protein crystal to the data obtained from a protein crystal with attached heavy atoms. The differences in the data sets result from the heavy atoms which allow to determine their position in the protein molecules. The positions can be used to determine the protein phase angles and, in combination with the amplitudes, the structure factors can be calculated.
The molecular replacement method [12] can be used if closely related protein structures are available. In this method, structure factors are calculated from the coordinates of the related protein structures and the corresponding phases are applied to the data set of the protein of interest. This method therefore results in a bias towards the model in the initial structure factor calculations.
Having both the amplitudes and initial phases, an initial model can be built. This initial model is usually not perfect but can be used to improve and refine the phases and thereby leading to an improved model. This step can be repeated until the agreement between the diffraction data and the model is maximised.

To measure the agreement between a model and the experimentally-observed X-ray diffraction data, the R-factor is used. The R-factor, defined as $R = \sum ||F_o| - |F_c|| / \sum |F_o|$, measures the difference between the structure factor amplitudes calculated from the model ($|F_c|$) and the ones obtained from the experimental diffraction data ($|F_o|$). Well refined models correspond to R-values of less than 20 percent whereof the best-refined structures are characterised by values below 10 percent [13]. However, the R-value is not completely reliable as structures with acceptable R-factor values can have significant errors [14]. High R-values might be a result of overfitting the experimental data by using too many parameters, e.g. in form of a large amount of water molecules, and do not necessarily indicate an accurate model [13].
In 1992, Brügner introduced the R_{free}-value, which is less prone to overfitting than the R-

factor [15]. Before carrying out the refinement, the experimental diffraction data is split into two sets: a small test set and a working set. The working set is used for the structure refinement while the test set is used to calculate the R_{free}-value in the same way as the standard R-factor. The R_{free} is therefore not influenced by the refinement procedure and a significant difference to the R-factor can therefore indicate that a model has been overfitted [13].

X-ray crystallography can provide very detailed atomic information, howeve, the process of crystallisation is very difficult and time consuming and can not be done for every type of protein. The crystallisation is therefore seen as the bottleneck for X-ray crystallography [16]. While being a great method for rigid proteins, it is not very suitable for studying flexible proteins.

2.2.2 NMR Spectroscopy

Nuclear magnetic resonance (NMR) is the second most used technique to study protein structures in structural biology, c.f. Table 2.1. The method can be used not only to provide atomic details about a protein's structure, but also for studying protein dynamics [17] and folding [18].
For NMR, as described in [19], a protein is purified, placed into a strong magnetic field, and then probed with radio waves. The list of observed resonances is then analysed to obtain a list of structural restraints. These restraints can be separated into three different classes [20]: distance restraints, e.g. gained from nuclear overhauser effect experiments (NOESY), torsion angle restraints obtained from vicinal coupling constants and orientational restraints derived from residual dipolar couplings (RDCs). These restraints can subsequently be used in structure calculations to generate an ensemble of models of the protein.
However, for large molecular structures it becomes difficult to resolve the individual signals (resonances) of the active nuclei due to increased spectral complexity and loss in sensitivity, putting a practical limit to the size that can be studied in detail by NMR [21].While the size limit has been pushed further over the years due to new techniques (TROSY, i.e. transverse relaxation-optimized spectroscopy) and labelling schemes [22], NMR structural studies of larger molecules are usually carried out with focus on accompanying NMR studies, such as defining ligand affinity and protein dynamics [23].

2.2.3 Electron Microscopy

Three-Dimensional Electron Microscopy (3DEM) can be used to determine the structures of large macromolecular complexes. A beam of electrons is used to image individual

proteins, each of them being observed in different orientations. These images are in the next step aligned and averaged to extract 3D information in form of a density map. A popular form of electron microscopy in structural biology is the so called cryo-electron microscopy where the sample is studied at cryogenic temperatures. It allows the observation of the sample without any prior staining or other fixation and therefore shows the specimen in their near-native environment. Until recently, the resolution of the resulting density maps was most of the time too low to see each atom and electron microscopy data was therefore mainly used in combination with other experimental methods, such as X-ray crystallography and NMR spectroscopy, or to dock atomic structures to obtain the complex. Electron microscopy, and more specifically Cryo-EM, is described in more detail in Part III Section 1.2.

2.3 Computational Structure Prediction

Computational structure prediction methods aim to predict the structure of a protein from its sequence with an accuracy that is comparable to the results achieved experimentally and thereby fill the large gap between known protein sequences and resolved protein structures. Experimental structure determination is not only expensive and time consuming, but also not possible for every protein: some proteins cannot be crystallised and can therefore not be used for X-ray diffraction while others are too large for NMR analysis. Computational protein modelling is therefore the only way to obtain structural information for proteins that are not suitable for any type of experimental structure determination.
The field of computational structure prediction has been a highly debated over the years and a lot of progress has been made so far. The three major approaches to three-dimensional structure prediction will be discussed in the following sections.

2.3.1 Comparative Modelling

Comparative modelling, also known as homology modelling or template-based modelling, is used to predict the structure of proteins that have a sequence that is similar to the one of a protein with previously determined structure [24].
The method is primarily based on the observation that, during evolution, the structure of a protein is more stable and undergoes less changes than its amino acid sequence, i.e. proteins of similar sequences adopt practically identical structures, and distantly related sequences still fold into similar structures [25].
In 1999, Rost analysed more than a million sequence alignments between protein pairs of known structures and defined an accurate rule to distinguish between pairs of similar

and non-similar structure based on both sequence identity and alignment length [26].

2.3.1.1 Method Overview

Comparative modelling or homology modelling is a multi-step process and can be broken down into four main steps [27]: template selection and target-template alignment, model construction, model refinement, and model assessment. In the initial step, possible modelling templates with high sequence similarity are identified by aligning the target sequence to all the sequences of known structures in the PDB [8] with alignment programs such as BLAST [28] and FASTA [29]. Having found potential templates, the best one is chosen based on criteria including sequence similarity, protein function and the predicted secondary structures. Having an alignment between the target sequence and the chosen template model, the model construction begins by copying the atoms of those template residues that are aligned to residues in the target sequence. The refinement step later on focuses on modelling the loops and side chains.

2.3.2 Fold Recognition

Fold recognition techniques, also known as protein threading methods, try to find a structural relationship between two proteins although there is no obvious sequence similarity and to make the prediction by threading each amino acid in the target sequence to a position in the template structure [30]. A good model will therefore have structural similarity to a known fold, the sequence however is not obviously similar.

2.3.2.1 Method Overview

Protein threading can roughly be divided into the following steps [31]: Initially, the target protein is aligned to each of the proteins in a structure database, e.g. PDB [8], using a scoring function. Based on these alignments, one or several templates are chosen and the aligned residues of the target are copied from the template. Finally, unaligned residues are predicted using loop modelling and side-chain packing.

2.3.3 *De-novo*/*Ab initio* Prediction

Ab initio or *de-novo* methods aim to predict a protein's tertiary structure from its amino acid primary sequence by relying primarily on physical principles and not on existing structural templates. These methods are of particular interest for the large number of protein sequences for which no homologs with three-dimensional structural information are known [32].

While Anfinsen's dogma [33] states that, at least for small proteins, the native structure of a protein is determined only by the protein's amino acid sequence and that, in the absence of large kinetic barriers, it corresponds to the state with the lowest free energy, no universal *ab-initio* method has been found so far due to one or a combination of the following reasons: 1) applied energy functions and force fields may not be realistic enough [34] 2) the number of possible conformations is too large to be efficiently sampled [35] 3) the native conformation may not be identical to the minimal energy structure of the utilised energy function [36].
The most detailed representations of a protein include all its atoms and the surrounding solvent molecules. Due to the large number of atoms, this representation is extremely expensive and most methods for protein structure prediction therefore have involved some significant complexity reduction [37]. These simplifications include united atom representations, a limited number of side-chain conformations and implicit solvent. The energy or scoring functions that are used to evaluate these low-complexity models therefore must be able to represent the forces responsible for protein structure, e.g. solvation, strand hydrogen bonding, in a manner robust to the model's limited accuracy [37]. Once a number of coarse-grained models that are potentially close to the native structure are identified, they can be refined using full-atom energy functions.

2.3.3.1 Fragment-based Methods

Fragment assembly-based methods are currently the most successful approaches for the *de-novo* protein structure prediction problem when no structural homologs are present. These methods rely on the assumption that the conformational space of the protein backbone is finite and stabilised by local interactions, and that the model of the target structure can therefore be generated by piecing together continuous subsets, so-called fragments, from known protein structures. The idea for using small fragments to assemble new structures was first introduced by Bowie and Eisenberg in 1994 [38]. Prominent methods using this approach are FRAGFOLD [39, 40] and Rosetta [41, 42]. Rosetta will be described in more detail in Part II Section 1.3 on page 17.

Part II

Structure Prediction using Evolutionary Information

1. Introduction

1.1 Motivation

In 2011, it has been shown that evolutionary information in form of correlated mutations in multiple sequence alignments of evolutionary related protein sequences can help to predict protein structures with explicit atomic coordinates quite accurately: The co-evolving residues give insight into which residues are in close proximity in the protein structure and therefore can be used, in form of distance restraints, to drive a computer simulation with an atomic-scale physical model of the protein structure from a random starting conformation to a native-like conformation. Since that breakthrough, much effort is being put into the improvement of these contact predictions. Due to the statistical nature of the predictions, their accuracy will however always be limited, e.g. contain a fraction of erroneously predicted contacts, and structure prediction protocols that are tolerant to incorrect distance restraints are needed.
In this project, we have combined evolutionary information with an iterative sampling strategy of the Rosetta molecular modelling suite and benchmarked it on a diverse set of globular proteins.

1.2 Structure Information based on Sequence Variation

As of February 2017[1], NCBI Reference Sequence database RefSeq [43], a non-redundant set of nucleotide and protein sequences, contains almost 80 million protein sequences. This vast number of available protein sequences allows for analysing multiple sequence alignments of subsets of these to infer homology and to study the evolutionary relationships between them.

The tertiary structure of proteins is usually better conserved than sequence in evolution [25, 44]. Hence, members of homologous protein families often have similar functions and structures while their sequence similarity can be comparatively low. During evolution, gene sequences undergo random changes and mutations, whereof some may be

[1] RefSeq Release 80 including 78,028,152 proteins, 17,862,608 transcripts and 66,224 organisms

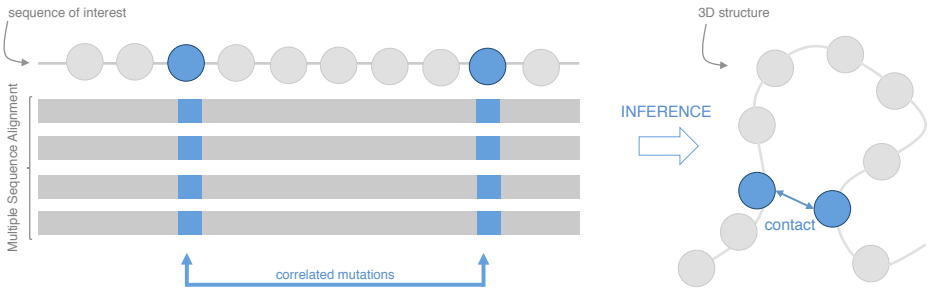

Figure 1.1 Using correlated mutations for protein structure prediction. Observed correlations in multiple sequence alignments of homologous protein sequences can be used to infer contacts between residues in the three-dimensional structure of a protein. Figure inspired by [47, 48].

detrimental for both function and structure, while others are neutral. To maintain structure and function, the evolution in families of proteins is therefore heavily constrained by structural and functional features such as maintaining a hydrophobic core, secondary structures, active sites, and buried and charged hydrogen bonds[45, 46].

In the early 1990s, patterns of correlated mutations in multiple sequence alignments of protein families have been interpreted for the first time as an indication of probable physical contact in the three-dimensional protein structure [49–52]. An illustration of the basic idea is shown in Figure 1.1. The accuracy of the predicted contacts has however not been good enough to drastically improve structure prediction methods as they were using "local" statistical models, e.g. mutual information scores, that were not able to separate direct from indirect contact information [48]. These indirect (transitive) correlations, explained in Figure 1.2, greatly limit the accuracy of predicted residue-residue contacts.

Using global statistical models instead, i.e. treating pairs of residues dependent on each other, reduces the effect of transitivity and thereby substantially increases the prediction accuracy [47, 53–56]. Since then, many methods that discriminate between direct and indirectly coupled mutations for residue-residue contact predicition have been developed: GREMLIN [57], CCMPred [58] and plmDCA [59, 60] use pseudolikelihood approaches while PSICOV[61] achieves the separation of direct and indirect correlations by inverting a residue-residue covariance matrix.

In 2011, residue-residue contacts predicted using global statistical models, e.g. maximum entropy, were shown to be accurate enough to fold a protein to reasonable accuracy using a method called EVFold [47]. Other methods that have been used in combination with evolutionary information for protein structure prediction include PconsFold [62], FRAG-

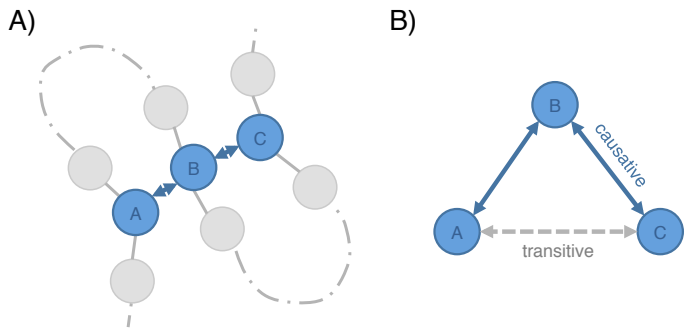

Figure 1.2 Transitive correlations. A) Physical contacts in a protein structure B) Observed correlations in a multiple sequence alignment. The correlations of residues A and B and B and C are causative, as they reflect direct physical contacts. The correlation between A and C is however transitive due to their mutual interaction with residue B.

FOLD [63], and CONFOLD [64]. Recently, in terms of the eleventh Critical Assessment of protein Structure Prediction (CASP11), structure predictions using coevolution derived residue-residue contact information as restraints were once again shown to considerably increase the accuracy of template-free structure modelling [65].
While the initial predictions were focusing on globular proteins, predicted contacts from evolutionary information have also been used to predict protein-protein complexes [66, 67] and the structures of membrane proteins [68, 69].

1.3 *De Novo* Structure Prediction with Rosetta

Rosetta is a large software suite that nowadays includes algorithms for computational modelling and analysis of protein structures, e.g. *de novo* protein design, enzyme design, ligand docking and structure prediction of proteins and protein complexes [70]. Originally, Rosetta was developed for *de novo* protein structure prediction and, since its release in 1997 [41], regularly is amongst the top performers of the biennial community-wide Critical Assessment of protein Structure Prediction (CASP) [42, 71–73]. In the subsequent sections, Rosetta's scoring functions, the standard protocol for *de novo* predictions and an iterative sampling strategy are briefly described.

1.3.1 The Rosetta Energy Functions

Rosetta applies two kinds of scoring functions: a centroid/low-resolution energy function and an all-atom energy function. Both functions, which are reviewed in detail in [74] and [70], are briefly summarised below:
In the low-resolution mode, the side chains are represented by a centroid that is located

at the side-chain center of mass. The energy function includes terms for solvation and electrostatics that are based on observed distributions in protein structures. Hydrogen bonding between β-strands is included by probabilistic descriptions and steric clashes between centroids and backbone atoms are penalised, representing the repulsive component of the van der Waals forces. The attractive forces of the latter are included by rewarding globally compact structures.
In the more physically realistic all-atom energy function, the van der Waals forces are modelled by the 6-12 Lennard-Jones potential. In addition, the energy function includes a solvation approximation, a structure- and orientation-dependent hydrogen bonding potential for explicit hydrogen bonding, a residue-based pair potential to model electrostatics (now, with the introduction of the Talaris energy function in 2013, replaced with an explicit Coulombic electrostatic term [75]) and a knowledge-based conformation-dependent amino acid internal free energy term.

1.3.2 Standard Rosetta *De Novo* Prediction Protocol

The standard Rosetta protocol for *de novo* structure prediction, described in detail in [35, 39, 76, 77], consists of two stages: a coarse-grained sampling stage and an optional fine-grained refinement stage. The sampling stage starts with a fully extended protein chain and "folds" the protein by replacing the torsion angles in randomly selected 3 or 9 residue windows with torsion angles of known protein fragments with similar local sequence in a series of thousands of Monte Carlo fragment replacement moves. The scores of the new conformations are evaluated using the low-resolution energy function using a centroid representation for the side chains as described above and are accepted or rejected based on the Metropolis criterion [78]. This Monte Carlo minimisation drives the structure toward the global minimum of the smoothed energy surface of the low-resolution energy function.
Once the coarse-grained sampling stage is finished and a full-atom refinement is desired, the side chains are added and optimised by minimising Rosetta's all-atom energy function.

1.3.3 Restraint-Guided Structure Calculations

Structure calculations in Rosetta can be guided towards the lowest energy conformations in the folding landscape by incorporating various types of restraints, including but not limited to restraints on atom-pair distances, angles, dihedral angles, and location in coordinate space. They are implemented as a biasing potential that usually causes an energy penalty if a specific parameter violates the ranges or values defined by the restraint.
Restraints derived from various sources, such as e.g. NMR [79–83], cross-links [84], and

coevolution information [65], have successfully been used for protein structure prediction in the past.

1.3.4 RASREC, an Iterative Sampling Strategy

For proteins with more than 100 amino-acids, the standard de novo protocol, generating thousands of completely independent protein models, generally runs into sampling issues and is not able to converge to the native fold [4]. While structural guidance, e.g. in form of sparse NMR data, helps to increase this size limit, the success rate for proteins over 15 kDA is still not robust [4, 81] and improved sampling methodologies are required.

The Resolution-Adapted Structural RECombination approach RASREC, described in [4, 81], seeks to improve the sampling near the native structure by recombining frequently occurring structural features in restraint-guided structure calculations during different resampling and refinement stages. In the subsequent paragraph, the RASREC protocol will be briefly described.

RASREC consists of 6 stages, one initial exploration stage and five resampling stages. The first four stages use the Rosetta low-resolution energy while the remaining two are using the Rosetta full-atom energy to generate well-refined fullatom models. During each of its six iteration stages, RASREC runs batches of independent structure calculations in parallel and stores and updates a user-specified number of all-time best models in a structural pool. The resampling stages (II-VI) are then seeded with structural information of the structural pool which intensifies the exploration of the most promising regions in the conformational space. The resampling stages differ the way how they use the structural information in current structural pool: Once the resampling technique of a stage has been sufficiently explored, i.e. the acceptance rate of structures into the pool decreases below 10 percent, the stage terminates and the next stage starts. After the final resampling stage is finished, the protocol is done and the structures in the structural pool represent the final models.

It was shown previously that this sampling methodology makes RASREC more robust and requires less data than the standard Rosetta prediction protocol [4, 81, 85].

2. Methods, Results, and Discussion

2.1 Publication 1: Combining Evolutionary Information and an Iterative Sampling Strategy for Accurate Protein Structure Prediction

We have combined the RASREC approach with evolutionary restraints and used the protocol for predicting globular protein structures with high accuracy. The work has been published in 2015 in *PLoS Computational Biology* [86]. The article is summarised below and both the article and the supplemental information are included as Appendix A (page 89 cont.) of this thesis.

2.1.1 Summary

It has been shown previously that the accuracy of *de novo* protein structure prediction can be significantly improved by integrating evolutionary information that can be used to infer spatial proximity in the three-dimensional structure.
In this study, we have combined the resolution-adapted structural recombination approach RASREC of the molecular modelling suite Rosetta with evolutionary information in form of intra-protein residue-residue contacts for accurate protein structure prediction. RASREC has been shown previously to converge faster to near-native models and to be more robust against incorrect distance restraints than standard prediction protocols and is therefore perfectly suited for restraints obtained from predicted residue-residue contacts with limited accuracy.
We have tested our protocol on 28 globular proteins and compared it to the results of the EVFold web server using identical contact predictions. Our method was able to converge for 26 of the 28 targets and improved the average TM-score of the entire benchmark set from 0.55 to 0.72 when compared to the EVFold web server. In addition, we showed that the improved sampling and high error tolerance of the underlying RASREC algorithm enables accurate structure prediction in cases where the accuracy of the predicted contacts is dropping below 50 percent.

2.1.2 Contribution

For this article, I participated in the design, performed the experiments, analysed the data, generated all figures, tables, protocol capture and supporting information and wrote the majority of the manuscript. I am corresponding author of this article.

Part III

De-Novo Modeling based on Cryo-EM Density Maps

1. Introduction

1.1 Motivation

Structure determination of biological macromolecules using cryo-electron microscopy has been limited to large complexes and low-resolution models for many years. Thanks to a new generation of direct electron detectors and powerful image processing routines, large macromolecules can now be obtained at near-atomic or even atomic resolution using cryo-electron microscopy (cryo-EM). These high resolution structures give rise to a need for methods for *de novo* model building into cryo-EM density maps.
This chapter will give a short summary of the current state-of-the art cryo-EM techniques and an outline of the current de novo modelling approaches.

1.2 Cryo-Electron Microscopy of Biological Macromolecules

Over the past two decades, single-particle cryo-electron microscopy (cryo-EM) has evolved to be a powerful technique to solve macromolecular structures at high resolution. It has become a popular technique in structural biology and was consequently chosen as Method of the Year 2015 by the peer-reviewed scientific journal *Nature Methods* [87]. This section gives a brief overview about the historical background, the general strategy, and recent advances in cryo-EM. If not stated otherwise, the information given in the following subsections is based on [88–90].

1.2.1 Background and General Information

The first electron microscope was introduced by M. Knoll and E.Ruska in 1931 [91]. This prototype had less resolving capabilities than light microscopes at that time [92]. Since then, a lot of progress has been made, making it possible to nowadays obtain complex molecular structures at near-atomic resolution [93].

There are two types of electron microscopes: transmission electron microscopes (TEM) and scanning electron microscopes (SEM). The latter scans the sample with a focused beam of electrons and is used to image surfaces of cells and small molecules in 3D. In

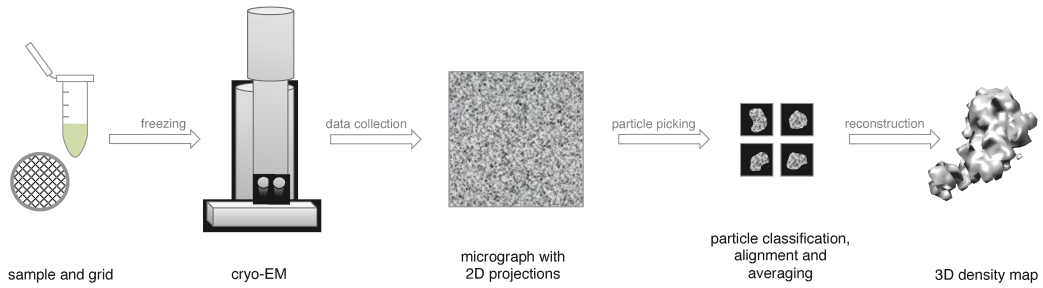

Figure 1.1 Single-particle cryo-EM workflow. The purified protein sample is applied to a grid and rapidly frozen (vitrified). Afterwards, an electron beam is used to illuminate the specimen. The resulting 2D projections of the individual protein molecules in different orientations are classified, oriented, averaged and combined into a 3D density reconstruction. The figure is inspired by [90].

TEM, electrons are transmitted through the specimen and the samples internal composition with up to atomic resolution can be imaged in 2D.

1.2.2 From Protein Sample to 3D Model

The principle for obtaining a 3D model from a protein sample using single-particle cryo-EM is illustrated in Figure 1.1. Briefly, the purified protein sample is applied to a grid and rapidly frozen (vitrified). Afterwards, an electron beam is used to illuminate the specimen. The resulting 2D projections of the individual protein molecules in different orientations are classified, oriented, averaged and combined into a 3D density reconstruction.

1.2.2.1 Sample Preparation and Image Formation

In cryo-EM, the sample is analysed in cryogenic temperature and a beam of electrons is transmitted through the specimen to form an image.

Sample Preparation

Cryo-EM analyses ice-embedded specimen. These are prepared by applying the purified protein sample in an aqueous solution to a grid that consists of tiny holes in a film, e.g. carbon, and a supporting metal frame. The grid is in the next step rapidly plunged into a cryogen - usually liquid ethane - for flash-freezing whereupon the particles get trapped in a thin film of vitrious ice. This kind of specimen preparation has several advantages when compared to other preparation techniques, e.g. negative staining. First, this technique makes it possible to obtain images of fully hydrated molecules, i.e. in their native environment, without being distorted by stain. In contrast to using extraneous contrasting agents, such as heavy atom salts in negative staining, these frozen-hydrated specimen

allow to measure image contrast that is related to the structure of the biological macromolecule itself. Additionally, the film of vitrious ice reduces the radiation damage by trapping free radicals produced by ionisation during electron irradiation (cf. subsequent paragraph).

Image Formation

For image formation, an electron beam passes through the frozen sample and during their transmission, individual electrons either get scattered by the specimen or pass through it unscattered. Scattering can take place in two different forms: Elastic scattering occurs with no loss of energy, while inelastically scattered electrons transfer energy to the electrons in the specimen. The electrons that are emerging from the sample, both unscattered and elastically scattered, are collected and focused by lenses to create a diffraction pattern that, for thin specimen, is directly related to density variations in the sample. All the structural information in the pattern is obtained from the elastically scattered electrons only.
The inelastically scattered electrons lead to radiation damage due to their energy transfer to the specimen which, after accumulation, can break molecular bonds and thereby destroy the sample. It is therefore necessary, even at the low temperatures of cryo-EM, to only use low doses of electrons, i.e. the weakest possible image exposures that enable obtaining a measurable signal. The signal-to-noise ratio (SNR) of the recorded images is therefore very low and limits the information that can be obtained from a single particle image. Therefore, to obtain high-resolution information, several images need to be averaged as described hereinafter.

1.2.2.2 Single Particle Analysis

The electron micrographs, generated as described in the previous section, consist of 2D representations of 3D molecules in various orientations and conformations. To obtain a three-dimensional structure from the micrographs, the 2D representations of the individual particles need to be combined. There are a large number of software suites, e.g. EMAN [94], SPIDER [95], and IMAGIC [96] that offer comprehensive sets of tools to carry out various strategies for image processing and 3D reconstruction.

Alignment and Classification

The images of the individual particles are very noisy. As mentioned in the previous section, only low electron doses can be used for imaging and the resulting images have high noise relative to the signal given by the particles, making them hard to interpret. To improve the signal-to-noise ratio, several images need to be averaged. In theory, images

of approximately 10,000 particles are needed to obtain an atomic resolution structure [97]. In practice, this number is even significantly larger. As the micrographs contain particles in multiple different orientations and (possibly) conformations, the images of the individual particles need to be sorted based on their structural features prior to the averaging step. For sorting, statistical methods, such as principal component analysis, multivariate analysis or covariance analysis are used. After the classification, related images are averaged to obtain characteristic projection views of the complex with much better signal-to-noise ratio than the original images.

3D Reconstruction

To obtain a full 3D structure, a sufficient number of angular views of the object need to be combined. In single-particle-analysis, the sufficient number of orientations of the object is obtained through the random distribution of orientations of several particles on the micrograph. To combine the averaged 2D projections to a 3D reconstruction, their relative orientations must be known. The steps to do so are based on the central projection theorem, which states that for a 3D object, the Fourier transform of each 2D projection is a central slice through the 3D Fourier transform of the object [98]. The orientations of the different particles are determined iteratively by comparing them against the projections of initial models. An initial model can be obtained by the random conical tilt method [99], models of related structures or angular reconstitutions [100].

1.2.2.3 Resolution

In theory, cryo-EM should be able to provide atomic resolution structures of biological molecules as small as 100kDa in molecular weight [97]. In practice however, the final resolution that can be obtained by single-particle analysis is dependent on several factors: the quality of the original data, the number of particles in the data, and the accuracy of the determined orientation parameters. All these aspects depend on the amount of high- and low-resolution information present in the images [101]. The high-resolution information determines the possible final resolution of a 3D reconstruction, while the low-resolution information, i.e. image contrast, is required to visualise and pick the particles. For frozen-hydrated biological samples, which are sensitive to radiation damage [102] and therefore need to be imaged with low electron doses, contrast is achieved by recording images in some defocus. This however reduces the high-resolution signal. Therefore, one has to find a good balance between defocus and contrast for best results [101].

In general, the resolution for cryo-EM density maps is assessed using the Fourier Shell Correlation (FSC) [103], measuring the normalised cross-correlation coefficient between

two 3-dimensional volumes as a function of spatial frequency. Therefore, to measure the FSC, the particle data set is split into two groups of the same size. A common approach is to use the odd particle images for one set and the other half for the second set. For each of the two datasets, independent 3D reconstruction are carried out and the resulting volumes are compared to determine the FSC curve.

1.2.3 Recent Advances in Cryo-Electron Microscopy

Starting in 2013, new structures, including ribosomes from human pathogens [104] or mitochondria [105], ribosomes in complex with a protein translocase [106], ion channels [107, 108], and a key enzyme in the biogenesis of methane [109] have been obtained by cryo-EM at near atomic resolution, marking the start of a new era in the field of cryo-EM [110]. This improvement of resolution arises from a combination of a new generation of direct electron detectors that record images with previously unseen quality and new image-processing tools that correct for sample movements and are able to classify images according to different structural states [101, 111].

1.2.3.1 Direct Electron Detectors

Poor image contrast is a major challenge in cryo-EM. To restrict radiation damage, the number of electrons that can be used is limited, leading to noisy images. It is therefore important to efficiently detect the available electrons. Direct electron detectors, i.e. monolithic active pixel sensors (MAPS), make it possible to see individual incident electrons [111, 112]. The new detectors have a better detective quantum efficieny (DQE), which describes the signal and noise performance in a digitally recorded image over the spatial frequency range, as conventional film [113, 114]. In addition, they have the advantage of immediate feedback and are able to record movies at a rate of many frames per second [110].

1.2.3.2 Advanced Image Processing Software

Two developments in image processing have complemented the new direct electron detectors: one addresses the problem that the electron beam induces movement of the sample, and the other addresses the difficulty in dealing with structurally heterogenous samples.

Once the electron beam hits the thin film of vitrious ice, chemical bonds in the sample are broken, cf. Section 1.2.2.1. Even early on in the exposure, these forces induce motion in the sample. Image recording on conventional media, such as CCD cameras or photographic film, takes a view seconds, which, due to these small movements, leads to

blurred high-resolution features. The new direct electron detectors, as described before, provide movie-functionality. Comparison of successive frames then allows to detect and trace movements that can later be reversed computationally [115–117]. This way, a motion corrected image can be produced which is much sharper than images obtained on conventional media.

Usually, macromolecular structures or complexes contain more than one unique 3D structure due to conformational heterogeneity. If projections of an heterogenous set of protein structures are combined into one 3D reconstruction without prior classification, the resulting map will be of low resolution. The classification is however complicated, as it is not obvious how to distinguish between projections from different orientations or projections from slightly different structures. Unsupervised classification has been made possible through development of statistical algorithms that were based on maximum-likelihood procedures [118–120]. Software packages including such maximum likelihood-based classification and refinement methods are FREALIGN [121] and RELION [122].

1.3 Structural Interpretation of Electron Microscopy Density Maps

This section briefly explains methods that are used to interpret cryo-EM density maps. Commonly used refinement methods, i.e. methods that optimise the side-chains of a protein structure to improve the fit to the density map, such as refinements using CNS [123, 124], Rosetta [125, 126], or PHENIX [127], are not explicitly discussed.
The methods used to analyse and interpret the structural information in a density map are highly dependent on the map's resolution [128]. Table 1.1 lists the features visible in a density map as a function of resolution as described in [129].

1.3.1 Segmentation

Cryo-EM reconstructions usually consist of several different molecular components of the same or different proteins and/or nucleic acids related to one another in large macromolecular complexes. Segmentation of cryo-EM maps, i.e. identification of regions belonging to individual proteins or subunits, is therefore an important task in their interpretation.
Nowadays, there exist several tools for automatic, model-free segmentation: *Segger*, developed as plugin for UCSF Chimera [130], uses the watershed method to partition a density map into regions [131]. Other methods use level sets [132] and elastic networks [133] for segmentation. The automatic segmentation is however non-trivial and compli-

Table 1.1 Visible features in a density map as a function of resolution

Resolution [Å]	Visible Features
> 8	possible identification of subunits and individual domains
≤ 8	α-helices are visible in form of rod-like densityies and β-sheets as thin continueos planes
< 8	connectivity between secondary structure elements becomes visible
≤ 5	developed features and pitch become visible in α-helices
≤ 4.5	individual strands of β-strands become visible
≤ 4	side-chain densities become recognizable and a relatively unambiguous trace can be seen

With increasing resolution, more features of a density map become visible. The information provided in this table is taken from [129].

cated by nonuniform resolutions throughout the map and tightly intertwined subunits [134].

1.3.2 Fitting of Atomic Structures

If the resolution of a density map is below 8 Å, no secondary structural elements are visible and only subunits and individual domains can be identified. One of the most common methods for analysing these low resolution density maps is the fitting of known atomic structures of individual proteins, usually obtained by X-ray crystallography or NMR, into these density maps to obtain atomistic representations of macromolecular assemblies. In general, one distinguishes between two different types of fitting methods: rigid-body and flexible fitting methods.

Rigid-body fitting methods [135–137], attempt to identify the best agreement between a rigid atomic model and a density map. This is generally achieved via an exhaustive rotational and translational search.

Flexible fitting methods, such as [138–143], consider the fact that proteins are intrinsically flexible by allowing the model to morph, or flexibly fit, into the density map. This approach becomes particularly useful when fitting structures of a different conformation or homologous structures.

1.3.2.1 DireX, a geometry-based algorithm for low-resolution refinement

DireX [138, 144] is a program that flexibly fits and refines protein structures into density maps obtained from X-ray crystallography or EM by efficiently sampling conformations under experimental restraints using a geometry-based sampling algorithm. During each

of its random move steps, which are based on the CONCOORD algorithm [145] that uses a large number of distance restraints computed from the starting structure to maintain correct stereochemistry and to prevent protein overlaps, DireX applies forces to the atoms that drive the structure into the density map. To avoid over-fitting, additional Deformable Elastic Network (DEN) restraints are applied during each step. The DEN approach aims to refine only those degrees of freedom that are defined by the data by combining prior structural information of the reference model (for the degrees of freedom without experimental data) with experimental data. The free map correlation C_{free} [146], can be used to optimise the use of restraints during the refinement. The C_{free} is a cross-validation approach for real space refinement against cryo-EM maps. To do so, the data is split according to frequency ranges and a band of high-frequency data with low SNR is used as test set while the low resolution range with significant SNR is used as training set. By generating a perfectly overfitted model, the correlations between the work and test set can be quantified.

1.3.3 Secondary Structure and Topology Determination

While the information contained in density maps at medium resolution, i.e. between 4 Å and 10 Å, is too low to build models with explicit atomic coordinates, it is possible to determine secondary structural elements and the overall topology of the protein.

1.3.3.1 Identification of Secondary Structural Elements

In medium resolution structures, helices are resolved as straight rods with comparatively high density. Computational detection of α-helices was first implemented in Helixhunter [147], a tool which uses an exhaustive cross-correlation search with a prototypical helix template over three translational and two rotational degrees of freedom. Contrary to α-helices, β-sheets are resolved as thin plates and are much more diverse in both size and shape, making correlation-based approaches impossible. Sheetmeter [148] therefore uses a morphological analysis of the density to identify regions thare are nearly flat, which are later filtered, clustered, and extended to provide a final β-sheet prediction with Sheetracer [149]. In 2007, SSEHunter [150], a single tool for detecting both α-helices and β-sheets, has been developed. The tool pairs an helix correlation routine with local geometry calculations and density skeletonization to detect both types of secondary structure elements.

1.3.3.2 Backbone Tracing

In 2008, Ludtke and Co-workers demonstrated that it is possible to achieve a Cα-trace with correct topology directly from a protein density map at ~4 Å resolution [151].

In their protocol, they combine secondary structural elements, as detected with SSE-Hunter, with a novel density skeletonization technique [152] and sequence-based secondary structure prediction.

Pathwalker [153, 154] is a tool that automatically enumerates putative configurations of protein structure models in sub-nanometer resolution density maps. In an initial step, Pathwalker populates the density map with a set of pseudo-atoms wich are in the next step connected with an approach derived from the Traveling Salesman Problem with the aim to minimise the deviation from the standardised Cα-Cα distance. The tracing is guided by both the distance between two pseudo-atoms, as well as their density connection.

1.3.4 *De-Novo* Model Building

De novo methods attempt to build an atomic model directly from the density map without structural templates such as solved structures from homologs. This is only possible if the resolution of the density map is high enough.

One such *de novo* method is EM-Fold [155, 156]. It uses medium resolution density maps (5-7 Å) as folding constraints for *de-novo* protein modelling. The protocol *de novo* assembles predicted secondary structural elements into the corresponding regions of a density map and subsequently uses Rosetta to build loops and side-chains.

1.3.4.1 *De-Novo* Model Building with Rosetta

In 2015, Wang et al. have described a *de-novo* modeling approach for cryo-EM maps at near-atomic resolution that matches and assembles sequence-based local backbone conformations, so called fragments, into the target density map [157]. As in the Rosetta *de novo* structure prediction [41, 42], the fragments for overlapping windows of amino acid sequence used in their protocol are 9 residues long and are taken from solved protein structures with similar local sequences and predicted secondary structure. Each fragment is subsequently subjected to a 6-dimensional search in the protein density map. To identify presumably correct placements, a mutually compatible subset of placements is identified using Monte Carlo simulated annealing (MCSA). The score function used to guide MCSA favors the following properties of fragment pairs: (1) good fit of the each partner to the density map (2) same residues that are approximately in the same place (3) residues that are close in sequence are nearby in space and (4) no two residues are occupying the same space. Fragment assembly and MCSA are applied iteratively until at least 70 percent of backbone residues have been assigned. Finally, this partial model is

completed through rebuilding and all-atom refinement using RosettaCM [158] guided by the density map.

1.3.4.2 Use of Crystallographic Determination Tools

In case of high-resolution density maps, it might also be possible to use methods that have originally been developed for structure modelling in X-ray crystallography. One of the most commonly used X-ray tools being used to determine the structures of cryo-EM reconstructions to date is COOT [159], a tool for model building and real-space refinement into density maps requiring a large amount of human intervention [134].
Automatic crystallographic structure determination tools, such as *autobuild* of the PHENIX software suite [160], Buccaneer [161], the SOLVE and RESOLVE packages [162], and ARP/wARP [163] have been developed and are widely used for crystallographic datasets of 3 Å or better. While they have been used for crystallographic data as low as 3.8 Å, the results at this resolution are inconsistent [134]. These tools are therefore not ideally suited for cryo-EM density maps in the 3.0 - 4.5 Å resolution range.

2. Materials and Methods

2.1 Method Overview

We have developed a *de-novo* model-building approach for cryo-EM maps at 3 to 4.8 Å resolution. Our approach called EMfasa is illustrated in Figure 2.1 and consists of five steps: (A) generation of a Cα-trace, (B) fitting of sequence non-specific local backbone conformations into the density map and identification of a consistent subsets of these fragments, (C) all-alanine structure generation by assembly of fragments via clustering, (D) automated sequence assignment by rotamer matching, (E) full-atom assembly and refinement.

In an initial step, given a segmented protein density map at 3 to 4.8 Å resolution and the primary sequence, a Cα-trace is generated that roughly describes the location of the residues in the protein and the protein's topology. The Cα-trace is in the next step used to spatially limit the fitting of a sequence non-specific 7-mer fragment library to each residue position in the trace. Subsequently, using a scoring function considering several terms such as the fit of a fragment to the density and the overlap to neighbouring fragments, a subset of nicely matching fragments is selected. This set of fragments is in the next step clustered to obtain a full-atom all-alanine structure. This all-alanine structure is then used to build a profile reflecting the fit of each of the 20 amino acids at each backbone position. The profile is subsequently used to align the protein sequence to the all-alanine structure. Based on this profile-sequence alignment, a full-atom protein structure including side chains is assembled by comparative modelling and the final trace is refined into the experimental density data. All steps are described in more detail in Sections 2.2 - 2.4.

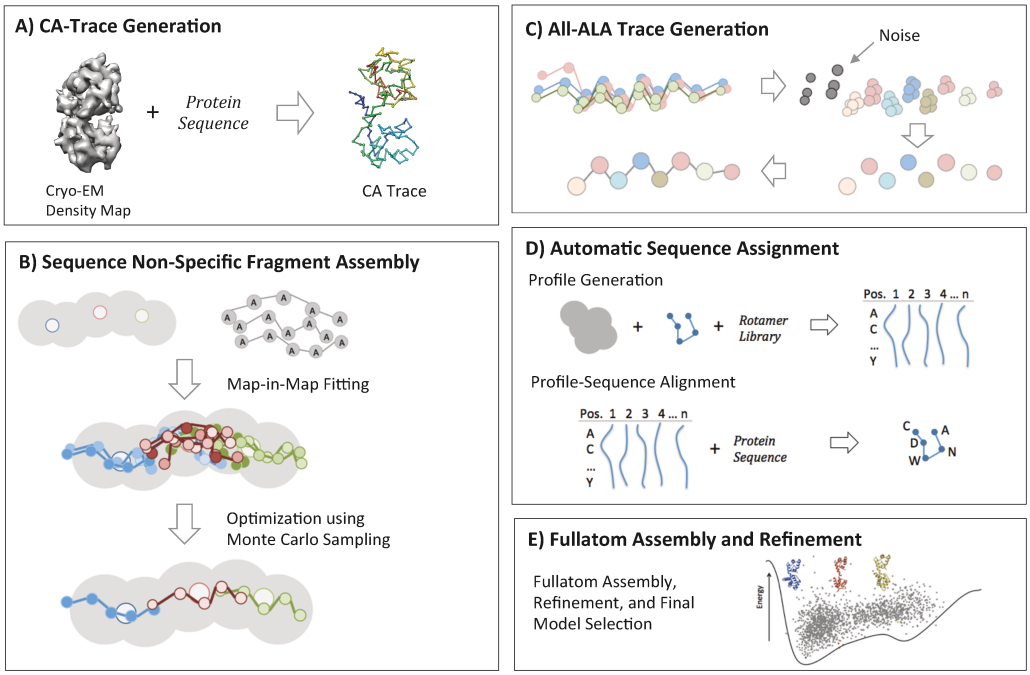

Figure 2.1 EMFasa protocol overview. (A) Generation of an Cα-trace using random bead placement and a combinatiorial optimization algorithm to connect them (B) sequence non-specific fragment fitting and identification of a good matching subset (C) all-alanine structure generation by clustering the good matching subset of fragments (D) automated sequence assignment by rotamer matching (E) fullatom assembly and refinement

2.2 Backbone Generation

Steps (A) to (C) in Figure 2.1 are used to build an all-alanine protein structure with explicit atomic coordinates for the entire backbone and the $C\beta$ of each alanine side chain. To enable a successful sequence assignment in the next step, each residue and its atoms need to be placed as precisely as possible.

2.2.1 Backbone Trace

In an initial step, backbone traces are generated using protocols developed by Zhe Wang and Gunnar Schröder. In a backbone trace, each residue of the protein of interest is represented by a pseudo Cα-atom (in the following referred to as "bead") that describes the residue's rough position in the density map. While specifying the topology of the protein by connecting potentially neighbouring residues, the trace does not contain any sequence information. The protocols and algorithms used to generate a backbone trace are shortly summarised below.

2.2.1.1 Bead Placement

In an initial step, the tool *dxbeadgen* of DireX [138] is used to randomly place twice as many beads as there are residues in the protein chain. This procedure is briefly described in [164]. To make sure that the beads are evenly distributed in the map, the randomly placed beads are subsequently refined with DireX using a low weight on the density map and with repulsive forces between the beads. It was chosen to use twice as many beads as residues in the protein structure to facilitate the connection step described below.

2.2.1.2 Trace Generation

The refined beads are in the next step connected using a protocol based on the combinatorial optimisation heuristic called Lin-Kernighan [165], using the LKH program [166]. The setup for the LKH program is prepared using Pathwalker [153], which, in its newest version, favours connections that lie within strong density regions.
The trace generation is carried out in three steps. Initially, the LKH program is used to generate 10 traces with potentially different connectivities from the refined beads and the density map. In the second step, a consensus Cα-trace is generated by feeding the LKH program an additional cost matrix that is reflecting the connections in the 10 traces that were generated in the step before. In the final step, the number of beads is decreased by a factor of two and the final trace is refined with DireX.

2.2.2 Fragment Assembly

The refined Cα-trace shows how the amino acid sequence propagates through the density, i.e. the protein's topology, and roughly describes the locations of the residues. These locations are however not accurate enough for allowing to directly infer the actual Cα-positions in the protein structure. Therefore, to improve the accuracy of the Cα-positions and to obtain a full-atom protein backbone, we are matching local backbone conformations of known protein structures into the density maps at the bead locations stored in the backbone trace. Other than Wang et al. [157], who are using fragments with similar local sequence and secondary structure, we are using sequence non-specific fragments.

2.2.2.1 Fragment Library

For successful fragment assembly, a library representing the most common backbone conformations in the protein universe is needed. A framework summarising the entire construction process of the fragment library used in our method is shown in Figure 2.2 and will be described in more detail below.

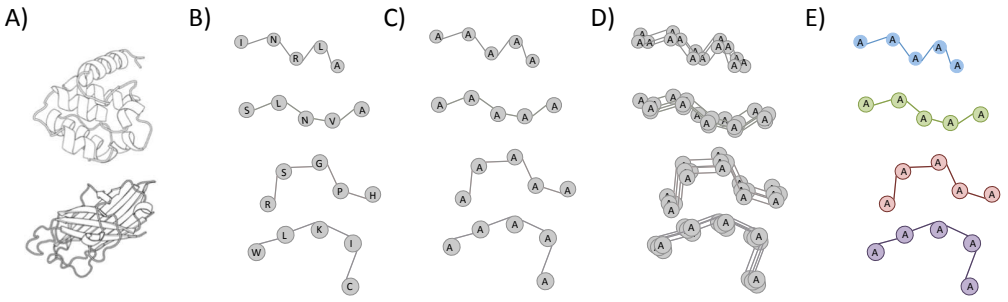

Figure 2.2 Construction of a sequence-nonspecific fragment library. To construct the fragment library, *n*-mer fragments (B) are extracted from a non-redundant set of high-resolution Xray protein structures (A). Subsequently, the residues of each fragment are mutated to alanine and bond angles, bond lengths, and torsion angles are idealized (C). Finally, the fragments are clustered (D) and the centroids of each cluster constitute the final fragment library (E).

Construction

We used the protein sequence culling sever PISCES [167] to construct a set of non-redundant protein structures, consisting of 4718 distinctive protein chains that all were determined by Xray crystallography and have a percentage sequence cutoff of 20 percent, a minimum resolution of 1.8 Å, and a R-factor cutoff of 0.25. The use of high-resolution X-ray structures ensures that only well resolved fragments are used to build the final fragment library. To build the fragment library, several thousands of 7-mer fragments (user specified number of continuous subsets of 7 amino acids) are extracted from the non-redundant set of protein structures described above. These extracted fragments are in the next step mutated to alanine with Scwrl4 [168] to obtain sequence-nonspecific fragments. Bond lengths, bond angles, and torsion angles of the fragments are idealised using Rosetta. In the final step, the idealised all-alanine fragments are clustered into X groups using ClusCo [169]. ClusCo offers four different clustering algorithms: k-means, single-, maximum-, and average-linkage. The core fragments of the final clusters describe the final fragment library. In the course of this thesis, if not stated otherwise, fragment libraries of size 100 are used. The fragment library has to be generated only once - the same library can be used to model any target protein structure.

2.2.2.2 Fragment Fitting

The fitting of the individual fragments into the density map is carried out using UCSF Chimera's [130] *fitmap* procedure. The fitting procedure is illustrated for one fragment and one potential backbone position in Figure 2.3. In an initial step, the input density map is cropped to a cube with an edge length of 20 Å around the bead of interest to ensure that the fragment will be fitted in close proximity to the latter. In the next step, a fragment

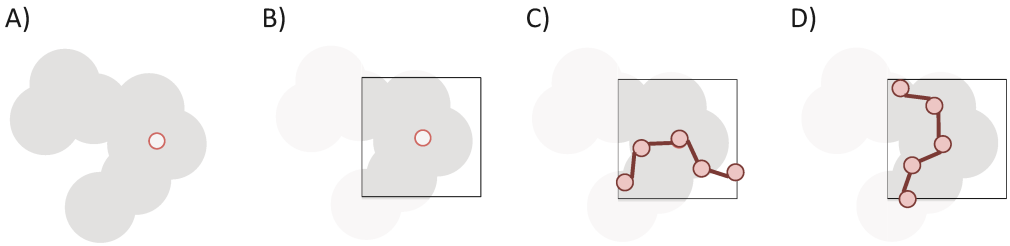

Figure 2.3 Fragment fitting procedure. The fragment fitting is carried out in several steps: a bead from the backbone trace is selected (A), the density map is cropped to a cube centering around the selected bead to enable a local fitting (B), a fragment from the fragment library is centered on the bead position (C), and finally fitted to the density using Chimera's $fitmap$ procedure (D).

from the fragment library is placed in the center of the cube (by positioning its central Cα on the bead) and 30 global search operations with the Chimera *fitmap* command are carried out. For each search operation, the Chimera fitting procedure generates a random initial placement of the fragment within 1 Å of its starting position and follows it by a local map-in-map optimisation. To do so, the Chimera procedure automatically generates a map with the user-specified resolution from the coordinates of the fragment by describing each atom as a Gaussian distribution of width proportional to the defined resolution and amplitude proportional to the atomic number. As metric for the map-in-map fitting, correlation about zero was chosen. For each fragment, the 5 top-scoring placements are stored, which, assuming a fragment library of size 100, results in 500 fragment placements per backbone position.

2.2.2.3 Fragment-Bead Assignment

After a completed fitting procedure, 500 fragment placements per backbone position have been stored. While having constrained the fitting procedure to a local optimization by using cropped density maps, a fragment that has been fitted at a certain bead position might end up closer to one of the neighbouring beads. The cropped density maps, i.e. cubes with an edge length of 20 Å, are large enough to not restrain the fragments to the corresponding beads, but allow sufficient space to also explore the neighbouring areas. Therefore, before being able to identify a set of mutually compatible fragments, the fragments need to be reassigned to the closest bead positions, as illustrated in Figure 2.4. Beads without any assigned fragments, will be excluded from the subsequent steps: no fragments have been placed in close proximity, indicating that the bead has been positioned outside of the density.

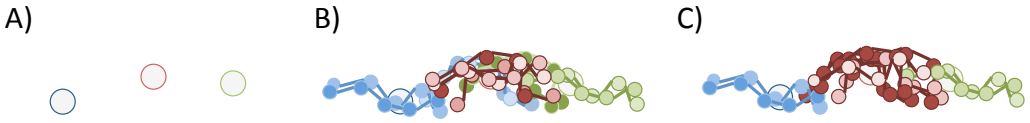

Figure 2.4 Fragment-bead assignment. For each bead (A), fragments are fitted and the top-scoring ones are stored (B). The placed fragments are in the next step reassigned to the beads closest to them (C) .

2.2.2.4 Monte Carlo Simulated Annealing

To find the set of fragments that are mutually compatible, Monte Carlo simulated annealing (MCSA) sampling, a method initially proposed by Kirkpatrick et al. in 1983 [170], is carried out. The applied MCSA procedure is illustrated in Figure 2.5. To reduce the search space, only a user-specified number X of top-scoring fragment placements per bead are considered. At the beginning of the procedure, each bead position is randomly assigned to one of its X top-scoring fragment placements. During each step of the MCSA, the assigned fragment at a random bead position is exchanged and the resulting score difference ΔE is calculated. The move is then either accepted or rejected based on the Metropolis criterion [78]: if $\Delta E < 0$, the move gets immediately accepted, otherwise the move is accepted with probability $P = e^{-\Delta E/T}$ depending on the energy difference and the current temperature T of the system. At large temperatures, uphill moves, i.e. conformations that lead to an increase in energy, are allowed to avoid getting trapped in local minima and thereby exploring a large area of the search space [171]. By gradually lowering the temperature, the sampling is guided towards the minimum of the scoring function. To address the fact that no good fragments might have been found at a position, "zero fragments" can be allowed as well. In that case, no fragment is assigned to the respective bead and its two neighbouring beads become neighbours themselves.

Evaluation of Fragment Compatibility

To determine whether a set of fragments is "better" than another, a scoring function is needed. Here, we evaluate the compatibility of a set of fragments with a scoring function consisting of 4 different terms:

$$\begin{aligned} score_{total} = {} & w_{corr}\ score_{corr} \\ & + w_{overlap}\ score_{overlap} \\ & + w_{clash}\ score_{clash} \\ & + w_{dir}\ score_{dir} \end{aligned} \tag{2.1}$$

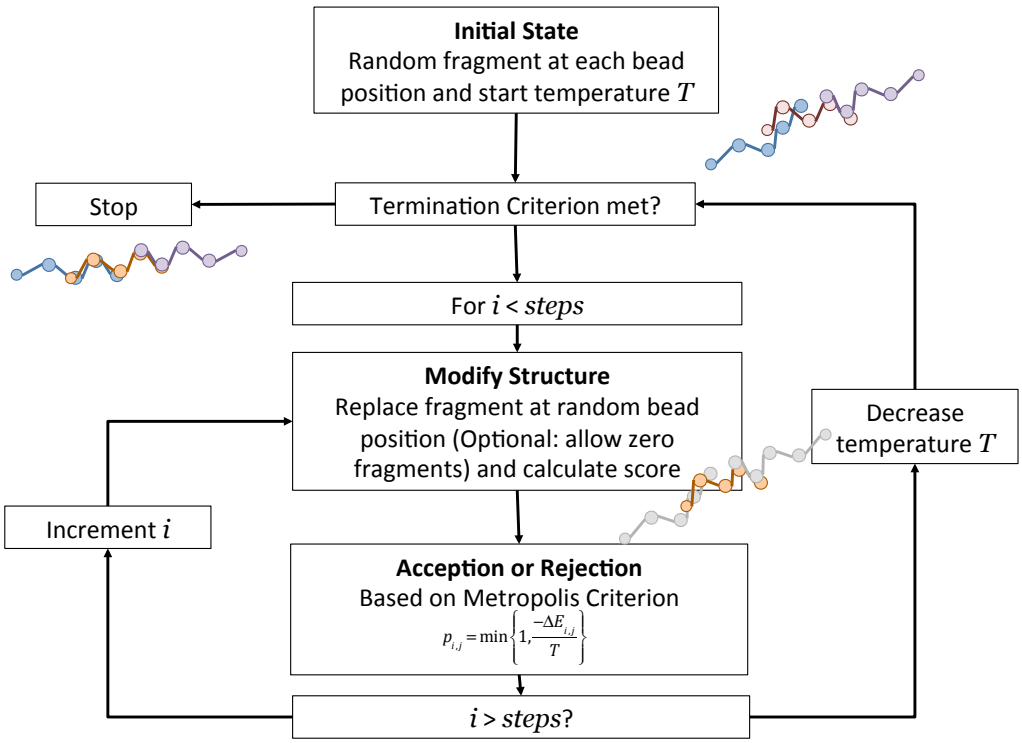

Figure 2.5 Monte Carlo simulated annealing. MCSA workflow to identify a set of mutually compatible fragments.

The term $score_{corr}$ reflects the fit of a fragment to the density and is based on the correlation between the generated density map for the coordinates of the fragment atoms and the density map, used as metric during the fitting procedure with UCSF Chimera, see Section 2.2.2.2. $Score_{overlap}$ describes how well two neighbouring fragments overlap, i.e. describe the same atoms. It is defined as the minimum RMSD over at least 2 Cα-atoms between the two fragments. All possible overlap combinations have to be evaluated, as, without any sequence information, it is not known how far the fragments are overlapping. $Score_{clash}$ evaluates whether there is contact (minimum distance < 2Å) between two fragments that are too far apart in sequence, i.e. assigned to beads that are at least 8 positions apart from each other. This score term is of special interest in areas of β-sheets. Finally $score_{dir}$ evaluates whether the fragment is oriented in the same way as the provided backbone trace or its neighbouring fragments by calculating the dot product of the vectors spanning the specific fragments or the backbone at the respective location. A visual interpretation of each score term is shown in Figure 2.6. The weight of each score term can be specified by the user.

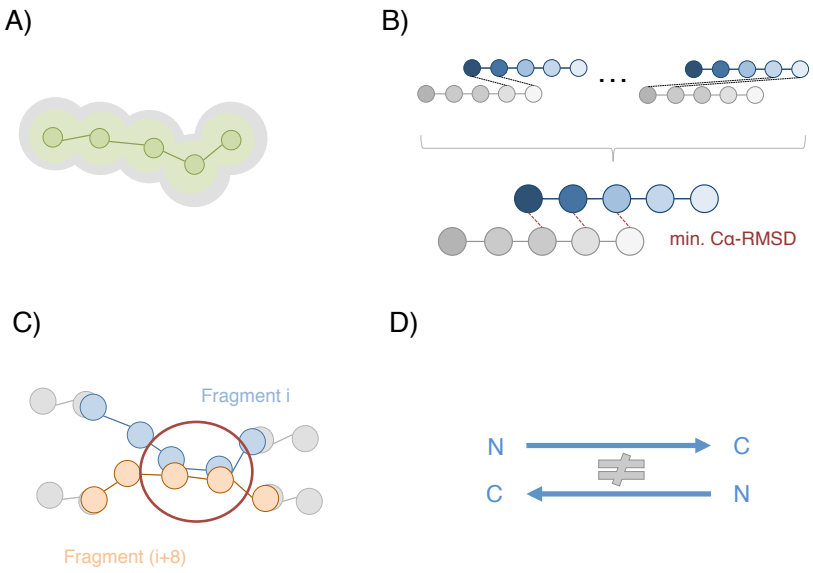

Figure 2.6 Score terms for evaluation of fragment compatibility. 4 different score terms are used to evaluate the fragment compatibility: (A) $score_{corr}$, (B) $score_{overlap}$, (C) $score_{clash}$, and (D) $score_{dir}$.

2.2.3 Consensus Trace Generation

After having found a set of mutually compatible fragments, these need to be assembled into a consensus trace. To determine which fragment positions are placed at identical residue positions, a density-based clustering of all fragment Cα-positions is carried out. The clusters, representing a residue in the final consensus trace, are in the next step connected based on the intra-fragment connections of the cluster members. An overview of the clustering and connection steps is shown in Figure 2.7 A).

2.2.3.1 Clustering

After the MCSA, the trace consists of fragments with different compatibility to each of their neighbours: while there are stretches of very good compatibility, usually situated in areas of well resolved density, the fragments in regions of low resolution are much more scattered. While we know how many residues are in the protein of interest, we do not know a priori how many of these residues could be identified using the fragment assembly. We therefore need to use a clustering algorithm that does not require the number of final clusters a priori.

The density-based clustering algorithm DBScan (Density-based spatial clustering of applications with noise) [172] groups points that are closely packed together and classifies

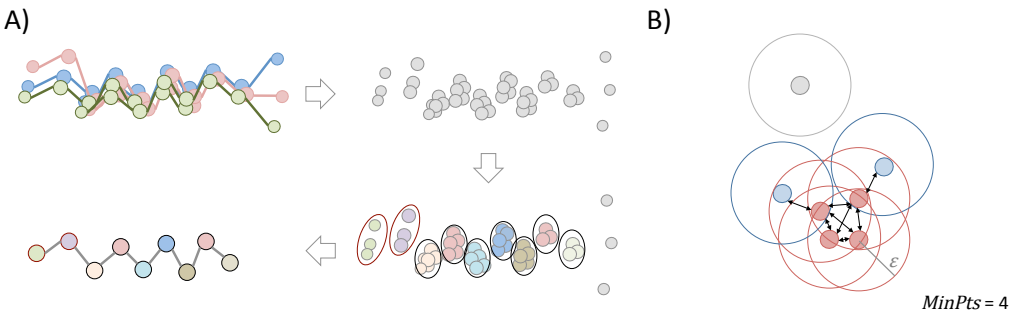

Figure 2.7 Consensus trace generation. (A) To generate a consensus trace, the Cα-positions of all fragments are iteratively clustered using the DBScan algorithm. The formed clusters are subsequently connected based on the intra-fragment connections of the cluster members. (B) The DBScan algorithm classifies each point into one of three groups: *core points* (red) have at least a minimum number ($MinPts$) of points in their ε-neighbourhood. *Border points* (blue) are in the ε-neighbourhood of a core point but do have less than $MinPts$ points in their ε-neighbourhood. All other points are classified as *noise* (grey).

points in low-density regions as noise. The algorithm requires two parameters: ε (eps) which defines the maximum distance between two points for them to be considered in the same neighbourhood, and the number of total samples in a neighbourhood ($minPts$) required to form a dense region. As shown in Figure 2.7 B), each point is classified into one of three groups: core points have at least $minPts$ points in their ε-neighborhood, border points are in the ε-neighborhood of a core point but have less than $MinPts$ points in their ε-neighborhood and therefore form the edge of the cluster. All other points are classified as noise or outliers.

In this case, the choice of both parameters is quite straightforward: The standard distance between two Cα atoms is 3.8 Å, the distance for density reachable points, ε, should therefore not exceed 2 Å so that two neighboring Cα positions will not end up in the same cluster but stay separate. The minimum number of points that is required to form a dense region is dependent on the number of fragmented traces that are used for clustering. Assuming that at least 2 fragments are overlapping, we define $minPts$ as $2x$ the number of input traces.

The clustering is carried out in two iterations: In the first round, very well defined Cα-positions are identified using DBScan with an ε-value of 1 Å. In the second round, ε is increased to 2 Å and used to cluster all the Cαs that are not part of any cluster formed in the first iteration. After each iteration process, two clusters that are closer than 2.5 Å are merged into one.

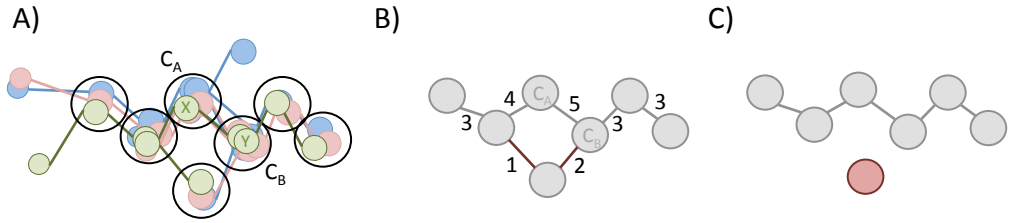

Figure 2.8 Cluster connectivity. The cluster connectivity is determined in several steps: (A) The clusters are connected based on the intra-fragment connection of their cluster members. (B) The number of connections are stored as the corresponding connection weight, and for each cluster, only the two most populated connections are kept. (C) To remove circles and branch-offs, lowest-weight connections are removed, until no cluster has more than two connections. Isolated clusters are removed.

2.2.3.2 Cluster Connectivity

The clustering step returns clusters representing potential residues in the protein of interest but does not provide any information about how these residues are connected. To connect the clusters, the intra-fragment connections of the cluster members are investigated: if Cα X in cluster C_A is connected to Cα Y in cluster C_B in the original fragment, a connection between clusters C_A and C_B is stored (Figure 2.8 A). This is done for all Cαs in each cluster, and in the final step, the two most populated connections for each cluster are kept (Figure 2.8 B). To remove circles and to avoid unwanted branch-offs, connections with the lowest weight are removed from each cluster with more than two connections until only two are left (Figure 2.8 C). Clusters without any connections will be removed and not be considered in subsequent steps.

2.2.3.3 Assembly of Final Consensus All-Alanine Structure

After the cluster connection step, the all-alanine structure still contains several gaps in regions where no nicely overlapping fragments have been found. To decrease the size and also the total number of gaps, several independent MCSA runs and cluster assembly steps are carried out. The resulting, potentially different, fragmented all-alanine structures are assembled to the final all-alanine structure as illustrated in Figure 2.9 A). To do so, DBscan is used with *minPts* set to 1, to keep all of the clusters stored in the fragmented structures and to make sure that no potential residue is lost. ε is set to 1 Å to group the clusters of the different traces occupying the same space. The resulting clusters are connected as described in the section before. To make sure that the topology of our generated consensus trace is correct, the trace is compared to the input Cα trace and wrong connections are removed. Remaining gaps in the trace are removed by connecting

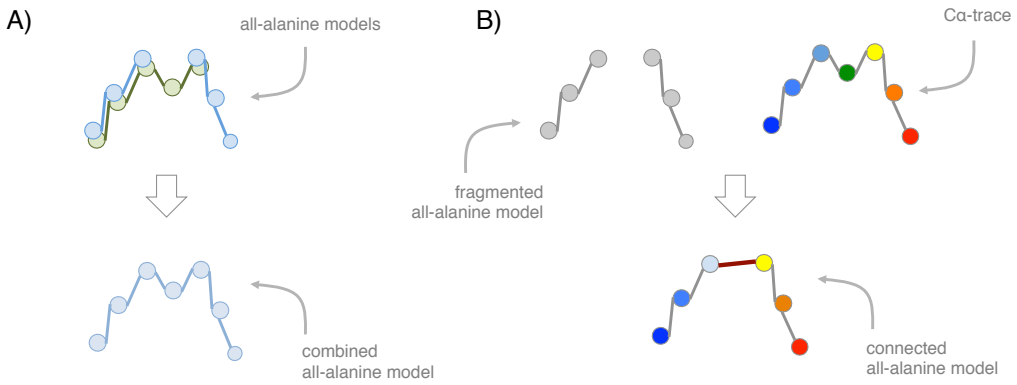

Figure 2.9 All-alanine model assembly. (A) Different all-alaine models are merged to a combined all-alanine model using DBScan. (B) Fragmented all-alanine models get connected by comaprison to the starting Cα-trace.

the corresponding edges, see Figure 2.9 B).
The coordinates of the atoms in this completed all-alanine structure represent the averaged coordinates of all the residues forming the respective clusters. This averaging can result in quite unrealistic local geometries and potential clashes. To solve this issue, a rapid real-space refinement of the all-alanine structure to the density is carried out using the *phenix.real-space-refinement* tool [173] of PHENIX [160].

2.3 Automatic Side-Chain Assignment

The section above described how a complete all-alanine structure is generated. For being able to build the final model of our protein of interest, the side-chains need to be assigned to the positions in the backbone. As shown in Figure 2.1 D), the automatic side-chain assignment is carried out in two steps: Initially, a profile reflecting the fit of each amino acid at each position in the all-alanine structure is generated. This profile is in the next step aligned to the protein sequence by using a dynamic programming algorithm, resulting in the final assignment.

2.3.1 Construction of Position-Specific Profile

The position-specific profile for the all-alanine structure is based on the fit of each amino-acid to the density at each of its residue positions. To determine the fit of each amino acid, a rotamer library is used and the correlation of each rotamer to the input density is calculated. These correlation values are in the next step used to build the profile.

2.3.1.1 Profile Generation

As rotamer library, the user can either choose the Dunbrack backbone-dependent rotamer library [174] or common-atom values from the Richardson backbone-independent rotamer library [175]. Both rotamer libraries are included in UCSF Chimera [130]. By using Chimeras *swapaa* command, each rotamer of the specified rotamer library is tested and the best matching rotamer for each of the 20 amino-acids is selected at each atom position. The best matching rotamer is either specified by maximising the values of the interpolated map at each atom position of the rotamer or by the best fit as determined by Chimera's *swapaa* command. The latter chooses the best fit based on three methods in the following specified order: the lowest clash score, best fit into density and the highest probability according to the rotamer library. Each successive method is only used if the previous one resulted in a tie.
In the next step, the residue-density compatibility for each of the selected rotamers is calculated. This is done by zoning the density map around either the atoms of the 20 best-fitting rotamers or the best-fitting rotamer of the amino acid that is currently evaluated and by calculating the correlation between this zoned map and the calculated map of the rotamer. This results in a matrix that lists a correlation value for each amino-acid at each position in the consensus backbone trace. These values are in the next step transformed to standard (Z) scores using the following equation

$$Z_{aa,i} = \frac{X_{aa,i} - M_{aa}}{SD_{aa}} \tag{2.2}$$

where $X_{aa,i}$ is the correlation of amino acid aa at backbone position i, and M_{aa} and SD_{aa} the mean and standard deviation of the correlation of that amino acid at each position in the backbone, respecively. The final profile P is calculated as follows:

$$P_{aa,i} = e^{\frac{3}{2}Z_{aa,i}} \tag{2.3}$$

2.3.2 Profile-Sequence Alignment

To align the protein sequence to the profile, a dynamic programming algorithm is used. As we are interested in a global alignment, we use the Needleman-Wunsch algorithm [176] with affine gap costs, as described by Gotoh [177]. The backtracking routine has been implemented according to [178]. The score of aligning residue aa to position i of the protein backbone is stored in the sequence profile $P_{aa,i}$. Affine gap penalties are described in the form $g_{gapOpen}+l*g_{gapExtend}$ where $g_{gapOpen}$ refers to the cost required to introduce a new gap and $g_{gapExtend}$ describes the cost to extend the length of an existing gap by 1. The gap costs can be specified by the user.

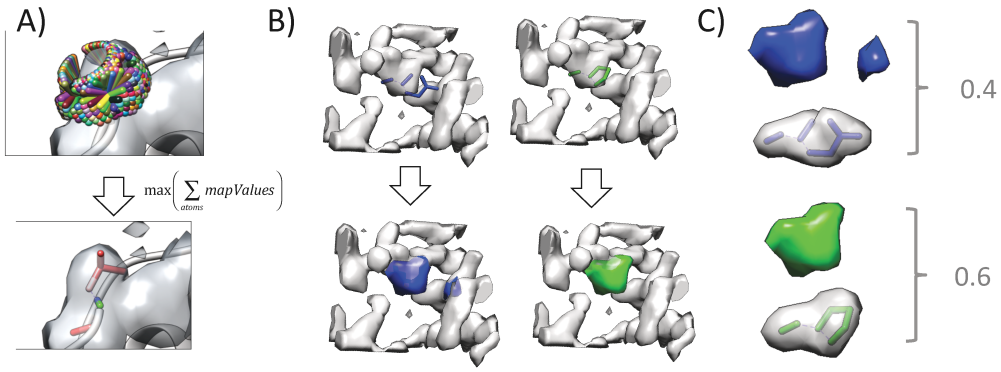

Figure 2.10 Rotamer selection and correlation calculation. (A) The best matching rotamer of each amino-acid is calculated by maximizing the values of the interpolated map at each position. (B) The input density map is zoned around the atoms of the current rotamer. (C) The correlation between the zoned input density map and the rotamer is calculated.

2.3.3 Identification of the correct N-C termini

For structures whose N- and C- termini are close to each other in 3D space, automatic and correct detection during the Cα-tracing and later steps is difficult and therefore might be incorrect. EMfasa can use a profile, generated as described before, to generate alignments of all possible sequence iterations (1-N, 2-N+1, ... N - N+1). The best-scoring alignment can subsequently be used to adapt and correct the termini of the all-alanine structure.

2.4 Final Model Assembly and Refinement

The profile-sequence alignment, obtained as described before, is in the next step used to build the final model of our protein as illustrated in Figure 2.11. To do so, the *automodel*

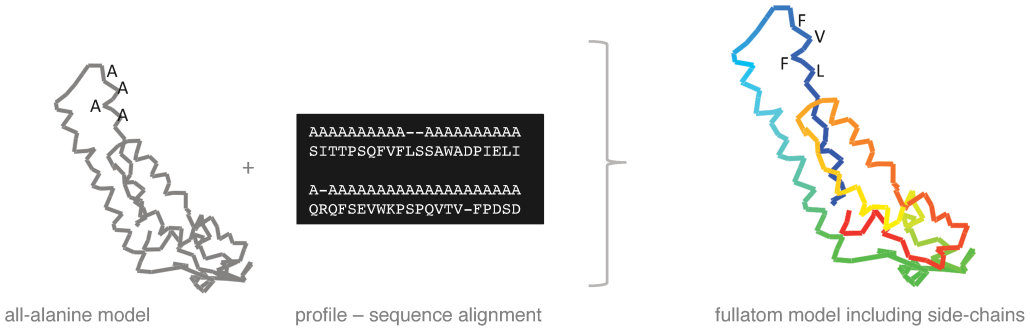

Figure 2.11 Final model assembly with Modeller. The final model, including side-chains, is assembled with Modeller by building a homology model using the profile-sequence alignment and the all-alanine structure as template structure.

class of Modeller [179] is used with additional absolute position restraints on each Cα of the all-alanine structure with a standard deviation of 2 Å. The *automodel* class automatically builds the 3D model of the target based on the target-template alignment and the template profile-sequence alignment[180]. In this case, the all-alanine structure is used as template and the profile-sequence alignment represents the target-template alignment. The absolute position restraints on the Cα-atoms are added to keep the aligned residues in place and thereby make sure that the majority of the model stays within the density.
During the comparative modelling with Modeller, no density information is used for which reason it can easily happen that the newly generated residues are not situated in the density map. Therefore, the model is in the next step rapidly refined to the density using the real space refinement of PHENIX [160] consisting of global minimization, local rotamer fitting, morphing and simulated annealing [173] .

2.5 Pool Generation and Selection of Final Models

A pool of protein structures is generated by using u topologically different Cα- traces in both directions[1] (A), combining several different MCSA trajectories obtained for each Cα-trace to v different all-alanine structures (B), calculating w different alignments by using varying gap penalties for each all-alanine structure (C), and generating x real space refinements for each model assembled based on one alignment (D). This results in a total pool of $2 * u * v * w * x$ different protein structures. The largest variability between the resulting models is obtained by using steps (A) through (C).
To evaluate the final models, we are currently using the correlation to the density map CC_{map} that has been calculated during the PHENIX real space refinement. In future, it might be of interest to use other scoring methods to improve the selection of the models closest to the native structure.

2.6 Metrics used for Evaluation

We have used several different metrics to evaluate the models obtained by EMfasa. A short description of each of these metrics and their application purposes are described below.

1) *Map Correlation* CC_{map}
CC_{map} is a real-space correlation coefficient that measures the similarity between a density map calculated from the protein model and the experimental density map across

[1] In absence of any additional information, models for both directions have to be built because the Cα trace itself does not provide any information about its N-C orientation.

the whole map volume as calculated by PHENIX. The real-space correlation coefficient (RSCC) between a calculated density map (ρ_{calc}) and the experimental density map (ρ_{obs}) is calculated as follows [181]:

$$RSCC = corr(\rho_{obs}, \rho_{calc}) = \frac{cov(\rho_{obs}, \rho_{calc})}{\sqrt{var(\rho_{obs})var(\rho_{calc})}} \tag{2.4}$$

2) *Root-Mean-Square-Deviation (RMSD)*
The root-mean-square deviation of atomic positions (RMSD) is the most commonly used measure of the similarity between the atoms of superimposed proteins [180]. The RMSD for N superimposed atoms is calculated as follows

$$RMSD = \sqrt{\frac{1}{N}\sum_{n=1}^{N} d_i^2} \tag{2.5}$$

with d representing the distance between the superimposed atoms at position i. Due to this formula, the RMSD is very sensitive to local structure variations. In this work, the RMSD has been only calculated on the Cα-atoms, also referred to as Cα-RMSD. Depending on what models were analyzed, the RMSD was either calculated for the entire structure or only for an aligned subset, cf. below.

3) *TMscore and TMalign*
The TMscore [182] measures the structural similarity of two proteins based on the global fold similarity in a sequence-order dependent way. It returns a value between 0 and 1, where 1 indicates a perfect match between to structures while values close to zero correspond to unrelated proteins. The TMscore is defined as follows

$$TMscore = Max\left[\frac{1}{L_N}\sum_{i=1}^{L_r}\frac{1}{1+(\frac{d_i}{d_0})^2}\right] \tag{2.6}$$

with L_n being the length of the native protein structure, L_r the length of the template protein structure, d_i the distance between the i-th atom pair and d_0 a scaling factor based on the average distance of corresponding residue pairs of random related proteins.

The TM-align algorithm [183] can be used for protein structures of unknown equivalence and different length. The algorithm initially generates an optimal superposition of the two structures using dynamic programming iterations and subsequently calculates the TMscore based on the alignment.
We have used TM-align to generate the alignments between our Cα-traces and all-alanine structures and the native models. For both model types, it is possible that some ar-

eas are overpopulated with residues, while some residues in noisy density regions might have been missed. It is therefore necessary to calculate an optimal alignment between both structures before calculating the TMscore or RMSD. TM-align normalizes the corresponding TMscores by the length of the respective native protein structure and calculates RMSD values on the aligned residues.

2.7 Implementation and Method Availability

The protocols for generating the Cα-traces have been developed by Zhe Wang and Gunnar Schröder and are available as part of DireX. The other parts, i.e. steps B) - E) of the EMfasa *de novo* modelling tool, have been developed as part of my PhD project and are implemented in form of a python package. EMfasa and a detailed tutorial describing all modelling steps and corresponding parameters can be downloaded at `www.simtk.org/projects/emfasa`.
Some EMfasa functions rely on external software. These programs, e.g. UCSF Chimera, Modeller, ClusCo, and PHENIX are not part of EMfasa and have to be installed separately for full functionality of the individual steps.

3. Results and Discussion

3.1 Protocol Discussion

This section illustrates interim results of the individual EMfasa steps. The section focuses on steps B (sequence non-specific fragment assembly) through E (fullatom assembly and refinement) of the protocol, as these are the ones that were developed throughout the course of this thesis.

3.1.1 Fragment Library

The performance of fragment-based structure prediction is usually highly dependent on the quality and nature of the underlying fragment library. Besides varying input structures and/or a different fragment representation, the largest differences in fragment libraries are credited to the actual fragment length and the number of representative fragments in the library. Given an identical set of input structures, one may therefore construct a large amount of different fragment libraries by simply varying these two parameters. The resulting libraries will differ in both quality and complexity and it is important to find a good interplay between both for the problem of interest [184]. In addition to these two parameters, the fragment library generated for EMfasa is also dependent on the clustering algorithm used.

To assess the quality of our fragment library, two metrics, based on the ones described in [185], were used: precision and coverage. For a protein of interest, precision is defined as the number of good matching fragments divided by the total number of fragments in the fragment library, while coverage describes how many residues are represented by at least one good match in the fragment library. As done in [185], several cutoffs in the range of 0.1 to 2 Å were used to distinguish between a good and a bad match. In this case, however, only the Cα-RMSD is evaluated. The validation data set comprising 41 proteins from different SCOP classes described in [185] can be found in Table D.1.

The results of the validation are shown in Figure 3.1. Subfigures A) and B) show the differences in the fragment libraries resulting from the use of four different clustering algorithms: the k-means algorithm and 3 hierarchical clustering algorithms, namely single-linkage, maximum-linkage and average-linkage. The Kmeans_100 library (naming

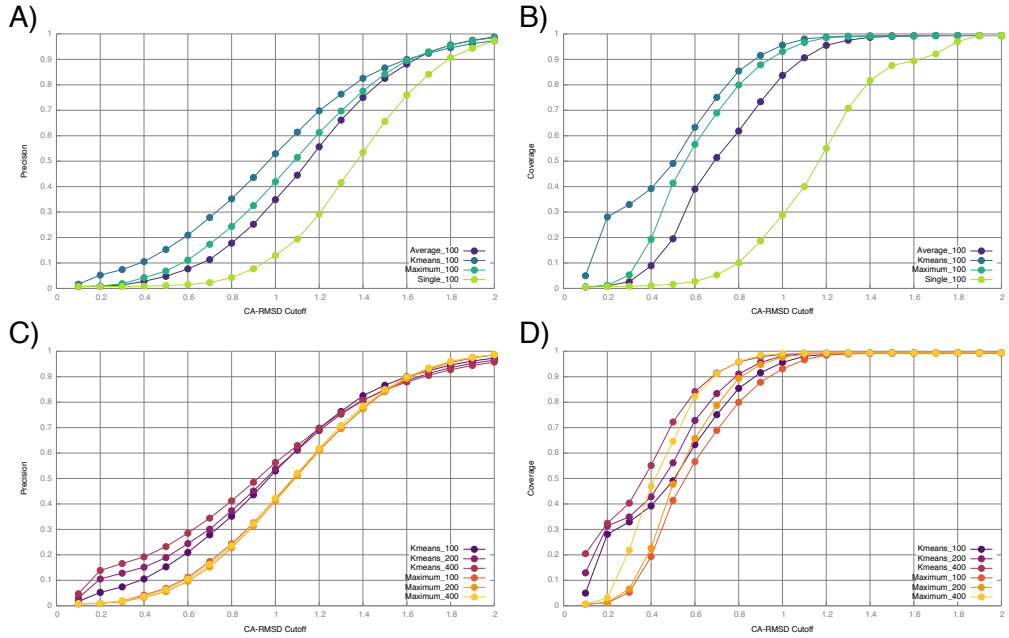

Figure 3.1 Precision and coverage for fragments of length 7. All fragment libraries evaluated in this plot were obtained by clustering 5000 fragments and calculated for several Cα-RMSD cutoffs. The precision and coverage values of all 4 plots were averaged over the 41 proteins of the validation set. A) and B) show precision and coverage for fragment libraries of size 100, generated with 4 different clustering algorithms. C) and D) show the results for fragment libraries of 3 different sizes (100, 200, and 400) for the fragment libraries obtained with k-means and maximum-linkage clustering.

scheme for fragment libraries: X_Y with X representing the clustering algorithm and Y the number of elements in the fragment library) performs best in terms of precision and coverage. The performances for Maximum_100 and Average_100 are similar, the performance of Single_100 however seems to be significantly worse. This can be explained by having a closer look at the fragments in the library. The K-means clustering returned a large amount of helix-fragments which explains the good coverage at RMSD-Cutoffs $<$ 0.5 Å. A closer examination of the single-linkage clustering showed that the clustering generated one extremely large cluster (containing more than 95 percent of all fragments), while the remaining 99 clusters only contain one to five fragments. The single-linkage clustering hence represents the majority of fragments by a single centroid fragment and is therefore not able to represent the most common backbone conformations.

Subfigures B) and C) show the influence of the library size on the quality. It is not surprising that coverage and precision are increasing with increasing library size. The more elements are part of the library, the more local backbone conformations are captured. The increase in precision and accuracy however does not seem to be very large and we therefore focused on fragment libraries of size 100 for modelling protein structures in the

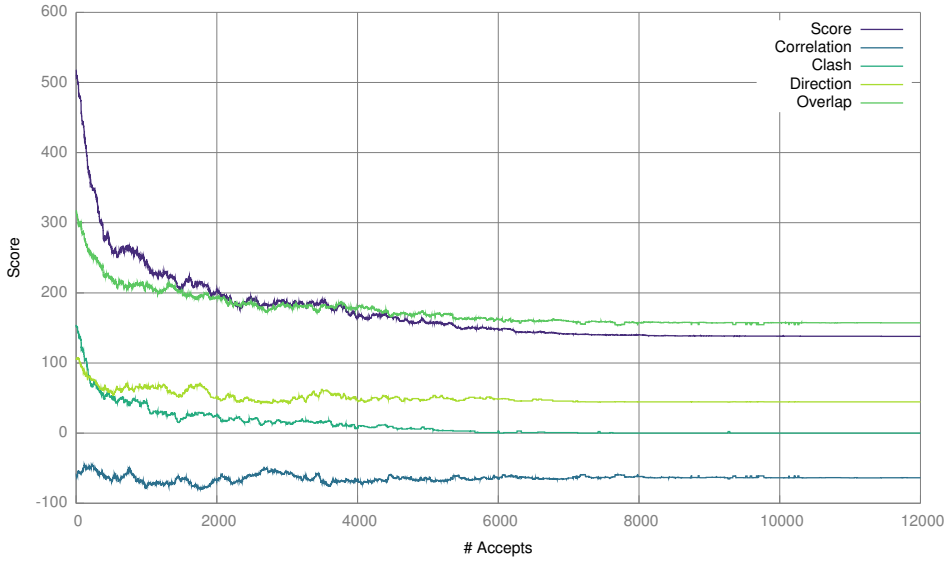

Figure 3.2 MCSA energy minimization. The plot shows the values for the individual score terms (correlation, clash, direction, overlap) and the overal score for each accepted structure during an MCSA run with 2000 iteration steps per temperature cycle for BMV.

course of this thesis.

3.1.2 MCSA Fragment Sampling

MCSA sampling is carried out as described in Section 2.2.2.4 to obtain a trace with nicely overlapping fragments. The plot in Figure 3.2 shows the individual score terms and the overall score for each accepted structure during an MCSA run for the 3.8 Å reconstruction of the Brome Mosaic Virus (BMV) [186]. During the MCSA sampling, only the 30 top-scoring fragments per bead position were considered to restrict the search space to fragments with good agreement to the density. It can be seen that the overall score is initially highly dominated by the clash and overlap score terms. This can be explained by the large amount of β-sheet content in the structure of BMV. Figure 3.3 shows the random starting set and the final set of fragments after the MCSA run. The starting set contains a lot of fragments that are crossing the sheet area, resulting in clashes and bad overlaps with its neighbouring fragments, i.e. fragments that are assigned to the neighbouring bead. In the final set however, the majority of the cross-sheet fragments was exchanged to extended sheet fragments which lowered both score terms by reducing the number of clashes and by improving the overlap to neighbouring fragments. The correlation score term stays relatively constant. This is due to the fact, that out of the 500 placements that were stored per bead position, only the 30 top-scoring ones were considered. This score term therefore is mainly responsible to filter the initial large set of fragments and thereby reducing the total search space.

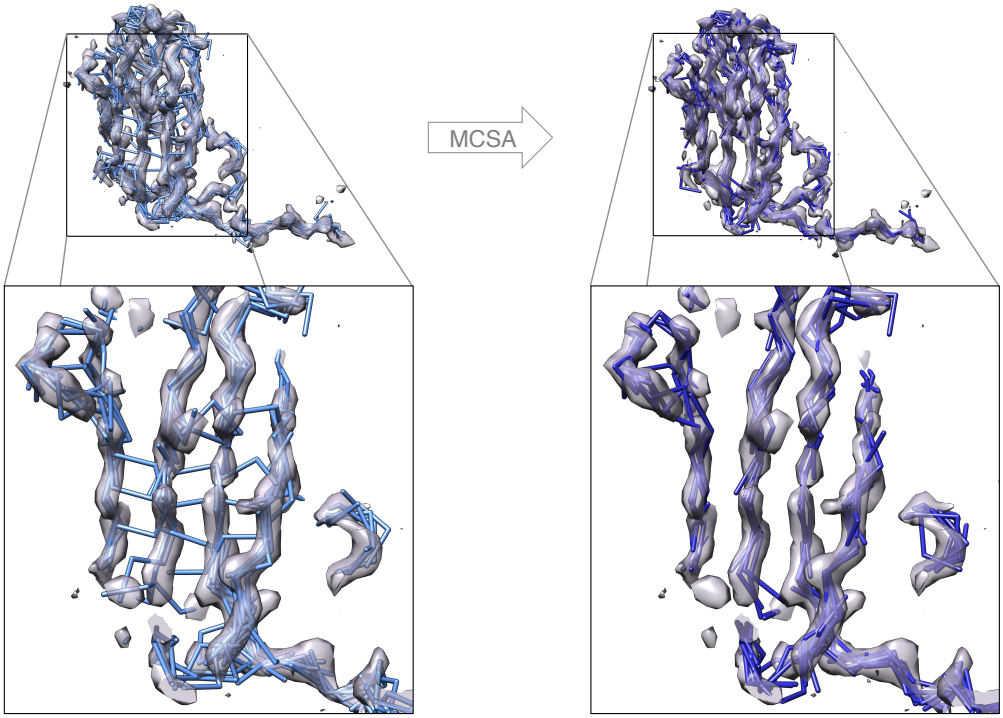

Figure 3.3 Fragment compatibility before and after MCSA. On the left, the random start configuration of an MCSA run for BMV is shown. During each step of the MCSA, one fragment either is randomly exchanged by one of the 30 possible fragments at that position or is removed, i.e. replaced by a "zero" fragment. The final configuration after the MCSA is shown on the right.

3.1.3 All-Alanine Structure Generation

Once a set of compatible fragments has been determined using MCSA sampling, their Cα-positions are clustered using DBScan as described in Section 2.2.3.1. Figure 3.4 A) shows the clusters found in a set of compatible fragments for BMV. The clustering has been carried out in two iterations: the majority of residues were clustered in an initial step. The clustered regions consist amongst others of the well defined sheet areas. The second clustering iteration was able to identify a few more clusters in more diverse regions at the outer areas of the protein. The comparison of the cluster locations, i.e. the averaged Cα-positions of the cluster members, and the native ribbon shows a remarkable overlap. This suggests, that our fragment library with 100 elements is good enough - while the 7-mers might not perfectly represent the stretches of the native protein (and therefore returning comparatively low precision and coverage, see Figure 3.1), it is sufficient, if a large number of residues from different fragments are fitted at the right location.
The found clusters are in the next step connected based on the intra-fragment connec-

tions of their participating residues, cf. Section 2.2.3.2 on page 44. Figure 3.4 B) shows the connections found based on the fragments and clusters that are illustrated in Figure 3.4 A). At this point, the structure consists of several all-alanine fragments of different length. The comparison to the native ribbon shows however, that the majority of found connections is in accordance with the native structure.
To reduce the number and size of the gaps in this fragmented all-alanine structure, several of them (as obtained by several independent MCSA runs with slightly different outcomes) can be combined. Figure 3.4 C) shows five fragmented all-alanine structures. While they are pretty similar in the sheet region of the protein core, they show some diversity in the outer regions and loop areas. The second column of Figure 3.4 C) shows the combined and refined all-alanine structure, generated as described in Section 2.2.3.3, in comparison to the native protein structure. While both structures are quite similar, it must be noted, that the all-alanine structure not necessarily has the same amount of residues as the native structure. In areas of low resolution of loop regions, individual residues might have gone lost.
The clustering is carried out on the Cα-atoms only. The remaining atoms of the all-alanine fragments are however carried along with their corresponding Cα-atom, leading to a full-atom all-alanine structure, see Figure 3.4 D). The overlap with the native structure reveals an almost perfect overlap in the β-sheets in the core region.

3.1.4 Sequence Assignment and Model Selection

To test how well the generated profile is able to identify the native residues in the experimental density map, we have built a profile for the native protein backbone. For BMV, in less than 10 percent of the cases, the native residue resulted in the best score. We are therefore clearly not able to assign the correct residues based on the residues with the highest score in the profile. We therefore treat the sequence assignment as a global problem and align the protein sequence to the backbone using the Needleman-Wunsch algorithm with affine gap costs. Based on the profile generated for one all-alanine structure, we usually generate 21 different alignments. The alignments differ in the applied gap costs that range from -2 to -12 for gapOpen and gapExtend with a step size of -2. Figure 3.5 shows the results based on one all-alanine structure and profile for BMV. In this case, several different alignment settings resulted in the same alignment and full-atom model. The red structure in Figure 3.5 was generated using very cheap gap penalties. The alignment therefore contained large gaps in both the sequence and profile resulting in a structure that is very different from the native one. Making gaps more expensive (cf. blue and green structures) resulted in models closer to the native structure.

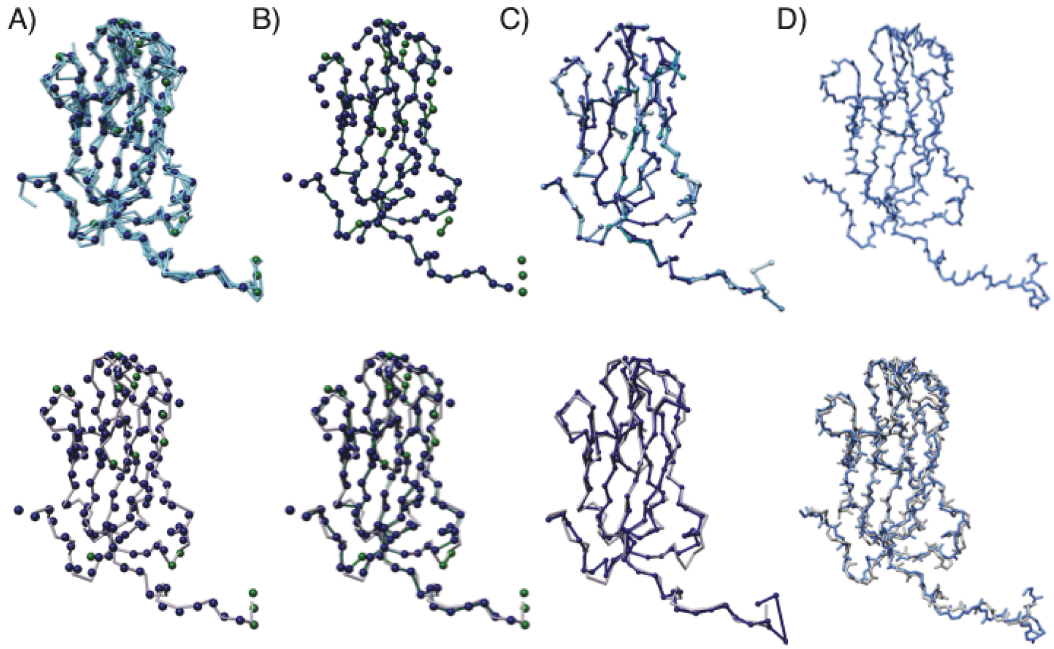

Figure 3.4 Fragment clustering and cluster connectivity. (A) Mutually compatible set of fragments (light blue) and cluster centers found after first (dark blue) and second (green) iteration. The second row shows the found clusters aligned to the native ribbon (grey) (B) Cluster connections based on intra-fragment connectivity. The second row shows the cluster connections aligned to the native ribbon (grey) (C) The first row shows 5 fragmented all-alanine structures. The combined trace is shown in the second row (blue) aligned to the native ribbon in grey. (D) Completed and refined all-alanine structure. The second row shows it in comparison to the native backbone structure (grey).

For building the profile, EMfasa evaluates the correlation of the different rotamers to the density at each backbone position. While doing so, no additional fit of the rotamer to the density is carried out and the input all-alanine backbone is kept fixed. Therefore, the profile is very sensitive to the used all-alanine structure and even small changes can lead to different profiles. We therefore build several all-alanine traces (as described in Section 2.5) and use these as basis to build multiple profiles. Each of these profiles and the corresponding all-alanine structures are in the next step used to assemble the full-atom models based on different alignments. Figure 3.6 shows the full-atom pool for BMV based on one Cα-trace with the correct topology and N-C orientation. The final model pool contains structures ranging from 2.8 Å to up to 20 Å. For selecting the final models close to the native structure, we are currently using the correlation between model and map, referred to as CC_{map}. Figure 3.6 shows that the models closer to the native structure correlate better with the density than the ones that are very different.

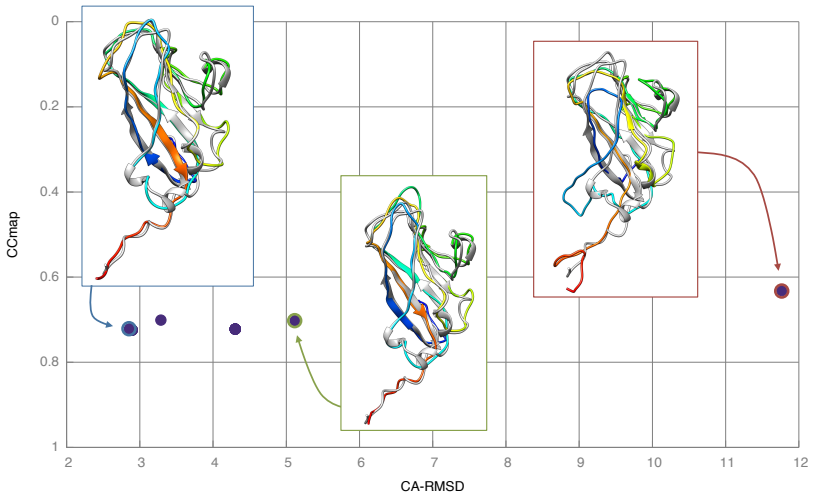

Figure 3.5 Full-atom models based on varying alignments. The plot shows the Cα-RMSD vs. CC_{map} plot of the 21 full-atom models of BMV based on one all-alanine structure and alignments with different gap-weights. For three structures, the final model (rainbow) is shown in comparison to the native structure (gray). The red structure was generated with gapOpen=-2 and gapExtend=-2, the green one with gapOpen=-4, gapExtend=-2 and the blue one with gapOpen=-10, gapExtend=-2.

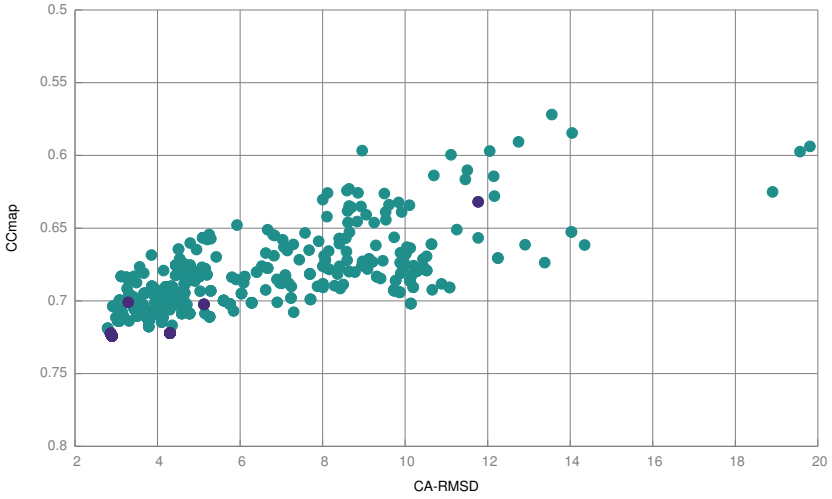

Figure 3.6 Full-atom pool for BMV. The plot shows the Cα-RMSD vs. CC_{map} plot for 525 models generated based on one Cα-trace with correct topology and the correct direction. In total, 25 all-alanine structures where generated and for each of these, 21 full-atom models. The full-atom models shown in Figure 3.5 are highlighted in blue.

3.2 Manuscript 1: Automatic Protein Structure Modelling into cryo-EM density maps using EMfasa

We have written a short manuscript about EMfasa and evaluated its performance on six experimental density maps. As of January/February 2017, the manuscript is submitted and under review. The article is summarised below and both the submitted article and the supplemental information are included in Appendix B (page 127 cont.) of this thesis.

3.2.1 Summary

The high-resolution density maps obtained by cryo-EM using direct electron detectors and powerful image processing routines allow to directly build atomic models. This task is challenging and has only been little explored so far.

We therefore have developed EMfasa, a fully automated protocol for *de novo* model-building into near-atomic cryo-EM density maps. As described in detail in Section 2.1, the protocol couples backbone tracing with sequence non-specific fragment assembly to rapidly generate highly accurate all-alanine structures that are subsequently completed to full-atom models via automated side-chain assignment. The method aims to rapidly build first models which subsequently can be refined using computationally more expensive methods.

EMfasa was tested on six experimental density maps with reported resolutions between 3.3 Å and 4.8 Å with sizes ranging from 149 to 665 residues. For all six targets, EMfasa was able to generate Cα-traces of correct topology and highly accurate all-alanine structures with Cα-RMSDs over at least 85 percent of the protein structure between 0.8 and 2.1 Å. The full-atom models obtained after the sequence-assignment step had on average 70 percent correctly assigned and placed residues for four of the targets. In case of the 4.8 Å resolution map, the information in the side chain density was, however, not good enough for a correct assignment.
In order to contrast our method with other recent approaches, we ran the iterative assembly steps of the Rosetta *de novo* modelling approach described in Section 1.3.4.1 and achieved better performance for two, and a similar performance for four of the targets.

3.2.2 Contribution

I developed steps B-E of EMfasa and implemented them in form of a python package. I generated all data, figures and tables and wrote the majority of the manuscript.

3.3 Publication 2: Archaeal flagellin combines a bacterial type IV pilin domain with an Ig-like domain

An initial version of EMfasa has been used as basis to generate an atomic model of a previously unknown protein structure which has been published in 2016 in *Proc Natl Acad Sci U S A* [187]. The article is summarised below and both the article and supplemental information are included as Appendix C (page 147 cont.) of this thesis.

3.3.1 Summary

While bacterial motility has been studied for many years, only little is known about the flagellar system providing motility in archaea. The archaeal flagellins, the proteins forming the filament, contain an N-terminal domain that is homolog to the N-terminal domain found in bacterial type IV pilin. This highly hydrophobic and conserved N-terminal α-helix has previously been described for the flagellar-like filament Iho670 of the archaeon *Ignicoccus hospitalis* at 7.5 Å resolution [188]. At this limited resolution, it was however not possible to obtain detailed structural information about the large globular domain of Iho670 and its sequence furthermore does not show homology to any other proteins.

Thanks to the direct electron detectors, our collaborators have now been able to reconstruct a density map of Iho670 at ~4Å resolution by using cryo-EM. This high-resolution reconstruction has allowed us to generate a nearly complete model of Iho670. To build the model, an initial version of EMfasa was used and the resulting models were additionally refined and manually combined. The atomic model shows that the globular domain is a β-sandwich and has the same fold expected for true archaeal flagellins. The model furthermore revealed that the archaeal flagellin's outer domains make extensive contacts with each other that largely determine the mechanical properties of these filaments, allowing them to flex. Our structure provides the basis for further studies to understand the archaeal flagellar motility in atomic detail.

3.3.2 Contribution

In this work, I contributed to generating the discussed model of the Iho670 filament using an initial version of EMfasa and additional optimisation steps. I furthermore carried out the interface analysis and the comparison of observed and predicted secondary structures for Iho670, FlaF and several archaeal flagellins.

3.4 2nd EMDataBank Model Challenge

EMDataBank [189] is a unified data resource for deposition and retrieval of 3-dimensional density maps, atomic models, and associated metadata. In addition, it serves as a resource for news, events, software tools, data standards, and validation methods. It is a joint effort of the Protein Data Bank in Europe (PDBe) [190], the Research Collaboratory for Structural Bioinformatics (RCSB) and the National Center for Macromolecular Imaging (NCMI).

In 2015/2016, EMDataBank hosted two community wide challenges with the goals of establishing benchmarks, comparing current practice, and evolving criteria for evaluation of results on reconstruction (Map Challenge) and modelling (Model Challenge) at moderate to high resolution. For the Model Challenge, a group of cryoEM specialists and model software developers established a benchmark set of eight 3DEM maps in the 3.0-4.5 Å resolution range. The Model Challenge therefore provided a great opportunity to test and evaluate our *ab-initio* method on a predefined benchmark set.

3.4.1 Model Challenge Targets

In total, the model challenge committee chose eight targets based on recently reported 3DEM structures, see Table 3.1. The reported resolutions range from 2.2Å (β-Galactosidase) to 4.1Å (GroEL).

3.4.2 Challenge Participation

We have used EMfasa to model the four targets highlighted in Table 3.1, namely TMV at 3.4 Å resolution, the β-subunit of the 20S proteasome at 3.3 Å resolution, BMV at 3.8 Å resolution and nicastrin at 3.4 Å resolution.

3.4.2.1 Map Preparation and Modelling

Each map was masked with the correct deposited PDB model and additionally normalised with *e2proc3d.py* of the EMAN2 package [94]. The Cα-traces, all-alanine models and full-atom models were generated with EMfasa as described in Section 2.1. For each target, a topologically correct Cα-trace was used to assemble the other types of models and the full-atom model with the best assignment was selected from the pool of several hundred structures (roughly 20 to 50 all-alanine structures per trace and 5 to 10 full-atom models per all-alanine structure).

Table 3.1 Model challenge targets

Target	EMDB-Entry	Resolution (Å)	Fitted Model (PDB Id)	Primary Citation	
Tobacco Mosaic Virus	EMD-2842	3.4	4udv	[191]	•
T20S Proteasome	EMD-5623	3.3	3j9i	[113]	•
	EMD-6287	2.8	-	[192]	
GroEL	EMD-6422	4.1	-	-	
TRPV1 Channel	EMD-5778	3.3	3j5p, 3j9j	[107]	
Brome Mosaic Virus	EMD-6000	3.8	3j7l, 3j7m, 3j7n	[186]	•
β-Galactosidase	EMD-5995	3.2	3j7h	[193]	
	EMD-2984	2.2	5a1a	[194]	
γ-Secretase	EMD-2677	4.5	4upc	[195]	
	EMD-3061	3.4	5a63	[196]	•
70S Ribosome	EMD-2847	2.9	5afi	[197]	
	EMD-6316	3.6	3ja1	[198]	

The eight targets chosen by the model challenge committee. Targets that were submitted to the challenge are highlighted (•).

3.4.2.2 Results

For each of the four targets, we submitted the full-atom model with the best assignment. The model with the best assignment was defined as the model maximizing the number of residues that are within 2 Å of their native residue partner. The results are shown in Table 3.2 and are illustrated in Figure 3.7. For all four targets, our method was able to automatically generate a Cα-trace with the correct topology (The low TM-score values for TMV are explained below).

The all-alanine structures for BMV, 20S(β), and nicastrin were of high accuracy with Cα-RMSDs between 1.0 and 1.6 Å. over 90 percent of the protein structure. This is especially impressive for the 665 residue long structure of nicastrin. The TM-scores for the Cα-trace and all-alanine structure of TMV are below 0.5 indicating that our models are comparatively far away from the native structure. In TMV, the N- and C-termini are very close in the three-dimensional space. Therefore, it was at that time (using an intermediate version of the tracing protocol) not possible to correctly identify them automatically. The reported Cα-trace and all-alanine structure therefore have their termini at the wrong location resulting in low TMscores. Using the profiles based on the all-alanine structures with wrong termini, EMfasa was however able to correct the termini as described in Section 2.3.3. The reported values for the final full-atom structure are therefore much better than the ones reported for the intermediate results.

The final full-atom models with the best assignment resulted in Cα-RMSDs between 1.8 and 4.9 Å over the entire structure. On average 64 percent of the residues were assigned

Table 3.2 Model challenge submission

	Cα-trace	all-alanine structure	submitted full-atom model	
Target (#Residues)	**TMscore / RMSD (#aligned)** [1]	**TMscore / RMSD (#aligned)** [1]	**Ca-RMSD (total)**	**Ca-RMSD (corr. assigned /#total)** [2]
BMV (149)	0.86 / 1.6 (142)	0.93 / 1.0 (145)	1.8	1.0 (119/149)
TMV (153)	0.45 / 1.3 (74)	0.46 / 1.3 (74)	2.9	0.8 (91/153)
20S beta (203)	0.84 / 1.6 (182)	0.89 / 1.3 (189)	4.9	0.9 (123/203)
Nicastrin (665)	0.88 / 2.37 (621)	0.94 / 1.6 (642)	3.5	0.9 (370/665)

[1] Structure alignment as generated by TMalign. The TMscore is normalized by the length of the respective native model. TMscore and RMSD are calculated over the aligned residues. [2] Residues that are within 2.0 Å of their native residue partner are considered to be correctly assigned and used for the RMSD calculation. The numbers in brackets indicate the number of correctly assigned residues (used for the RMSD calculation) and the total number of residues in the full-atom model

correctly. The final models are therefore not perfect, but were still able to model the majority of the protein structure correctly. At this point, it is important to keep in mind that EMfasa does not aim to build the perfect structure with zero deviation from the native model, but provides a fast way to build a diverse set of initial models that can be further refined and used to interpret 3DEM reconstructions.

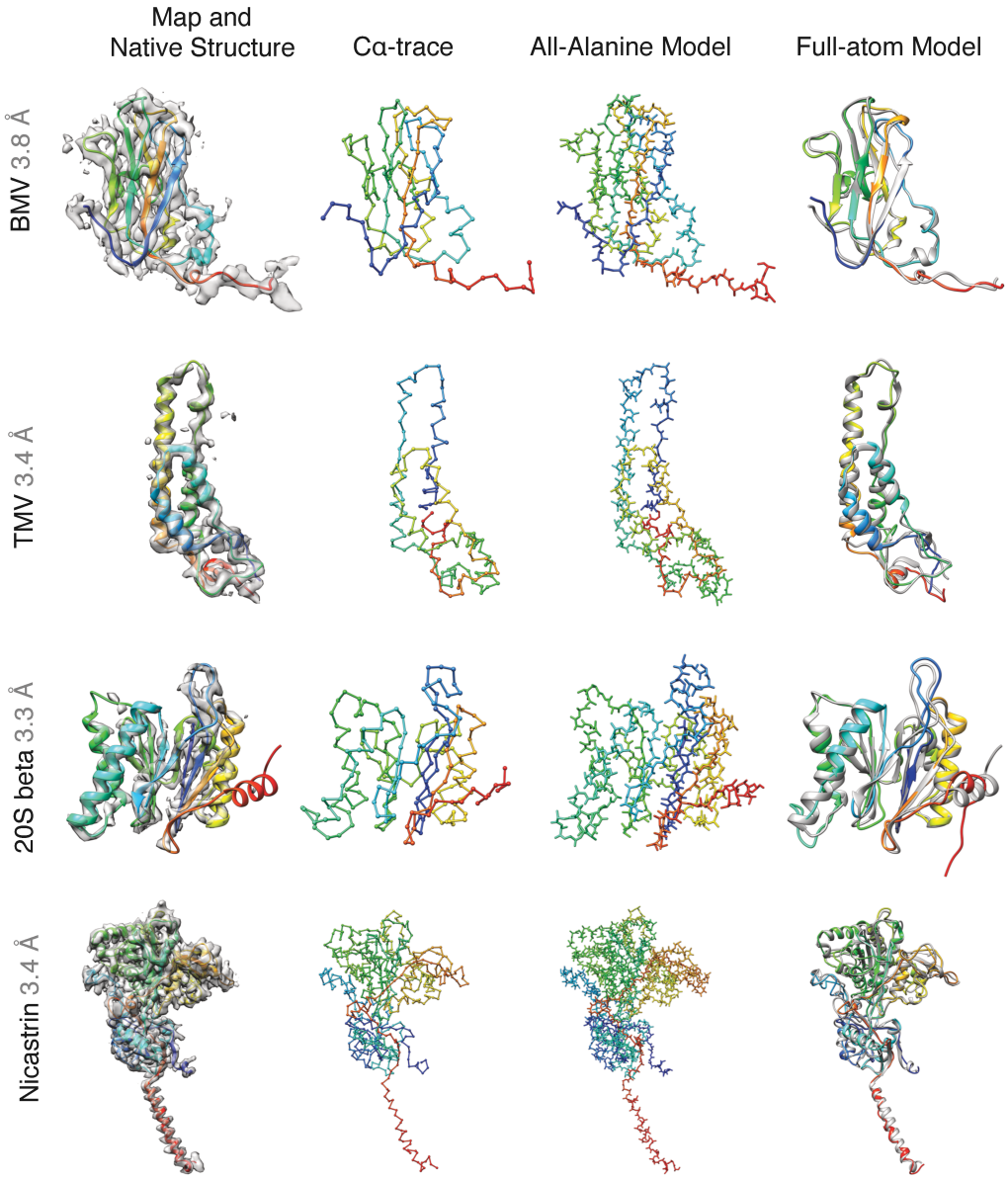

Figure 3.7 Model challenge submission. For each target, the native structure and the experimental density map are shown in column 1) The Cα-trace and the corresponding all-alanine model are shown in columns 2) and 3), respectively. Column 4) shows the EMfasa model with the best assignment compared to the native structure (gray).

4. Discussion

With EMfasa, we have introduced a new approach that rapidly builds three types of models into cryo-EM density maps at near-atomic resolution. In an initial step, EMfasa builds a Cα-trace that roughly describes the positions of the individual models and how the protein chain propagates through the density map, i.e. the protein's topology. Based upon the residue positions and their connections in the Cα-trace, the second type of model is generated via sequence non-specific fragment assembly and clustering: full-atom all-alanine backbone structures. The all-alanine structures can subsequently be used to build full-atom models by aligning the protein sequence to the all-alanine structure based on the fit of each amino acid at each residue position.

In Section 3.2 and 3.4 we have shown that our method was able to build Cα-traces with the correct topology for all tested systems. The resulting all-alanine structures were of very high accuracy, resulting in a Cα-RMSD between 0.8 and 2.1 Å over at least 85 percent of the native protein for the targets discussed in Section 3.2. For four targets of the six discussed targets in this section, the all-alanine structures, the corresponding profiles and resulting profile-sequence alignments could be used to build models whereof the lowest-RMSD models have on average 70 percent correctly assigned and placed residues. For the density map of 4.8 Å, EMfasa, in its current version, reached its limit: only 23 residues were assigned correctly. While the side chain information was not good enough for the side chain assignment, EMfasa was still able to build fairly accurate all-alanine structures (Cα-RMSD of 1.9 Å for 82 % of the residues).
An unfinished version of EMfasa was furthermore used to build an initial model for the flagellar-like filament Iho670 of the archaeon *Ignicoccus hospitalis* from a 4 Å resolution density map [187] as discussed in Section 3.3. The final deposited model was generated by additionally using the cryo-EM refinement protocol of Rosetta [126] and by manually combining several models with Coot [199].

Given that EMfasa produces three different kinds of models that carry different information, we think that it can greatly assist and reduce human efforts at various stages. The pool of full-atom models can be used as basis for structure refinement and determination studies while in cases, where the side chain densities are too low to provide sufficient in-

formation for accurate side chain determination, the all-alanine structures and Cα-traces can be of great interest. They provide important information about the protein's structure and may be used to identify structurally related proteins. Treating the sequence assignment as a global problem, especially in cases of weak side-chain identity, might be of advantage when compared to methods that try to localise residues individually in the density map. In case of unknown proteins with near-atomic resolution density maps, the all-alanine structures built by EMfasa could be used to generate profiles that can in the next step be used to search protein sequence databases for matches.
EMfasa uses a single set of sequence non-specific fragments and therefore scales approximately linearly with the number of amino acids. This fact makes the protocol very interesting for large protein structures, e.g. nicastrin.

Several improvements and future enhancements could furthermore improve the accuracy and enlarge the application spectrum of the initial EMfasa version presented in this thesis. So far, EMfasa has only been used on the segmented density map of the individual monomers. Tracing algorithms, such as Pathwalker, are now able to model multiple subunits without prior segmentation [154]. Including that into EMfasa's tracing step would make prior segmentation redundant. While the all-alanine structures obtained with EMfasa are of remarkable accuracy (cf. Section 3.2 and 3.4), the sequence-assignment step could be further improved. So far, we use a very simplistic approach with a fixed backbone to generate the profiles and the different rotamers are not additionally refined into the density. Hence, a slightly wrong backbone structure can greatly influence the scores of the individual rotamers. A more extensive exploration of the local densities could therefore lead to improved profiles. The alignment between the profile and the protein sequence is currently carried out without any additional information. Secondary structure information of both the sequence and backbone structure and knowledge about probable gap locations could furthermore improve the alignment step. While building the full-atom models with Modeller, no density map information is used. Incorporating the density map directly into the homology modelling step could additionally improve the models. At the moment, we evaluate the final structures based on their correlation with the density map. Including extra refinement steps with accurate full-atom force fields would enable one to select the best structures based on their energy, which might improve the final selection step.

The number of cryo-EM density maps in the specified resolution range is nowadays constantly growing, making the development of methods that are tailored to the near-atomic resolution range highly relevant and we therefore think that, despite the limitations discussed above, EMfasa is a valuable addition to the currently only limited-number of ex-

isting tools that model protein structures into density-maps in the specified resolution range. Analogously, the method can be expected to be useful for the interpretation of electron density maps from X-ray crystallography.

Part IV

Conclusions and Outlook

Conclusions and Outlook

In this thesis, two different approaches for computational structure prediction and calculation were introduced. Both methods rely on two fundamentally different sources of input data and can therefore be used to calculate or predict the structures of targets of completely different nature. Both methods showed promising results and can therefore contribute to closing the ever increasing gap between known protein sequences and protein structures.

In Part II, a method was introduced that combines evolutionary information derived from correlated mutations in multiple sequence alignments of homologous proteins with the iterative sampling protocol RASREC of the Rosetta molecular modelling software suite for accurate protein structure prediction. The protocol was tested on a benchmark set of 28 globular proteins and was compared to another state-of-the-art method of that time. Due to RASRECs robustness against erroneous distance restraints and its iterative sampling strategy, our protocol outperformed the other method by predicting structures to higher accuracies for the majority of the benchmark set.
The accuracy of residue-residue contact predictions derived from multiple sequence alignments of homologous structures greatly depends on the total amount of available sequences. The presented protocol was able to predict high resolution models based on contact predictions with accuracies as low as 45 percent. Due to the ability to efficiently use the sparse information contained in contact predictions with low accuracy, the error robustness and the efficient sampling strategy of the underlying RASREC algorithm, the presented method should be able to predict accurate models in cases where other structure prediction methods would most likely fail to predict the correct fold.
In the past few years, the accuracy of the contact prediction methods was increased even further [200] and the number of known protein sequences is still growing. With that increase in accuracy and the ever growing sequence space, protein structure prediction using evolutionary information can be used for more and more proteins and protein families, making our study highly relevant for the future.

In Part III we explained that the recent increase in resolution of 3DEM reconstructions requires protocols that are tailored to the 3-4.5 Å resolution range of the resulting density

maps. To tackle that problem, we have developed a new approach, called EMfasa, that rapidly builds models into near-atomic resolution density maps that can serve as basis for intensive refinement using computationally more expensive methods.

We showed that EMfasa was able to build Cα-traces with correct topology and highly accurate all-alanine structures for all tested systems. The final pool of full-atom models furthermore contained models close to the native structure for all targets of resolution better than 4 Å.

EMfasa builds three types of models, i.e. Cα-traces, all-alanine structures, and full-atom models, and therefore can assist and reduce human efforts during protein structure determination at various stages. In cases where the density resolution is not high enough to provide sufficient sequence information for side-chain assignment, c.f. the 4.8 Å resolution map of 20S(α) discussed in Section 3.2, both the generated all-alanine structures and Cα-traces provide valuable information about the protein structure. EMfasa does not aim to build perfect full-atom models but rather provides a rapid way to generate a diverse set of full-atom models that can be further refined using computationally more expensive methods.

The so called resolution revolution in cryo-EM [201] resulted in a constant growth of EM map depositions in the Electron Microscopy Data Bank over the past few years with a notable increase in the number of structures at better than 4 Å resolution starting in 2014 [202]. This constant growth, which will not abate in the near future, makes protein modelling in the near-atomic resolution range highly relevant. We therefore think that our method is a valuable contribution to the field of cryo-EM and structural biology in general.

Bibliography

1. Milo, R. What is the total number of protein molecules per cell volume? A call to rethink some published values. *Bioessays* **35,** 1050–1055 (2013).
2. Berg, J. M., Tymoczko, J. L. & Stryer, L. Protein structure and function. *Biochemistry* **5** (2002).
3. Schwede, T. Protein modeling: what happened to the "protein structure gap"? *Structure* **21,** 1531–40 (2013).
4. Lange, O. F. & Baker, D. Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. *Proteins* **80,** 884–95 (2012).
5. Ambrogelly, A., Palioura, S. & Söll, D. Natural expansion of the genetic code. *Nature chemical biology* **3,** 29–35 (2007).
6. Scheeff, E. D. & Fink, J. L. in *Structural Bioinformatics* (eds Bourne, P. E. & Weissig, H.) 15–39 (John Wiley Sons, Inc., 2003).
7. Pauling, L, Corey, R. B. & Branson, H. R. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A* **37,** 205–11 (1951).
8. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28,** 235–42 (2000).
9. *PDB Current Holdings Breakdown* http://www.rcsb.org/pdb/statistics/holdings.do.
10. Smyth, M. S. & Martin, J. H. J. x Ray crystallography. *Molecular Pathology* **53,** 8–14 (Feb. 2000).
11. Ke, H. [25] Overview of isomorphous replacement phasing. *Methods in enzymology* **276,** 448–461 (1997).
12. Rossmann, M. G. The molecular replacement method. *Acta Crystallographica Section A: Foundations of Crystallography* **46,** 73–82 (1990).
13. Wlodawer, A., Minor, W., Dauter, Z. & Jaskolski, M. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS J* **275,** 1–21 (2008).

14. Bränd'en, C.-I. & Alwyn Jones, T. Between objectivity and subjectivity. *Nature* **343,** 687–689 (1990).

15. Brunger, A. T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355,** 472 (1992).

16. Dessau, M. A. & Modis, Y. Protein crystallization for X-ray crystallography. *JoVE (Journal of Visualized Experiments),* e2285–e2285 (2011).

17. Ishima, R. & Torchia, D. A. Protein dynamics from NMR. *Nature Structural & Molecular Biology* **7,** 740 (2000).

18. Dobson, C. M. & Hore, P. J. Kinetic studies of protein folding using NMR spectroscopy. *Nature Structural & Molecular Biology* **5,** 504–507 (1998).

19. Kwan, A. H., Mobli, M., Gooley, P. R., King, G. F. & Mackay, J. P. Macromolecular NMR spectroscopy for the non-spectroscopist. *FEBS J* **278,** 687–703 (2011).

20. Nabuurs, S. B., Spronk, C. A., Vriend, G. & Vuister, G. W. Concepts and tools for NMR restraint analysis and validation. *Concepts in Magnetic Resonance Part A* **22,** 90–105 (2004).

21. Yu, H. Extending the size limit of protein nuclear magnetic resonance. *Proc Natl Acad Sci U S A* **96,** 332–4 (1999).

22. Wüthrich, K. The second decade–into the third millenium. *Nat Struct Biol* **5 Suppl,** 492–5 (1998).

23. Frueh, D. P., Goodrich, A. C., Mishra, S. H. & Nichols, S. R. NMR methods for structural studies of large monomeric and multimeric proteins. *Curr Opin Struct Biol* **23,** 734–9 (2013).

24. Krieger, E., Nabuurs, S. B. & Vriend, G. Homology modeling. *Methods Biochem Anal* **44,** 509–23 (2003).

25. Chothia, C & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J* **5,** 823–6 (1986).

26. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng* **12,** 85–94 (1999).

27. Xiang, Z. Advances in homology protein structure modeling. *Current Protein and Peptide Science* **7,** 217–227 (2006).

28. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215,** 403–410 (1990).

29. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences* **85,** 2444–2448 (1988).

30. Godzik, A. Fold recognition methods. *Methods Biochem Anal* **44,** 525–46 (2003).

31. Xu, J., Jiao, F. & Yu, L. Protein structure prediction using threading. *Protein Structure Prediction,* 91–121 (2008).

32. Chivian, D., Robertson, T., Bonneau, R. & Baker, D. Ab initio methods. *Methods Biochem Anal* **44,** 547–57 (2003).

33. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181,** 223–30 (1973).

34. Zhang, Y. Progress and challenges in protein structure prediction. *Current opinion in structural biology* **18,** 342–348 (2008).

35. Bradley, P., Misura, K. M. S. & Baker, D. Toward high-resolution de novo structure prediction for small proteins. *Science* **309,** 1868–71 (2005).

36. Lü, Q., Xia, X.-Y., Chen, R., Miao, D.-J., Chen, S.-S., Quan, L.-J. & Li, H.-O. When the lowest energy does not induce native structures: parallel minimization of multi-energy values by hybridizing searching intelligences. *PloS one* **7,** e44967 (2012).

37. Bonneau, R & Baker, D. Ab initio protein structure prediction: progress and prospects. *Annu Rev Biophys Biomol Struct* **30,** 173–89 (2001).

38. Bowie, J. U. & Eisenberg, D. An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. *Proc Natl Acad Sci U S A* **91,** 4436–40 (1994).

39. Jones, D. T. Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins* **Suppl 1,** 185–91 (1997).

40. Jones, D. T. Predicting novel protein folds by using FRAGFOLD. *Proteins* **Suppl 5,** 127–32 (2001).

41. Simons, K. T., Kooperberg, C, Huang, E & Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* **268,** 209–25 (1997).

42. Simons, K. T., Bonneau, R, Ruczinski, I & Baker, D. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* **Suppl 3,** 171–6 (1999).

43. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35,** D61–5 (2007).

44. Illergård, K., Ardell, D. H. & Elofsson, A. Structure is three to ten times more conserved than sequence–a study of structural response in protein cores. *Proteins* **77,** 499–508 (2009).

45. Overington, J, Johnson, M. S., Sali, A & Blundell, T. L. Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc Biol Sci* **241,** 132–45 (1990).

46. Worth, C. L., Gong, S. & Blundell, T. L. Structural and functional constraints in the evolution of protein families. *Nat Rev Mol Cell Biol* **10,** 709–20 (2009).

47. Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R. & Sander, C. Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6,** e28766 (2011).

48. Marks, D. S., Hopf, T. A. & Sander, C. Protein structure prediction from sequence variation. *Nature biotechnology* **30,** 1072–1080 (2012).

49. Göbel, U, Sander, C, Schneider, R & Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins* **18,** 309–17 (1994).

50. Hatrick, K & Taylor, W. R. Sequence conservation and correlation measures in protein structure prediction. *Comput Chem* **18,** 245–9 (1994).

51. Neher, E. How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci U S A* **91,** 98–102 (1994).

52. Shindyalov, I. N., Kolchanov, N. A. & Sander, C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng* **7,** 349–58 (1994).

53. Burger, L. & van Nimwegen, E. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput Biol* **6,** e1000633 (2010).

54. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A* **108,** E1293–301 (2011).

55. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci U S A* **106,** 67–72 (2009).

56. Lapedes, A., Giraud, B. & Jarzynski, C. Using sequence alignments to predict protein structure and stability with high accuracy. *arXiv preprint arXiv:1207.2484* (2012).

57. Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. *Proceedings of the National Academy of Sciences* **110,** 15674–15679 (2013).

58. Seemayer, S., Gruber, M. & Söding, J. CCMpred–fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics* **30,** 3128–30 (2014).

59. Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys Rev E Stat Nonlin Soft Matter Phys* **87,** 012707 (2013).

60. Ekeberg, M., Hartonen, T. & Aurell, E. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *Journal of Computational Physics* **276,** 341–356 (2014).

61. Jones, D. T., Buchan, D. W. A., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28,** 184–90 (2012).

62. Michel, M., Hayat, S., Skwark, M. J., Sander, C., Marks, D. S. & Elofsson, A. PconsFold: improved contact predictions improve protein models. *Bioinformatics* **30,** i482–8 (2014).

63. Kosciolek, T. & Jones, D. T. De novo structure prediction of globular proteins aided by sequence variation-derived contacts. *PLoS One* **9,** e92197 (2014).

64. Adhikari, B., Bhattacharya, D., Cao, R. & Cheng, J. CONFOLD: Residue-residue contact-guided ab initio protein folding. *Proteins* **83,** 1436–49 (2015).

65. Ovchinnikov, S., Kim, D. E., Wang, R. Y.-R., Liu, Y., DiMaio, F. & Baker, D. Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta. *Proteins* **84 Suppl 1,** 67–75 (2016).

66. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* **3,** e02030 (2014).

67. Hopf, T. A. *et al.* Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3** (2014).

68. Hopf, T. A., Colwell, L. J., Sheridan, R., Rost, B., Sander, C. & Marks, D. S. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **149,** 1607–21 (2012).

69. Hayat, S., Sander, C., Marks, D. S. & Elofsson, A. All-atom 3D structure prediction of transmembrane ß-barrel proteins from sequences. *Proc Natl Acad Sci U S A* **112,** 5413–8 (2015).

70. Kaufmann, K. W., Lemmon, G. H., DeLuca, S. L., Sheehan, J. H. & Meiler, J. Practically Useful: What the Rosetta Protein Modeling Suite Can Do for You. *Biochemistry* **49,** 2987–2998 (Apr. 2010).

71. Bradley, P. *et al.* Free modeling with Rosetta in CASP6. *Proteins* **61 Suppl 7,** 128–34 (2005).

72. Raman, S. *et al.* Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* **77 Suppl 9,** 89–99 (2009).

73. Kim, D. E., Dimaio, F., Yu-Ruei Wang, R., Song, Y. & Baker, D. One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins* **82 Suppl 2,** 208–18 (2014).

74. Rohl, C. A., Strauss, C. E., Misura, K. M. & Baker, D. Protein structure prediction using Rosetta. *Methods in enzymology* **383,** 66–93 (2004).

75. Bazzoli, A., Kelow, S. P. & Karanicolas, J. Enhancements to the Rosetta Energy Function Enable Improved Identification of Small Molecules that Inhibit Protein-Protein Interactions. *PLoS One* **10,** e0140359 (2015).

76. Simons, K. T., Ruczinski, I, Kooperberg, C, Fox, B. A., Bystroff, C & Baker, D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* **34,** 82–95 (1999).

77. Bonneau, R. *et al.* De novo prediction of three-dimensional structures for major protein families. *J Mol Biol* **322,** 65–78 (2002).

78. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast computing machines. *The journal of chemical physics* **21,** 1087–1092 (1953).

79. Rohl, C. A. & Baker, D. De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *J Am Chem Soc* **124,** 2723–9 (2002).

80. Rohl, C. A. Protein structure estimation from minimal restraints using Rosetta. *Methods Enzymol* **394,** 244–60 (2005).

81. Raman, S. *et al.* NMR structure determination for larger proteins using backbone-only data. *Science* **327,** 1014–8 (2010).

82. Reichel, K., Fisette, O., Braun, T., Lange, O. F., Hummer, G. & Schäfer, L. V. Systematic evaluation of CS-Rosetta for membrane protein structure prediction with sparse NOE restraints. *Proteins* (2016).

83. Ovchinnikov, S., Park, H., Kim, D. E., Liu, Y., Wang, R. Y.-R. & Baker, D. Structure prediction using sparse simulated NOE restraints with Rosetta in CASP11. *Proteins* **84 Suppl 1,** 181–8 (2016).

84. Kahraman, A., Herzog, F., Leitner, A., Rosenberger, G., Aebersold, R. & Malmström, L. Cross-link guided molecular modeling with ROSETTA. *PLoS One* **8,** e73411 (2013).

85. Lange, O. F. *et al.* Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *Proc Natl Acad Sci U S A* **109,** 10873–8 (2012).

86. Braun, T., Koehler Leman, J. & Lange, O. F. Combining Evolutionary Information and an Iterative Sampling Strategy for Accurate Protein Structure Prediction. *PLoS Computational Biology* **11** (ed Marks, D. S.) e1004661 (Dec. 2015).

87. Method of the Year 2015 [Editorial]. *Nature Methods* **13** (2016).

88. *Structural Bioinformatics (Methods of Biochemical Analysis, V. 44)* (eds Bourne, P. E. & Weissig, H.) chap. Electron Microscopy (Wiley-Liss, 2003).

89. Frank, J. *Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state* (Oxford University Press, 2006).

90. Doerr, A. Single-particle cryo-electron microscopy. *Nat Methods* **13,** 23 (2016).

91. Knoll, M. & Ruska, E. Das elektronenmikroskop. *Zeitschrift für Physik* **78,** 318–339 (1932).

92. Dykstra, M. Biological electron microscopy. *Theory, techniques and troubleshooting. Plenum Publishing Corporation.: Biological electron microscopy,* 219–220 (1992).

93. Kühlbrandt, W. Cryo-EM enters a new era. *Elife* **3,** e03678 (2014).

94. Tang, G., Peng, L., Baldwin, P. R., Mann, D. S., Jiang, W., Rees, I. & Ludtke, S. J. EMAN2: an extensible image processing suite for electron microscopy. *J Struct Biol* **157,** 38–46 (2007).

95. Frank, J, Radermacher, M, Penczek, P, Zhu, J, Li, Y, Ladjadj, M & Leith, A. SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *J Struct Biol* **116,** 190–9 (1996).

96. Van Heel, M, Harauz, G, Orlova, E. V., Schmidt, R & Schatz, M. A new generation of the IMAGIC image processing system. *J Struct Biol* **116,** 17–24 (1996).

97. Henderson, R. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q Rev Biophys* **28,** 171–93 (1995).

98. Penczek, P. A. Fundamentals of three-dimensional reconstruction from projections. *Methods Enzymol* **482,** 1–33 (2010).

99. Radermacher, M. Three-dimensional reconstruction of single particles from random and nonrandom tilt series. *J Electron Microsc Tech* **9,** 359–94 (1988).

100. Van Heel, M. Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction. *Ultramicroscopy* **21,** 111–23 (1987).

101. Cheng, Y. Single-Particle Cryo-EM at Crystallographic Resolution. *Cell* **161,** 450–7 (2015).

102. Henderson, R. & Glaeser, R. M. Quantitative analysis of image contrast in electron micrographs of beam-sensitive crystals. *Ultramicroscopy* **16,** 139–150 (1985).

103. Harauz, G. & van Heel, M. Exact filters for general geometry three dimensional reconstruction. *Optik* **73,** 146–156 (1986).

104. Wong, W. *et al.* Cryo-EM structure of the Plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine. *Elife* **3** (2014).

105. Amunts, A. *et al.* Structure of the yeast mitochondrial large ribosomal subunit. *Science* **343,** 1485–9 (2014).

106. Voorhees, R. M., Fernández, I. S., Scheres, S. H. W. & Hegde, R. S. Structure of the mammalian ribosome-Sec61 complex to 3.4 Å resolution. *Cell* **157,** 1632–43 (2014).

107. Liao, M., Cao, E., Julius, D. & Cheng, Y. Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* **504,** 107–12 (2013).

108. Cao, E., Liao, M., Cheng, Y. & Julius, D. TRPV1 structures in distinct conformations reveal activation mechanisms. *Nature* **504,** 113–8 (2013).

109. Allegretti, M., Mills, D. J., McMullan, G., Kühlbrandt, W. & Vonck, J. Atomic model of the F420-reducing [NiFe] hydrogenase by electron cryo-microscopy using a direct electron detector. *Elife* **3,** e01963 (2014).

110. Kühlbrandt, W. Biochemistry. The resolution revolution. *Science* **343,** 1443–4 (2014).

111. Bai, X.-c., McMullan, G. & Scheres, S. H. W. How cryo-EM is revolutionizing structural biology. *Trends Biochem Sci* **40,** 49–57 (2015).

112. McMullan, G, Clark, A. T., Turchetta, R & Faruqi, A. R. Enhanced imaging in low dose electron microscopy using electron counting. *Ultramicroscopy* **109,** 1411–6 (2009).

113. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat Methods* **10,** 584–90 (2013).

114. McMullan, G, Faruqi, A. R., Clare, D & Henderson, R. Comparison of optimal performance at 300keV of three direct electron detectors for use in low dose electron microscopy. *Ultramicroscopy* **147,** 156–63 (2014).

115. Bai, X.-C., Fernandez, I. S., McMullan, G. & Scheres, S. H. W. Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. *Elife* **2,** e00461 (2013).

116. Brilot, A. F. *et al.* Beam-induced motion of vitrified specimen on holey carbon film. *J Struct Biol* **177,** 630–7 (2012).

117. Campbell, M. G. *et al.* Movies of ice-embedded particles enhance resolution in electron cryo-microscopy. *Structure* **20,** 1823–8 (2012).

118. Scheres, S. H. W., Gao, H., Valle, M., Herman, G. T., Eggermont, P. P. B., Frank, J. & Carazo, J.-M. Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. *Nat Methods* **4,** 27–9 (2007).

119. Sigworth, F. J. A maximum-likelihood approach to single-particle image refinement. *J Struct Biol* **122,** 328–39 (1998).

120. Scheres, S. H. W. Classification of structural heterogeneity by maximum-likelihood methods. *Methods Enzymol* **482,** 295–320 (2010).

121. Lyumkis, D., Brilot, A. F., Theobald, D. L. & Grigorieff, N. Likelihood-based classification of cryo-EM images using FREALIGN. *J Struct Biol* **183,** 377–88 (2013).

122. Scheres, S. H. W. A Bayesian view on cryo-EM structure determination. *J Mol Biol* **415,** 406–18 (2012).

123. Brünger, A. T. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallographica Section D: Biological Crystallography* **54,** 905–921 (1998).

124. Brunger, A. T. Version 1.2 of the Crystallography and NMR system. *Nature protocols* **2,** 2728–2733 (2007).

125. DiMaio, F., Tyka, M. D., Baker, M. L., Chiu, W. & Baker, D. Refinement of protein structures into low-resolution density maps using rosetta. *J Mol Biol* **392,** 181–90 (2009).

126. DiMaio, F. *et al.* Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. *Nat Methods* **12,** 361–5 (2015).

127. Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr D Biol Crystallogr* **68,** 352–67 (2012).

128. Esquivel-Rodríguez, J. & Kihara, D. Computational methods for constructing protein structure models from 3D electron microscopy maps. *J Struct Biol* **184,** 93–102 (2013).

129. Baker, M. L., Baker, M. R., Hryc, C. F. & Dimaio, F. Analyses of subnanometer resolution cryo-EM density maps. *Methods Enzymol* **483,** 1–29 (2010).

130. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. UCSF Chimera–a visualization system for exploratory research and analysis. *J Comput Chem* **25,** 1605–12 (2004).

131. Pintilie, G. D., Zhang, J., Goddard, T. D., Chiu, W. & Gossard, D. C. Quantitative analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions. *J Struct Biol* **170,** 427–38 (2010).

132. Baker, M. L., Yu, Z., Chiu, W. & Bajaj, C. Automated segmentation of molecular subunits in electron cryomicroscopy density maps. *J Struct Biol* **156,** 432–41 (2006).

133. Burger, V. & Chennubhotla, C. Nhs: network-based hierarchical segmentation for cryo-electron microscopy density maps. *Biopolymers* **97,** 732–41 (2012).

134. DiMaio, F & Chiu, W. Tools for Model Building and Optimization into Near-Atomic Resolution Electron Cryo-Microscopy Density Maps. *Methods Enzymol* **579,** 255–76 (2016).

135. Rossmann, M. G., Bernal, R & Pletnev, S. V. Combining electron microscopic with x-ray crystallographic structures. *J Struct Biol* **136,** 190–200 (2001).

136. Rossmann, M. G. Fitting atomic models into electron-microscopy maps. *Acta Crystallogr D Biol Crystallogr.* **56(Pt 10),** 1341–9 (2000).

137. Woetzel, N., Lindert, S., Stewart, P. L. & Meiler, J. BCL::EM-Fit: rigid body fitting of atomic structures into density maps using geometric hashing and real space refinement. *J Struct Biol* **175,** 264–76 (2011).

138. Schröder, G. F., Brunger, A. T. & Levitt, M. Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. *Structure* **15,** 1630–41 (2007).

139. Topf, M., Lasker, K., Webb, B., Wolfson, H., Chiu, W. & Sali, A. Protein structure fitting and refinement guided by cryo-EM density. *Structure* **16,** 295–307 (2008).

140. Tama, F., Miyashita, O. & Brooks, C. L. 3rd. Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM. *J Struct Biol* **147,** 315–26 (2004).

141. Suhre, K., Navaza, J. & Sanejouand, Y. H. NORMA: a tool for flexible fitting of high-resolution protein structures into low-resolution electron-microscopy-derived density maps. *Acta Crystallogr D Biol Crystallogr* **62,** 1098–100 (2006).

142. Tan, R. K.-Z., Devkota, B. & Harvey, S. C. YUP.SCX: coaxing atomic models into medium resolution electron density maps. *J Struct Biol* **163,** 163–74 (2008).

143. Trabuco, L. G., Villa, E., Schreiner, E., Harrison, C. B. & Schulten, K. Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and X-ray crystallography. *Methods* **49,** 174–80 (2009).

144. Wang, Z. & Schröder, G. F. Real-space refinement with DireX: from global fitting to side-chain improvements. *Biopolymers* **97,** 687–97 (2012).

145. De Groot, B. L., van Aalten, D. M., Scheek, R. M., Amadei, A, Vriend, G & Berendsen, H. J. Prediction of protein conformational freedom from distance constraints. *Proteins* **29,** 240–51 (1997).

146. Falkner, B. & Schröder, G. F. Cross-validation in cryo-EM-based structural modeling. *Proc Natl Acad Sci U S A* **110,** 8930–5 (2013).

147. Jiang, W, Baker, M. L., Ludtke, S. J. & Chiu, W. Bridging the information gap: computational tools for intermediate resolution structure interpretation. *J Mol Biol* **308,** 1033–44 (2001).

148. Kong, Y. & Ma, J. A structural-informatics approach for mining beta-sheets: locating sheets in intermediate-resolution density maps. *J Mol Biol* **332,** 399–413 (2003).

149. Kong, Y., Zhang, X., Baker, T. S. & Ma, J. A Structural-informatics approach for tracing beta-sheets: building pseudo-C(alpha) traces for beta-strands in intermediate-resolution density maps. *J Mol Biol* **339,** 117–30 (2004).

150. Baker, M. L., Ju, T. & Chiu, W. Identification of secondary structure elements in intermediate-resolution density maps. *Structure* **15,** 7–19 (2007).

151. Ludtke, S. J., Baker, M. L., Chen, D.-H., Song, J.-L., Chuang, D. T. & Chiu, W. De novo backbone trace of GroEL from single particle electron cryomicroscopy. *Structure* **16,** 441–8 (2008).

152. Ju, T., Baker, M. L. & Chiu, W. Computing a family of skeletons of volumetric models for shape description. *Comput Aided Des* **39,** 352–360 (2007).

153. Baker, M. R., Rees, I., Ludtke, S. J., Chiu, W. & Baker, M. L. Constructing and validating initial CÎś models from subnanometer resolution density maps with pathwalking. *Structure* **20,** 450–63 (2012).

154. Chen, M., Baldwin, P. R., Ludtke, S. J. & Baker, M. L. De Novo modeling in cryo-EM density maps with Pathwalking. *Journal of Structural Biology* **196,** 289–298 (2016).

155. Lindert, S., Staritzbichler, R., Wötzel, N., Karakaş, M., Stewart, P. L. & Meiler, J. EM-fold: De novo folding of alpha-helical proteins guided by intermediate-resolution electron microscopy density maps. *Structure* **17,** 990–1003 (2009).

156. Lindert, S., Hofmann, T., Wötzel, N., Karakaş, M., Stewart, P. L. & Meiler, J. Ab initio protein modeling into CryoEM density maps using EM-Fold. *Biopolymers* **97,** 669–77 (2012).

157. Wang, R. Y.-R. *et al.* De novo protein structure determination from near-atomic-resolution cryo-EM maps. *Nat Methods* **12,** 335–8 (2015).

158. Song, Y. *et al.* High-resolution comparative modeling with RosettaCM. *Structure* **21,** 1735–42 (2013).

159. Emsley, P, Lohkamp, B, Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66,** 486–501 (2010).

160. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66,** 213–21 (2010).

161. Cowtan, K. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr D Biol Crystallogr* **62,** 1002–11 (2006).

162. Terwilliger, T. C. SOLVE and RESOLVE: automated structure solution and density modification. *Methods Enzymol* **374,** 22–37 (2003).

163. Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat Protoc* **3,** 1171–9 (2008).

164. Wang, Z. *De novo protein backbone modeling with low-resolution density maps* PhD thesis (Heinrich Heine University Düsseldorf, 2014).

165. Lin, S. & Kernighan, B. W. An effective heuristic algorithm for the traveling-salesman problem. *Operations research* **21,** 498–516 (1973).

166. Helsgaun, K. An effective implementation of the Lin–Kernighan traveling salesman heuristic. *European Journal of Operational Research* **126,** 106–130 (2000).

167. Wang, G. & Dunbrack, R. L. Jr. PISCES: a protein sequence culling server. *Bioinformatics* **19,** 1589–91 (2003).

168. Krivov, G. G., Shapovalov, M. V. & Dunbrack, R. L. Jr. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* **77,** 778–95 (2009).

169. Jamroz, M. & Kolinski, A. ClusCo: clustering and comparison of protein models. *BMC Bioinformatics* **14,** 62 (2013).

170. Kirkpatrick, S, Gelatt, C. D. Jr & Vecchi, M. P. Optimization by simulated annealing. *Science* **220,** 671–80 (1983).

171. Vanderbilt, D. & Louie, S. G. A Monte Carlo simulated annealing approach to optimization over continuous variables. *Journal of Computational Physics* **56,** 259–271 (1984).

172. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *AAAI Press,* 226–231 (1996).

173. Afonine, P., Headd, J., Terwilliger, T. & Adams, P. *Computational Crystallography Newsletter* **4,** 43–44 (2013).

174. Dunbrack, R. L. Rotamer Libraries in the 21 st Century. *Current opinion in structural biology* **12,** 431–440 (2002).

175. Scouras, A. D. & Daggett, V. The dynameomics rotamer library: Amino acid side chain conformations and dynamics from comprehensive molecular dynamics simulations in water. *Protein Science* **20,** 341–352 (2011).

176. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48,** 443–53 (1970).

177. Gotoh, O. An improved algorithm for matching biological sequences. *J Mol Biol* **162,** 705–8 (1982).

178. Blazewicz, J., Frohmberg, W., Kierzynka, M., Pesch, E. & Wojciechowski, P. Protein alignment algorithms with an efficient backtracking routine on multiple GPUs. *BMC Bioinformatics* **12,** 181 (2011).

179. Sali, A & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234,** 779–815 (1993).

180. Webb, B. & Sali, A. Protein structure modeling with MODELLER. *Protein Structure Prediction,* 1–15 (2014).

181. Tickle, I. J. Statistical quality indicators for electron-density maps. *Acta Crystallographica Section D: Biological Crystallography* **68,** 454–467 (2012).

182. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics* **57,** 702–710 (2004).

183. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33,** 2302–9 (2005).

184. Kolodny, R., Koehl, P., Guibas, L. & Levitt, M. Small libraries of protein fragments model native protein structures accurately. *J Mol Biol* **323,** 297–307 (2002).

185. De Oliveira, S. H. P., Shi, J. & Deane, C. M. Building a better fragment library for de novo protein structure prediction. *PLoS One* **10,** e0123998 (2015).

186. Wang, Z. *et al.* An atomic model of brome mosaic virus using direct electron detection and real-space optimization. *Nat Commun* **5,** 4808 (2014).

187. Braun, T. *et al.* Archaeal flagellin combines a bacterial type IV pilin domain with an Ig-like domain. *Proc Natl Acad Sci U S A* **113,** 10352–7 (2016).

188. Yu, X., Goforth, C., Meyer, C., Rachel, R., Wirth, R., Schröder, G. F. & Egelman, E. H. Filaments from Ignicoccus hospitalis show diversity of packing in proteins containing N-terminal type IV pilin helices. *J Mol Biol* **422,** 274–81 (2012).

189. Lawson, C. L. *et al.* EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res* **39,** D456–64 (2011).

190. Velankar, S. *et al.* PDBe: Protein Data Bank in Europe. *Nucleic Acids Res* **39,** D402–10 (2011).

191. Fromm, S. A., Bharat, T. A. M., Jakobi, A. J., Hagen, W. J. H. & Sachse, C. Seeing tobacco mosaic virus through direct electron detectors. *J Struct Biol* **189,** 87–97 (2015).

192. Campbell, M. G., Veesler, D., Cheng, A., Potter, C. S. & Carragher, B. 2.8 Å resolution reconstruction of the Thermoplasma acidophilum 20S proteasome using cryo-electron microscopy. *Elife* **4** (2015).

193. Bartesaghi, A., Matthies, D., Banerjee, S., Merk, A. & Subramaniam, S. Structure of ß-galactosidase at 3.2-Å resolution obtained by cryo-electron microscopy. *Proc Natl Acad Sci U S A* **111,** 11709–14 (2014).

194. Bartesaghi, A., Merk, A., Banerjee, S., Matthies, D., Wu, X., Milne, J. L. S. & Subramaniam, S. 2.2 Å resolution cryo-EM structure of ß-galactosidase in complex with a cell-permeant inhibitor. *Science* **348,** 1147–51 (2015).

195. Lu, P. *et al.* Three-dimensional structure of human γ-secretase. *Nature* **512,** 166–70 (2014).

196. Bai, X.-c. *et al.* An atomic structure of human γ-secretase. *Nature* **525,** 212–7 (2015).

197. Fischer, N., Neumann, P., Konevega, A. L., Bock, L. V., Ficner, R., Rodnina, M. V. & Stark, H. Structure of the E. coli ribosome-EF-Tu complex at <3 Å resolution by Cs-corrected cryo-EM. *Nature* **520,** 567–70 (2015).

198. Li, W., Liu, Z., Koripella, R. K., Langlois, R., Sanyal, S. & Frank, J. Activation of GTP hydrolysis in mRNA-tRNA translocation by elongation factor G. *Sci Adv* **1** (2015).

199. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D: Biological Crystallography* **60,** 2126–2132 (2004).

200. Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput Biol* **13,** e1005324 (2017).

201. Kühlbrandt, W. The resolution revolution. *Science* **343,** 1443–1444 (2014).

202. Nogales, E. The development of cryo-EM into a mainstream structural biology technique. *Nature methods* **13,** 24–27 (2016).

Part V

Appendices

A. Embedded Publication 1

Braun T, Koehler Leman J, Lange OF. *Combining Evolutionary Information and an Iterative Sampling Strategy for Accurate Protein Structure Prediction.* PLoS Comput Biol. 2015 Dec 29;11(12):e1004661.

A.1 Copy Permissions

A.2 Full Article

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Combining Evolutionary Information and an Iterative Sampling Strategy for Accurate Protein Structure Prediction

Tatjana Braun[1¤]*, **Julia Koehler Leman**[2], **Oliver F. Lange**[1]

1 Biomolecular NMR and Munich Center for Integrated Protein Science, Department Chemie, Technische Universität München, Garching, Germany, **2** Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, Maryland, United States of America

¤ Current address: Institute of Complex Systems, Structural Biochemistry (ICS-6), Forschungszentrum Jülich, Jülich, Germany
* tatjana.braun@tum.de

OPEN ACCESS

Citation: Braun T, Koehler Leman J, Lange OF (2015) Combining Evolutionary Information and an Iterative Sampling Strategy for Accurate Protein Structure Prediction. PLoS Comput Biol 11(12): e1004661. doi:10.1371/journal.pcbi.1004661

Editor: Debora S Marks, Harvard Medical School, UNITED STATES

Received: March 15, 2015

Accepted: November 17, 2015

Published: December 29, 2015

Data Availability Statement: All data can be reproduced given the scripts and instructions provided in the protocol capture (available as Supporting Information and with the current Rosetta release) and the files in the Supporting Information.

Funding: This work was supported by DFG grant LA 1817/3-1. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Recent work has shown that the accuracy of *ab initio* structure prediction can be significantly improved by integrating evolutionary information in form of intra-protein residue-residue contacts. Following this seminal result, much effort is put into the improvement of contact predictions. However, there is also a substantial need to develop structure prediction protocols tailored to the type of restraints gained by contact predictions. Here, we present a structure prediction protocol that combines evolutionary information with the resolution-adapted structural recombination approach of Rosetta, called RASREC. Compared to the classic Rosetta *ab initio* protocol, RASREC achieves improved sampling, better convergence and higher robustness against incorrect distance restraints, making it the ideal sampling strategy for the stated problem. To demonstrate the accuracy of our protocol, we tested the approach on a diverse set of 28 globular proteins. Our method is able to converge for 26 out of the 28 targets and improves the average TM-score of the entire benchmark set from 0.55 to 0.72 when compared to the top ranked models obtained by the EVFold web server using identical contact predictions. Using a smaller benchmark, we furthermore show that the prediction accuracy of our method is only slightly reduced when the contact prediction accuracy is comparatively low. This observation is of special interest for protein sequences that only have a limited number of homologs.

Author Summary

Recently, a breakthrough has been achieved in modeling the atomic 3D structures of proteins from their sequence alone without requiring any experimental work on the protein itself. To achieve this goal, a database of evolutionary related sequences is analyzed to find co-evolving residues, giving insight into which residues are in close proximity to each other. These residue-residue contacts can help to drive a computer simulation with an atomic-scale physical model of the protein structure from a random starting conformation

to a native-like 3D conformation. Although much effort is being put into the improvement of residue-residue contact predictions, their accuracy will always be limited. Therefore, structure prediction protocols with a high tolerance against incorrect distance restraints are needed. Here, we present a structure prediction protocol that combines evolutionary information with the iterative sampling approach of the molecular modeling suite Rosetta, called RASREC. RASREC has been shown to converge faster to near-native models and to be more robust against incorrect distance restraints than standard prediction protocols. It is therefore perfectly suited for restraints obtained from predicted residue-residue contacts with limited accuracy. We show that our protocol outperforms other currently published structure prediction methods and is able to achieve accurate structures, even if the accuracy of predicted contacts is low.

"This is a *PLOS Computational Biology* Methods paper"

Introduction

The computational prediction of protein structures from their amino acid sequence is an ongoing challenge that has occupied scientists for more than four decades. While Anfinsen's dogma [1] suggests that for most proteins the information contained in their amino acid sequence is sufficient to define their three-dimensional structure, the problem still remains largely unsolved. For some small proteins (<80 residues), current *ab initio* prediction methods are successful in predicting the corresponding 3D structures with high accuracy. One such method is the Rosetta *ab initio* protocol, which assembles short fragments of known proteins by a Monte Carlo strategy [2,3]. With increasing protein size however, sampling of the large conformational space becomes a major challenge [4] and combination with experimental data is required to achieve accurate protein models [5,6].

As experimental data is not always available and may be difficult or costly to obtain, researchers have focused on reducing the search space of possible protein conformations in other ways, for instance by including evolutionary information found in patterns of correlated mutations in protein sequences. The underlying assumption is that these correlated pairs indicate spatial proximity in the protein structure and can therefore be used to guide *ab initio* protein structure prediction [7].

The idea has already been introduced in the early 1990s [8–11], however, until recently, the accuracy of the predicted contacts was not sufficient to significantly improve structure prediction methods. Pairs of correlated mutations have been calculated using 'local' statistical models, e.g. mutual information scores, which are not able to separate direct from indirect contact information. While direct contacts reflect actual contacts in the protein structure, indirect contacts are false positives that arise from connections through a third residue. These transitive (indirect) pair correlations greatly limit the accuracy of predicted residue-residue contacts [7].

Recently, a substantial increase in prediction accuracy has been achieved by using 'global' statistical models [12–16] that are able to reduce these effects of transitivity by treating pairs of residues dependent on each other. Another important factor for the recent boost in prediction accuracy is the rapid growth of available protein sequences due to advances in DNA sequencing technology [7].

In 2011, it has been shown that the information contained in maximum-entropy derived residue-residue contacts is sufficient to predict protein folds with explicit atomic coordinates quite accurately (Cα-RMSDs of 2.7–4.8Å over at least two-thirds of the protein) using the method EVFold [13]. Since then, a lot of research focused on improving the contact predictions and new methods for residue-residue contact prediction emerge regularly [17–21]. In addition to the initial predictions of mostly soluble proteins [13], predicted contacts from evolutionary information have been used to predict protein-protein complexes [22–24], and the structures of membrane proteins [25,26].

While much effort is put into the improvement of contact predictions, there is also a substantial need to investigate how this information is best exploited in structure prediction. The accuracy of contact predictions is limited by the statistical nature of the prediction methods, distracting sources of co-evolution (e.g. active sites and protein-protein interaction sites), and limited numbers of homologous sequences. Due to the noisy nature of the predicted residue-residue contacts, structure prediction protocols with a high tolerance against incorrect distance restraints are needed.

EVFold uses the CNS molecular dynamics software suite [27,28] for structure prediction. It starts with the fully extended amino-acid sequence and folds the protein by applying standard distance geometry techniques and simulated annealing with bonded and non-bonded potentials [13].

The fragment-based folding algorithm FRAGFOLD [29,30] was used in combination with the contact prediction method PSICOV [17] for *ab initio* structure prediction [31]. The restraints were scored with a square well function with exponential decay.

Michel and coworkers applied the *ab initio* structure prediction protocol of the molecular modeling software suite Rosetta [32] with a smoothed square well restraint scoring function to predict structures within the PconsFold pipeline [33]. A comparison between Rosetta and CNS indicated that with similar contact predictions, models of similar quality were generated [33]. Improvements in structure prediction were mainly credited to improved residue-residue contact predictions obtained with the combined prediction method PconsC [34].

The CONFOLD webserver uses the CNS suite [27,28] for a two-stage modeling approach. Both, restraints derived from predicted contacts and secondary structure, are used and after the initial round of model generation, unsatisfied restraints are filtered out. The method has been shown to be especially powerful when using true contacts [35].

In this work we combine evolutionary information, obtained from predicted residue-residue contacts, with the Resolution-Adapted Structural RECombination approach RASREC [36] (cf. Fig 1). RASREC is an iterative sampling protocol of Rosetta that carries out restraint-guided fragment assembly during six different resampling and refinement stages. The main idea behind the protocol is the iterative recombination of frequently reoccurring structural features and promising strand pairings. It has been shown previously that RASREC requires less data and is more robust against incorrect distance restraints than the standard Rosetta prediction protocol [5,6,36]. These properties make RASREC the ideal starting point for developing a protocol for structure prediction guided by evolutionary restraints, the latter containing a fraction of incorrectly predicted protein-protein contacts.

For our method, evolutionary information was added to the RASREC protocol by translating the top scoring residue-residue contact pairs into sigmoidal distance restraints. This initial RASREC prediction was furthermore followed by an additional refinement run using distance information from both the previous run and the predicted residue-residue contacts.

To investigate the performance of our method, we carried out a benchmark on 28 globular proteins using state-of-the-art contact predictions (generated using a pseudo-likelihood maximization approach). To test the impact of increasing numbers of false restraints, we

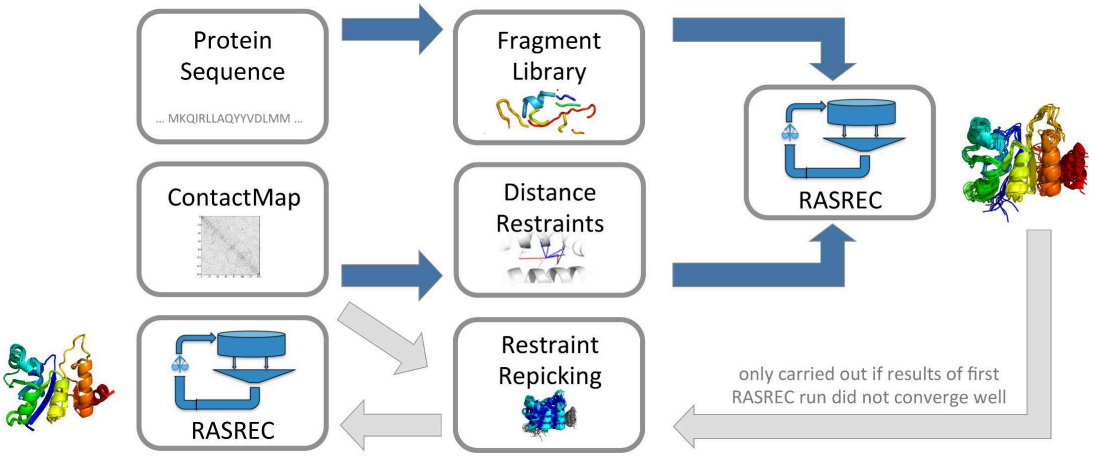

Fig 1. Protocol pipeline. Our protocol consists of one core step (blue) and an optional refinement step (light grey). Core step: The top scoring residue pairs of a predicted contact map are translated into distance restraints and used for structure prediction in combination with the RASREC protocol. Refinement step: Restraints are repicked from the results of the core step and used in a second RASREC run combined with additional contact map restraints.

doi:10.1371/journal.pcbi.1004661.g001

additionally predicted the structures of a smaller benchmark set using less accurate residue-residue contact predictions (calculated with a mean-field direct coupling analysis).

In this manuscript we report the results of the benchmark using both types of residue-residue contact predictions and contrast the performance of our protocol with results obtained by the EVFold webserver using identical contact predictions. We furthermore illustrate the contribution of the optional refinement run to the final results of our method and investigate the benefits of including predicted residue-residue contacts to the standard RASREC sampling method in general.

Materials and Methods

Datasets

We have benchmarked our protocol on two previously published datasets, namely the 14 globular proteins from the EVFold benchmark set published in [13] and the 14 globular proteins used as test set for developing Pconsfold [33]. The structures vary in sequence length between 58 and 247 residues and cover the three structured CATH classes i.e. mainly α, mainly β, and mixed α/β. An overview of all targets in our benchmark set can be found in Table 1.

In case of the EVFold benchmark set, the protein sequences of the models published in [13] (available at http://evfold.org/evfold-web/datasets.do) were used to enable a direct comparison between EVFold and our method. For the Pconsfold dataset, the sequences deposited in the RCSB Protein Data Bank [37] were used. FASTA sequences for all targets in our benchmark set are available in S2 File.

Contact prediction

We used two sets of contact predictions, generated with the PLM (pseudo-likelihood maximization) and DI (direct information/ mean field approximation) scoring method, respectively.

Table 1. Benchmark set. Positive predictive values (PPV) have been calculated for two restraint sets (calculated with the pseudo-likelihood maximization approach (PLM) and direct coupling analysis (DI), respectively) by comparing the potential contacts to the actual Cβ-Cβ distances in the reference structure with a cutoff of 8 Å.

Benchmark set	Target	Fold (CATH)	Model Size	# Restraints	PPV Distance Restraints	
					PLM	DI
EVFold benchmark set	2hda	β	58	50	0.78	0.52
	5pti	few ss	63	60	0.67	0.65
	1wvn	α/β	73	70	0.64	0.39
	1g2e	α/β	81	80	0.84	0.65
	1odd	α	87	80	0.54	0.28
	1rqm	α/β	105	100	0.61	0.55
	1r9h	α/β	105	100	0.79	0.64
	2o72	β	110	110	0.76	0.65
	1bkr	α	117	110	0.45	0.33
	2it6	α/β	117	110	0.68	0.49
	1e6k	α/β	124	120	0.73	0.61
	1f21	α/β	147	140	0.69	0.44
	5p21	α/β	170	170	0.48	0.48
	3tgi	β	226	220	0.79	0.50
Pconsfold benchmark set	1jo8	β	58	50	0.80	-
	1bdo	β	80	80	0.51	-
	1fqt	β	112	110	0.85	-
	2cua	β	135	130	0.57	-
	1vp6	β	138	130	0.66	-
	1a3a	α/β	148	140	0.79	-
	1ihz	β	149	140	0.78	-
	1jwq	α/β	179	180	0.65	-
	1im5	α/β	180	180	0.72	-
	1atz	α/β	189	180	0.81	-
	1chd	α/β	203	200	0.81	-
	1hdo	α/β	206	200	0.43	-
	1o1z	α/β	234	230	0.71	-
	1tqh	α/β	247	240	0.68	-

doi:10.1371/journal.pcbi.1004661.t001

The PLM method uses a pseudo-likelihood maximization approach [19,38] for finding the maximum entropy set of correlated interactions. This approach is one of the most accurate prediction methods to date [20]. Residue contacts based on this scoring method were predicted for the entire benchmark set using the EVFold webserver (available at http://www.evfold.org/) with default parameters. EVFold returns, along with the predicted 3D models, a list of all-by-all residue pairings computed with EVcouplings-PLM. Restraints based on these contact predictions will be referred to as PLM-restraints in the remainder of this manuscript.

The DI method, as published in [13], uses a less accurate mean field approximation. The contact predictions used in [13] are provided as downloadable content on the EVFold website. Restraints extracted from these contact predictions will be referred to as DI-restraints in the remainder of this manuscript.

In EVFold, contact predictions are further processed by applying several filters based on residue conservation, secondary structure prediction and cysteine pairings [13] before being translated to distance constraints. In contrast, we used the predicted contacts without any

filters to see how much information they provide by themselves. For both restraint sets, the predicted contacts were ordered by their assigned confidence score and the L top-ranked contacts with a minimum distance of 5 residues were selected (with L being the length of the protein sequence rounded down to the nearest multiple of 10). Unless mentioned otherwise, predicted residue contacts refer to these L top-ranked contacts.

The accuracy of the contact predictions was assessed in form of the positive predictive value (PPV) by comparing a potential contact to the actual Cβ-Cβ distance in the reference structure. A contact was counted as a true positive if the Cβ-Cβ distance in the native structure is ≤ 8 Å.

Structure generation with RASREC

To generate the three-dimensional structures, we used the RASREC protocol as described previously [36]. For objective benchmarking and mimicking real application cases, homologous structures (with a PSI-BLAST [39] e-score < 0.05) were excluded in creating the fragment library of each target.

Instead of using experimentally derived distance restraints, we used the predicted residue contacts as source of residue-residue distance information. For this purpose, the L top scoring contact predictions were translated into Rosetta specific Cβ-Cβ distance restraints as described below.

To account for the fact that the predicted contacts might be noisy and might contain a varying number of incorrectly predicted contacts (i.e. false positives), the distance restraints were scored with a shallow sigmoidal potential [23]:

$$f_{\text{Sigmoid}}(x) = \frac{1}{1 + e^{-m\cdot(x-x_0)}} - 0.5 \text{ with } x_0 = 8.0 \text{ and } m = 1 \quad (1)$$

Satisfied distance restraints (Cβ-Cβ distance ≤ 8 Å) add a bonus to the final energy term, while unsatisfied distance restraints are ignored. This greatly reduces the influence of incorrectly predicted residue contacts and the structure prediction will not be misguided. Using bounded restraints in this step instead, i.e. punishing each violated restraint with an energy penalty, often resulted in misfolded and unconverged structures in initial test runs.

As in [36],the pool size of RASREC, specifying the number of best scoring models maintained during each iteration stage, was set to 500. The total number of models generated during a RASREC run depends on how fast the different iteration stages terminate and cannot be directly controlled. For the EVFold benchmark set, the total number of generated models per target ranges from 13,000 to 65,000. For a detailed description of all options and parameters used, please refer to S1 Supporting Information and the Protocol Capture in S1 Text and S1 File.

RASREC requires substantial computer resources. For the EVFold benchmark set, the average computation time was ~2600 cpu hours using 2.6 GHz AMD Opteron processors, see Fig A in S1 Supporting Information. The computation time is dependent on several factors, which include sequence length, fold complexity, and instructiveness of the restraints.

Optional refinement step. If the results of the first RASREC run did not converge in all parts of the protein structure (fraction of converged residues $< 90\%$ in the 30 lowest energy models), an optional refinement run (ReRASREC) was carried out to increase both accuracy and convergence. For this purpose, converged substructures from the initial RASREC run were rebuilt and non-converged regions were refined using additional contact information:

To easily re-establish the converged core of the initial RASREC run, we derived distance restraints for the converged regions in the following way: Distances between all Cα-Cα pairs were calculated, and those that are short-range ($\leq$ 8Å) and have a standard deviation (SD)

below 1Å in the 30 low-energy RASREC models were kept. These converged distances were enforced during ReRASREC using the strict bounded potential as in [6]:

$$f_{\text{Bounded}}(x) = \begin{cases} \left(\frac{x-lb}{sd}\right)^2 \text{ for } x < lb \\ 0 \text{ for } lb \leq x \leq ub \\ \left(\frac{x-ub}{sd}\right)^2 \text{ for } ub < x \leq ub + 0.5 * sd \\ \frac{1}{sd}(x - (ub + 0.5 * sd)) + \left(\frac{0.5 * sd}{sd}\right)^2 \text{ for } x > ub + 0.5 * sd \end{cases} \text{with } sd = 1 \quad (2)$$

To reflect the average distance *d* in the converged region, the lower bound *lb* was set to *(d–1)* and the upper bound *ub* to *(d+1)*.

The structural models from the first RASREC run allowed us to select additional low-ranked predictions from the contact map: Prior to having any structural knowledge we could only choose contact predictions with very high confidence in the attempt to avoid frustrating the calculations with too many erroneous restraints. In the second iteration however, we were able to use the lowest-energy models of the first RASREC run to filter out contact predictions that clearly disagree with these models. Hence, lower-confidence predictions could be incorporated as well. To refine the unconverged regions (residue-residue distance, SD > 1 Å in 30 low-energy structures), we therefore chose additional residue-residue pairings from the predicted contact map that affect these regions and do not totally disagree (i.e. are short range with an average distance d ≤ 8 Å) with the lowest-energy models of the first run. The restraints were scored with a wide bounded potential with lower and upper bound set to 1.5 Å and 8 Å, respectively. This wide range was chosen to allow these regions to adapt to energetically favorable conformations. To reduce the influence of potentially incorrect restraints in this set, we furthermore combined random pairs into ambiguous restraints [6]. For each model new random pairs were generated.

Identifying unsuccessful predictions by backbone convergence. For "blind" structure predictions it is important to discern whether the final result of a prediction method is reliable or not. Here, we used the backbone convergence of the 30 lowest-energy models as a criterion to decide whether a prediction is classified as successful or not. The backbone of a residue was considered converged if the corresponding Cα-atoms in the 30 lowest-energy structures had less than 2 Å coordinate variability. If less than half of the residues of the 30 lowest-energy structures converged, a prediction was regarded as unsuccessful. In those cases, our protocol was not able to find a consistent low energy state.

Model ranking. The models predicted by RASREC were ranked according to their resulting Rosetta Energy Units (REU). Distance restraints were included with a weight of 0.1 in this full-atom energy function. The ensemble of the 10 lowest-energy structures is considered as the final result of our protocol. Therefore, if not stated otherwise, the metrics used for performance evaluation are averaged over the 10 lowest-energy structures.

Structure prediction with EVFold

The EVFold webserver offers to directly fold the protein of interest based on its predicted residue-residue contacts. Structure prediction is accomplished using the CNS software [27,28] with the protocol described in [13]. The webserver predicts structures for different amounts of filtered restraints, starting with only a few and increasing to *L* in 10 steps with *L* being the domain length. As output, the 3D coordinates of all 50 predicted structures are provided. We

used the web interface to generate the models along with the predictions based on the PLM approach. These models are referred to as EVFold-PLM models. Further, we used the structures published in [13] (available at http://evfold.org/evfold-web/datasets.do), which are based on the residue-residue contact predictions with the less accurate DI approach and are referred to as EVFold-DI models.

Model ranking. EVFold ranks its models with a score based on inherent properties and extent of constraint satisfaction. We consider the single top-ranked structure as the final result of EVFold, irrespective of the number of distance restraints used. In addition, results averaged over the 10 top-ranked structures can be found in Table C in S1 Supporting Information.

Metrics used for performance evaluation

To evaluate the performance of our method, several different metrics were used: 1) Cα-RMSD calculated over all residues present in the reference structure (RMSD), 2) Cα-RMSD calculated over all residues in secondary structural elements in the crystal structure as assigned by Stride [40] called $RMSD_{SSE}$, and 3) TM-Score [41] over all Cα-atoms in the reference structure. The template modeling score (TM-Score) evaluates the global fold similarity and is less sensitive to local structural variations than the RMSD. It ranges from 0 (random similarity) to 1 (perfect similarity) [41].

In contrast to e.g. RMSD values calculated with PyMOL [42], which excludes outliers in a series of refinement cycles, these three metrics are easily reproducible and consider the same residues for each model evaluated.

Results and Discussion

We have developed a protocol (Fig 1) that combines RASREC with evolutionary sequence information in form of predicted residue-residue contacts for accurate protein structure prediction. We benchmarked this protocol on a diverse set of 28 globular proteins and compared its results with the ones from the EVFold web server, to our knowledge one of the best methods currently available.

Models generated with ReRASREC have higher accuracies

Fig 2 shows the performance of our protocol (ReRASREC-PLM) compared to the one of the EVFold web server (EVFold-PLM) on the basis of three different metrics. Our protocol converged (fraction of converged residues > 0.5 in the 30 low-energy structures) for 26 out of the 28 targets and correctly predicted the fold for each of the converged targets (TMscore > 0.5 or RMSD < 5Å). For the majority of the benchmark set, the final models were of high structural accuracy resulting in an average TM-score of 0.74, an average RMSD of 4.4 Å, and an average $RMSD_{SSE}$ of 3.3 Å over all 26 converged targets.

The overall performance of our protocol was significantly higher than that of EVFold-PLM using identical contact predictions (however not necessarily identical distance restraints, see section Structure Prediction with EVFold). With an average TM-score of 0.72 over the entire benchmark set, ReRASREC-PLM lead to an improvement of 0.17 when compared to EVFold-PLM, whose average TM-score was only 0.55. ReRASREC-PLM furthermore increased the number of targets with a TM-score > 0.7 from 6 to 20. In terms of RMSD and $RMSD_{SSE}$, using our method lead to an average improvement from 7.3 Å to 4.9 Å and from 5.7 Å to 3.7 Å respectively. Moreover, EVFold-PLM failed to predict the correct fold for 6 out of 28 targets (TM-score < 0.5 and RMSD > 5Å) while our protocol predicted very accurate models (TM-Score 0.62) with correct folds for all of these targets.

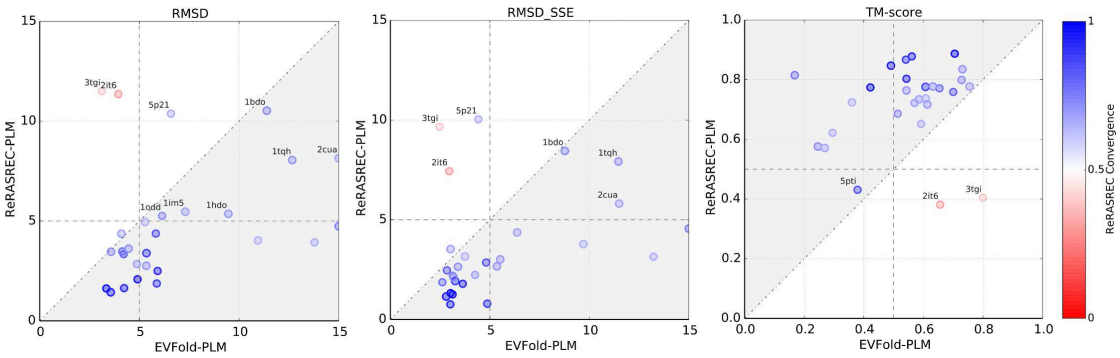

Fig 2. Comparison between ReRASREC-PLM and EVFold-PLM. In case of ReRASREC-PLM, the similarity measures are averaged over the 10 lowest-energy models, while for EVFold-PLM the single top ranked model is evaluated. The color represents the fraction of converged residues in the 30 lowest-energy models of ReRASREC-PLM. The gray areas indicate an improvement of ReRASREC-PLM over EVFold-PLM.

doi:10.1371/journal.pcbi.1004661.g002

Based on our backbone convergence criteria (see Materials and Methods) our protocol failed for targets 2it6 and 3tgi. Both targets consist of long loop regions (fraction of secondary structural content is only 0.54 and 0.37 respectively) and are therefore challenging for RASREC as it is mainly focusing on the recombination of reoccurring structural features such as secondary structure elements.

Fig 2 reveals that predictions for two converged targets, namely 5p21 and 1bdo, resulted in models with an RMSD $>$ 10 Å. The TM-Score is however above 0.5 in both cases, i.e. 0.65 and 0.58, respectively, showing that the majority of the protein structure was predicted correctly. The good accordance between the top-scoring models and the corresponding native structures can furthermore be seen in Fig B in S1 Supporting Information.

ReRASREC-PLM was not only able to predict the correct fold for a larger number of targets, but also significantly improved the accuracy within the set of targets with correctly predicted folds. Excluding the 8 targets where either EVFold-PLM (6) or RASREC-PLM (2) had difficulties, ReRASREC-PLM still increased the average TM-Score by 0.18 over EVFold-PLM from 0.60 to 0.78. In terms of RMSD and $RMSD_{SSE}$, RASREC-PLM improved them from 5.6 Å to 3.9 Å and from 4.2 Å to 2.9 Å, respectively.

We also compared the accuracy of ReRASREC-PLM with two other recently published methods (PconsFold [33] and FRAGFOLD [31]) on the subset of targets where each publication reported actual numbers on. We found that, although both methods improve upon EVFold-PLM, ReRASREC-PLM still outperforms both (Table A in S1 Supporting Information).

ReRASREC models have accurate side chains in the protein core

Fig 3 further indicates that the models generated with our protocol do not only have high accuracy in their backbones, but also a high rotamer recovery of core side-chain conformations. A superposition of the lowest-energy model and the corresponding crystal structure of each target can be found in Fig B in S1 Supporting Information.

Table 2 shows that on average 84% of the converged core side chains in the RASREC models are in the same χ_1 rotamer well, and 46% have the same set of rotamer states for all χ angles as the corresponding crystal structures. An analysis of the single top-ranked models of EVFold-PLM and ReRASREC-PLM furthermore shows that ReRASREC-PLM predicts higher numbers

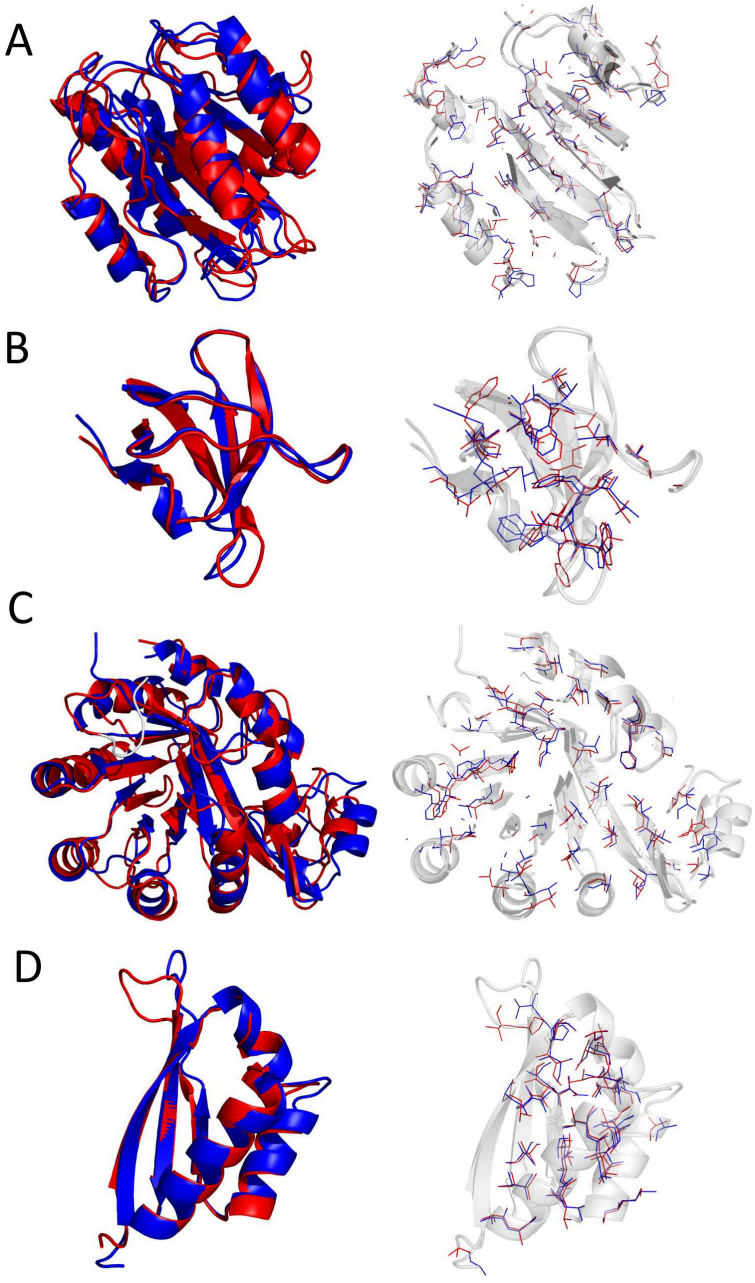
A
B
C
D

Fig 3. Superposition of top ranked models and corresponding crystal structures. Top-energy ReRASREC structures (red) for 1atz (A), 1jo8 (B), 1o1z(C), and 1wvn(D) are superimposed with the corresponding crystal structures (blue). For each target, a cartoon representation of the lowest-energy structure (left) and a close-up showing non-polar side-chains (right) is shown.

doi:10.1371/journal.pcbi.1004661.g003

of buried side chains with native χ_1 romater assignment than EVFold-PLM, see Table B in S1 Supporting Information.

Table 2. Accuracy of sidechain χ1 rotamers in the final ReRASREC models. Buried and converged side chains are selected and their adopted rotamer assignments are compared to those in the reference crystal structure. Alanine and Glycine are excluded from this analysis.

Benchmark set	Target	Number of side chains			Fraction of recovered rotamers	
		buried*	converged & buried**	recovered χ1 ***	χ1 only%	all χ angles~
EVFold benchmark set	1bkr	42	8	7	0.88	0.50
	1e6k	49	20	18	0.90	0.55
	1f21	53	20	19	0.95	0.45
	1g2e	25	11	10	0.91	0.64
	1odd	27	10	9	0.90	0.70
	1r9h	36	8	7	0.88	0.63
	1rqm	42	12	7	0.58	0.33
	1wvn	19	14	13	0.93	0.50
	2hda	16	11	6	0.55	0.27
	2it6	48	7	6	0.86	0.29
	2o72	27	8	7	0.88	0.38
	3tgi	101	27	22	0.81	0.37
	5p21	71	20	19	0.95	0.60
	5pti	14	5	3	0.60	0.20
Pconsfold benchmark set	1a3a	56	19	16	0.84	0.63
	1atz	72	9	8	0.89	0.33
	1bdo	25	10	8	0.80	0.70
	1chd	74	23	19	0.83	0.43
	1fqt	44	21	19	0.90	0.52
	1hdo	84	21	15	0.71	0.38
	1ihz	51	7	6	0.86	0.29
	1im5	68	22	19	0.86	0.32
	1jo8	15	10	8	0.80	0.50
	1jwq	76	17	15	0.88	0.53
	1o1z	99	29	26	0.90	0.34
	1tqh	106	36	32	0.89	0.53
	1vp6	50	19	18	0.95	0.63
	2cua	46	15	11	0.73	0.27
Average	N/A	N/A	N/A	N/A	0.84	0.46

* Side chains that are buried in the reference structure (SASA < 40Å)

** Side chains that are buried (SASA < 40Å) and converged (χ1 angle, SD < 10 degrees in 10 low-energy structures).

*** Subset of converged and buried residues that adopt the same χ1 rotamer state as in the reference structure.

% Ratio of column 2 (correct) and column 1 (converged and buried)

~ Fraction of sidechains in column 1 (converged and buried) for which all side-chain torsion angles adopt the same rotamer state as in the reference structure.

doi:10.1371/journal.pcbi.1004661.t002

ReRASREC is more robust against incorrect distance restraints

It has been shown previously [5,6,36] that RASREC is more robust against incorrect distance restraints than the standard Rosetta *ab initio* protocol. A high tolerance against false positives is of special interest for proteins where only a limited number of homologous sequences are available. In those cases, the fraction of false positives in the corresponding contact predictions is comparably high, hence making structure prediction for standard prediction methods difficult.

To investigate how our protocol performs with an elevated amount of incorrectly predicted residue contacts, we used it in combination with the contact predictions published in [13]. These predictions were generated with the less accurate mean field approach (DI–direct information) and therefore contain an increased number of incorrectly predicted protein contacts as compared to the restraints obtained with the PLM approach (see Table 1). With an average PPV of 0.51, the accuracy of the DI-restraints drops by 0.17 compared to the average PPV of the PLM-restraints.

Given these restraints with a significantly lower accuracy, our protocol was able to converge for 12 out of 14 targets (see Fig C in S1 Supporting Information) and predicted the correct fold for all of the converged targets with an average TM-score of 0.70 and an average RMSD of 4.0 Å (see Table 3). The results obtained with our protocol significantly outperform the top ranked results generated with EVFold using DI-restraints: Using our protocol lead to an increase in average TM-score of 0.17 when compared to the average TM-score of 0.47 of the corresponding EVFold results. In terms of RMSD, the use of ReRASREC-DI improved the prediction from 7.2 Å to 5.6 Å. For 6 targets, the top-ranked EVFold models furthermore displayed the incorrect fold (TM-score < 0.5 and RMSD > 5 Å).

Table 3. Results for the EVFold benchmark set using different methods and different restraint sets. For ReRASREC, the metrics were calculated and averaged over the 10 lowest-energy models while for EVFold, the single top ranked structure was used. For both methods, results generated with both PLM- and DI-restraints are shown. For each double column and target, the 'better' performance is highlighted.

	TM-score				RMSD			
	PLM-restraints		DI-restraints		PLM-restraints		DI-restraints	
Target	ReRASREC-PLM	EVFold-PLM	ReRASREC-DI	EVFold-DI	ReRASREC-PLM	EVFold-PLM	ReRASREC-DI	EVFold-DI
1bkr	**0.62**	0.30	**0.68**	0.29	**3.93**	13.79	**3.67**	13.20
1e6k	**0.89**	0.71	**0.87**	0.63	**1.62**	3.34	**1.78**	4.76
1f21	**0.76**	0.70	**0.59**	0.51	**3.34**	4.21	**6.87**	8.16
1g2e	**0.88**	0.56	**0.84**	0.54	**1.64**	4.23	**1.83**	5.23
1odd	**0.69**	0.51	**0.49**	0.37	**5.26**	6.14	**6.20**	9.40
1r9h	**0.72**	0.57	**0.68**	0.48	**2.84**	4.87	**5.47**	7.19
1rqm	**0.80**	0.54	**0.78**	0.55	**2.50**	5.91	**2.46**	4.72
1wvn	**0.87**	0.54	**0.82**	0.28	**1.87**	5.87	**2.09**	8.21
2hda	**0.77**	0.42	**0.72**	0.40	**2.08**	4.91	**2.47**	6.59
2it6*	0.38	**0.66**	0.38	**0.39**	11.36	**3.94**	10.62	**10.54**
2o72	**0.77**	0.65	**0.69**	0.54	**3.48**	4.14	**4.41**	6.07
3tgi*	0.40	**0.80**	0.19	**0.53**	11.50	**3.12**	20.19	**7.66**
5p21	**0.65**	0.59	0.66	**0.70**	10.38	**6.58**	7.99	**3.64**
5pti	**0.43**	0.38	**0.62**	0.45	**4.37**	5.82	**2.77**	4.75
Mean	**0.69**	0.57	**0.64**	0.47	**4.73**	5.49	**5.63**	7.15

* unconverged targets

doi:10.1371/journal.pcbi.1004661.t003

Using the less accurate DI-restraints had less of an impact on accuracy for ReRASREC than for EVFold; the average TM-score of the EVFold benchmark set decreased by 0.05 and by 0.1 points for ReRASREC and EVFold, respectively (Table 3). While ReRASREC predicted the correct fold for all 12 converged targets with both restraint sets, EVFold increased the number of incorrect folds from 2 to 6 when using the less accurate DI-restraints instead of PLM-restraints.

This suggests that our protocol can predict structures with restraints of mediocre accuracy better than the CNS protocol used by EVFold.

Successful model ranking with full-atom energy function

For realistic application cases the ranking of the predicted structural models is of great importance as it will be the single criterion for selecting the final predicted models. The models generated with our protocol were ranked with the full-atom energy function of Rosetta. All-atom energy functions are very sensitive to correct packing of side chains due to the steep gradient of the Lennard-Jones repulsive term. Correct packing of side chains is hard to achieve, in particular, if the backbone structure is not sufficiently accurate. Selection based on this energy function is therefore only possible if the backbone accuracy is very high.

Fig 4 shows the full-atom energies and RMSD values for each model generated during the different stages of a single RASREC run for one exemplary target. The energy funnel at the low RMSD area shows that the all-atom energy function is able to discriminate between correct and incorrect structural models.

This observation is further reinforced by comparing the lowest-RMSD models to the lowest-energy models (Table C in S1 Supporting Information): The average TM-score of the lowest-RMSD models is with 0.77 only 0.05 higher than the one of the lowest-energy models generated by ReRASREC with 0.72.

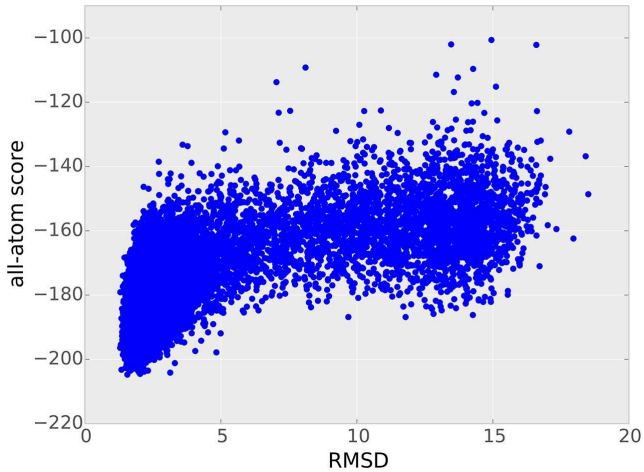

Fig 4. RMSDs and all-atom scores of each structure generated during a single RASREC run. All structures generated during the initial RASREC run of target 1e6k are shown. A simple structural refinement was carried out for each model to convert the centroid models (the first four RASREC stages use the Rosetta low-resolution energy) into full atom models with packed side chains.

doi:10.1371/journal.pcbi.1004661.g004

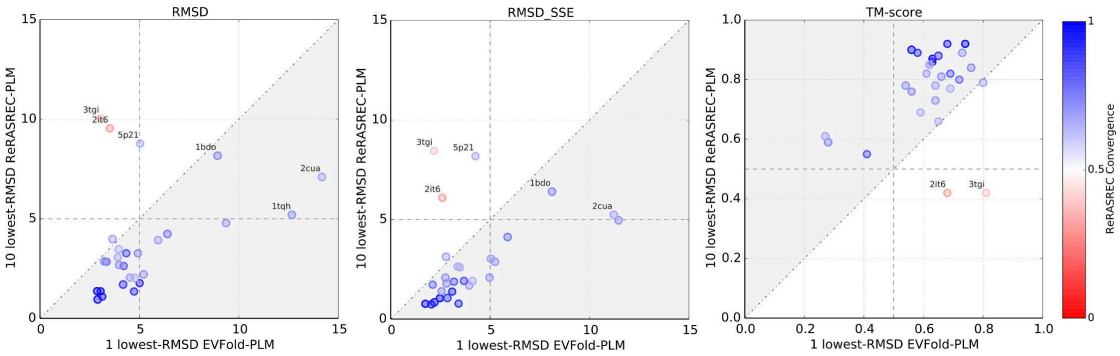

Fig 5. Comparison of ReRASREC's lowest-RMSD models to the lowest-RMSD models generated with EVFold. The single most accurate EVFold structure (lowest RMSD) has been selected among all 50 provided models and is compared to the average of the 10 models of a RASREC refinement run with lowest RMSD.The color represents the fraction of converged residues in the 30 lowest energy models of ReRASREC-PLM. Gray shaded areas indicate an improvement of ReRASREC-PLM over EVFold-PLM.

doi:10.1371/journal.pcbi.1004661.g005

In contrast, EVFold ranks its models based on inherent geometrical properties and constraint satisfaction. Choosing the lowest-RMSD models instead of the top ranked ones increases the average TM-score from 0.55 to 0.62 and improves the RMSD from 7.3 Å to 5.2 Å.

Investigating these results more closely, one can observe that the top ranked structures of EVFold-PLM adapt the incorrect fold (RMSD $>$ 5 Å and TM-score $<$ 0.5) for two targets, namely 1bkr and 1o1z, although models with correct topologies were generated as well. For those two targets, the ranking of EVFold-PLM therefore fails. For ReRASREC-PLM using the full-atom score function, no such discrepancy was observed.

Gain in accuracy due to high quality structural models

In this section, we analyze the accuracy of the models generated by EVFold-PLM and ReRASREC-PLM irrespective of their ranking schemes. Therefore, we have compared the most accurate models (average of the 10 lowest-RMSD models) of ReRASREC to the single lowest-RMSD models generated by the EVFold web server within its 50 reported models. As shown in Fig 5, the ReRASREC models with lowest RMSD outperform the lowest-RMSD models of EVFold for each converged target. Overall, the ReRASREC models show an increase in TM-score of 0.15 when compared to the average TM-score of 0.62 of the single most accurate EVFold models.

We have shown in the previous section that the difference in accuracy between the lowest-energy and lowest-RMSD models of ReRASREC-PLM is small. The lowest energy models of ReRASREC-PLM are therefore more accurate than any models obtained with the EVFold web-server (see Fig D in S1 Supporting Information). On average, the lowest-energy models of ReRASREC-PLM lead to an increase in TM-score of 0.1 when compared to the TM-score of 0.62 of the single lowest-RMSD models of EVFold-PLM. This shows that our method generates models of higher structural quality than EVFold-PLM.

Refinement run leads to small improvements in model accuracy

If the backbone of the first RASREC run did not converge within 2 Å for over 90 percent of the residues, a refinement run (see Materials and Methods) was carried out. To see to what extent

the refinement run contributes to the final performance of our protocol, we compared the results of the initial RASREC run to the results obtained after the refinement run (ReRASREC).

Fig 6A and 6B show that the accuracy of the top ten scoring models after the refinement run did not significantly improve. However, Fig 6C indicates that the pairwise RMSD between all models in the ensemble of the 10 lowest-energy structures decreased by up to 1.4 Å after applying the refinement run, indicating better convergence. On average, the pairwise RMSD decreased by 0.5 Å. In addition, **Fig 6D** plots the average RMSD of the 10 lowest-energy models against their pairwise RMSD for both RASREC and ReRASREC. In both cases, a similar correlation between RMSD and pairwise RMSD can be observed. This shows that the refinement run does not lead to an artificial over-convergence but that the relation between both, as explored by RASREC individually, is kept.

This comparison shows that while the models have high accuracies after the initial RASREC run, the refinement run improves the overall prediction by increasing the precision and convergence of the final models.

Convergence predicts accuracy

Fig 6D shows that there is a reasonable correlation between the pairwise RMSD and the overall performance of each target (pearson correlation coefficient of 0.83 and 0.73 for RASREC and ReRASREC respectively), meaning that low pairwise RMSD values correlate with low RMSD values and vice versa. The same trend can be observed when relating the backbone convergence (as defined previously) of a prediction to its performance, see Fig E in S1 Supporting Information: High backbone convergence corresponds to low RMSD values with a pearson correlation coefficient of -0.77. These strong correlations indicate that the accuracy of our final models can be predicted by their convergence. Highly converged structures (low pairwise RMSD) indicate an accurate prediction while a highly diverse ensemble suggests that the prediction is incorrect. This observation further reinforces our choice deeming predictions with a convergence lower than 50% as unsuccessful.

Increase in prediction accuracy due to residue-residue contact information

To identify to what extent the RASREC protocol benefits from residue-residue contact information, we have compared RASREC runs without evolutionary information to RASREC runs including them in form of distance restraints for the 14 proteins of the EVFold benchmark set. For this test, we considered the results after a single RASREC run without the optional refinement step. As shown in Fig 7, without the use of evolutionary contact information, RASREC only predicted the fold of 3 out of 14 proteins correctly (TM Score > 0.5 or RMSD < 5Å) with an average TM-score of 0.41. However, if restraints derived from predicted residue-residue contacts were included, RASREC improved the coordinate accuracy for all targets of the benchmark set significantly, yielding an average TM-score over all 14 targets of 0.69. This shows that the additional data provided by the predicted residue-residue contacts enables RASREC to predict models in a near-native conformation, which would not be possible otherwise.

To investigate to what extend the RASREC protocol uses the available contact information, we compared the fraction of satisfied restraints (PPV), i.e. Cβ-Cβ distance ≤ 8 Å, in the top-scoring models of our protocol and the native structure (Fig F in S1 Supporting Information). On average, the fraction of satisfied restraints in the top-scoring models after the initial RASREC run (0.72) is very similar to the one of the native models (0.69). Overall, the RASREC models satisfy 88% of all restraints that are satisfied in the native structures, see Table D in S1 Supporting Information. RASREC furthermore correctly violates 63% of the incorrect distance

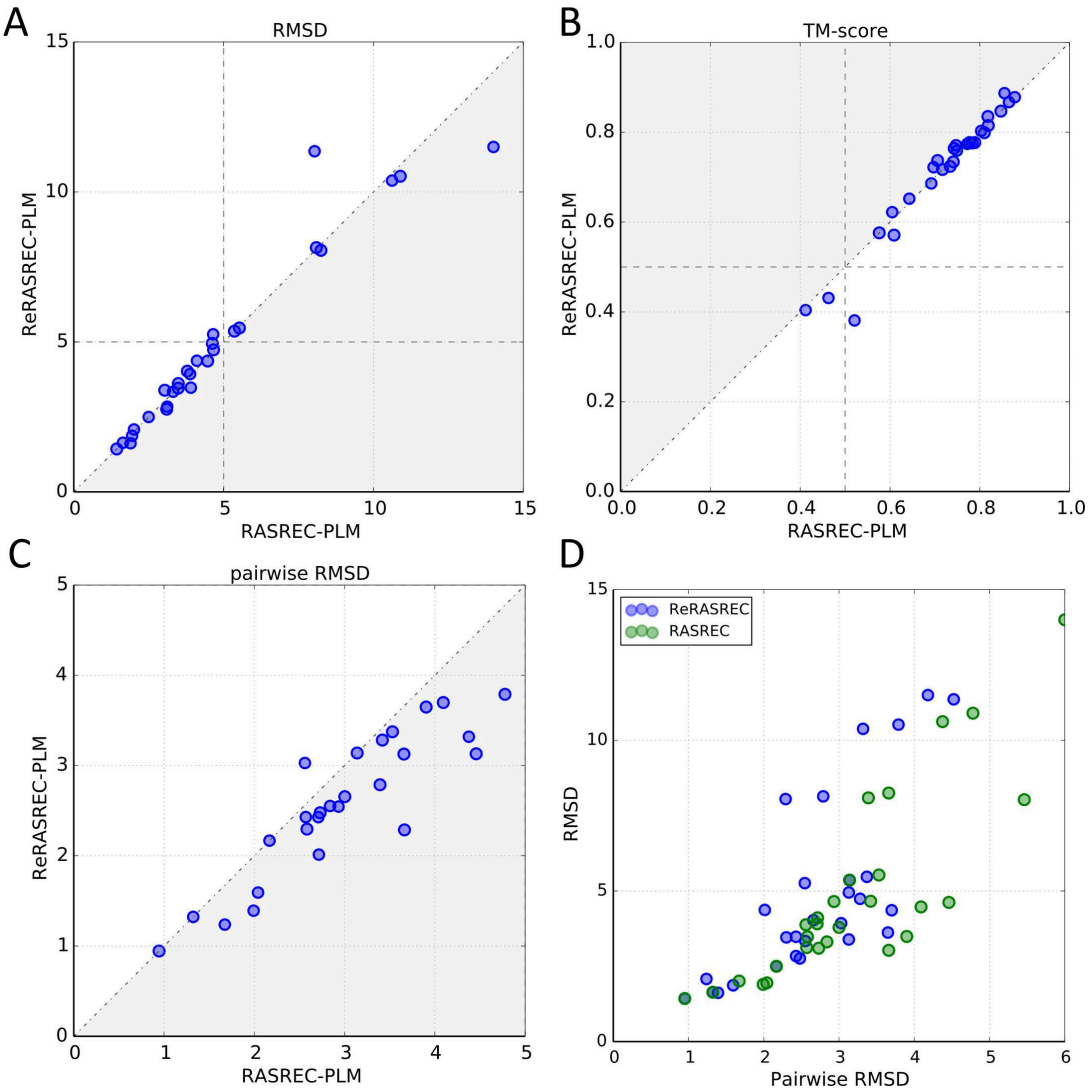

Fig 6. Comparison between initial RASREC results (RASREC-PLM) and refinement results (ReRASREC-PLM). A) RMSD and B) TM-scores of the 10 lowest-energy models of RASREC-PLM and ReRASREC-PLM C) Averaged pairwise RMSD of 10 lowest-energy models in ReRASREC-PLM and RASREC-PLM D) Average RMSD plotted against the average pairwise RMSD of the 10 lowest-energy models for both RASREC-PLM and ReRASREC-PLM.

doi:10.1371/journal.pcbi.1004661.g006

restraints. The good correspondence between the PPVs on the native structure and the RASREC models, as well as the large fraction of satisfied "correct" restraints shows that RASREC is

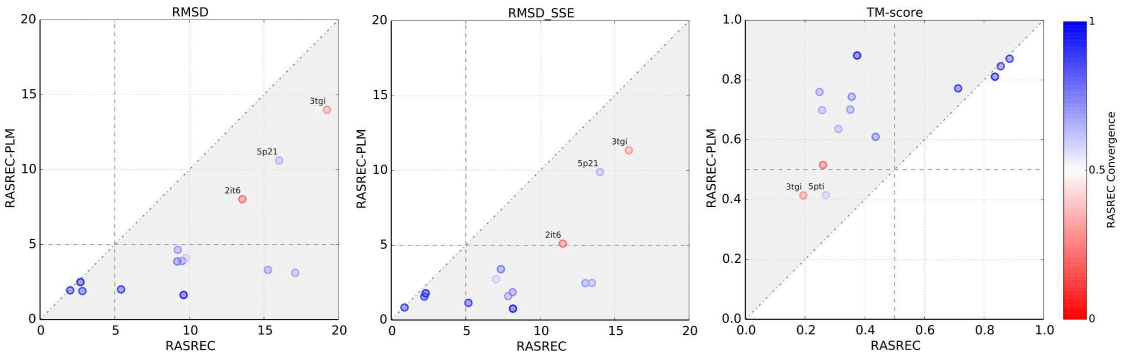

Fig 7. Comparison between RASREC runs without using contact information (RASREC) and RASREC runs using contacts predicted with the PLM approach (RASREC-PLM). For both methods, a single RASREC run without the optional refinement was carried out and the ensemble of the 10 lowest-energy models was considered as the final result. The color represents the fraction of converged residues in the 30 lowest energy models of RASREC-PLM.

doi:10.1371/journal.pcbi.1004661.g007

able to efficiently use the provided contact information. However, ignoring a larger amount of incorrect distance restraints might improve the prediction even further.

Comparing the PPVs, calculated for the restraints used by EVFold, on the top-ranked EVFold models and the native structures suggests that EVFold does not use the provided contact information as well as RASREC, see Fig F in S1 Supporting Information.

Conclusions

In this study, we demonstrated that RASREC combined with evolutionary information is a powerful tool to predict the structures of globular proteins with high accuracy. Tested on a benchmark set of 28 globular proteins, we showed that our protocol is able to outperform latest state-of-the-art methods by predicting structures to higher accuracies for the majority of the benchmark set.

We further showed that the combination of improved sampling and high error tolerance of RASREC enables structure prediction in cases where the accuracy of predicted contacts is comparatively low, e.g. dropping below 50 percent. Robustness against erroneous distance restraints is of special interest for proteins for which only a limited amount of homologous sequences are known. The accuracy of residue-residue contact prediction is highly dependent on the number of available sequences in the multiple sequence alignment. For multiple sequence alignments with a small number of sequences, the accuracy is in general too low to significantly improve structure prediction using standard prediction protocols. We find that our protocol is able to more efficiently use the sparse information contained in contact predictions with low accuracy, due to the error robustness and iterative sampling strategy of the underlying RASREC algorithm. Our protocol should therefore be able to predict accurate models in cases where other currently published methods would most likely fail to predict the correct fold.

In addition, we have shown that integrating evolutionary information into the RASREC protocol is essential for accurate protein structure prediction for 9 out of 12 proteins in the EVFold benchmark set. Even adding contact predictions with accuracies as low as 45% can be sufficient to predict high resolution models that would not be possible using RASREC alone.

The optional refinement run improves the prediction by increasing the precision of the final models. Future work focusing on this step might further increase accuracy and convergence of the final models.

Overall, we have shown how evolutionary information can be efficiently used for predicting accurate protein structures. The rapid growth of sequence information and the current advances in statistical sequence analysis have made protein structure prediction using evolutionary information highly relevant. Finding a way to reliably and efficiently use the distance information contained in multiple sequence alignments will be a first step to fill the increasing gap between the large number of known protein sequences and the significantly smaller number of known protein structures.

Supporting Information

S1 Supporting Information. This Supporting Information file (PDF) contains supporting Figs A-F, Tables A-D, and Methods A-C. Fig A, Computational expense for the initial RASREC run. Fig B, Lowest-Energy ReRASREC-PLM Structures. Fig C, Comparison of ReRASREC-DI and EVFold-DI. Fig D, Comparison of top ranked ReRASREC models and lowest RMSD EVFold models, Fig E, Analysis of prediction performance and convergence. Fig F, Fraction of satisfied restraints in native structures and top-ranked models. Table A, TM-scores for EVFold, RASREC, PconsFold and FRAGFOLD. Table B, Accuracy of side-chain $\chi 1$ rotamers. Table C, Comparison between top ranked and lowest-RMSD structures. Table D, Restraint classification performance of RASREC. Method A, Contact Prediction and Restraint File Generation. Method B, Structure Prediciton with the RASREC protocol. Method C, Refinement with Rasrec.
(PDF)

S1 Text. Protocol capture. This protocol capture describes the steps necessary to reproduce the results presented in the manuscript "Combining evolutionary information and an iterative sampling strategy for accurate protein structure prediction". Exemplary input files and scripts to carry out the steps outlined in this protocol capture as well as exemplary output files are provided in S1 File. For simplification, we only describe structure prediction using our protocol for target 1wvn in this protocol capture. The supplementary materials are also included with Rosetta under the directory "Rosetta/demos/protocol_capture/2015/ rasrec_evolutionary_restraints"
(PDF)

S1 File. Files for protocol capture. Input files for target 1wvn and scripts to carry out the steps outlined in the protocol capture in S1 Text as well as exemplary output files are provided in this attachement. The supplementary materials are also included with Rosetta under the directory "Rosetta/demos/protocol_capture/2015/ rasrec_evolutionary_restraints"
(ZIP)

S2 File. Dataset. FASTA sequences and PLM contact predictions for all targets of the benchmark set.
(ZIP)

Acknowledgments

The authors thank Johannes Soeding, Stefan Seemayer and Hetu Kamisetty for inspiring discussions. We also thank the John von Neumann Institute for Computing (NIC) for access to the JUROPA supercomputer at Jülich Supercomputing Centre (JSC).

Author Contributions

Conceived and designed the experiments: TB OFL. Performed the experiments: TB. Analyzed the data: TB. Contributed reagents/materials/analysis tools: TB OFL. Wrote the paper: TB OFL. Protocol testing, documentation review and helpful comments on the manuscript: JKL.

References

1. Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181: 223–230. PMID: 4124164
2. Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, Kim DE, et al. (2005) Free modeling with Rosetta in CASP6. Proteins 61 Suppl 7: 128–134. PMID: 16187354
3. Bradley P, Misura KM, Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. Science 309: 1868–1871. PMID: 16166519
4. Kim DE, Blum B, Bradley P, Baker D (2009) Sampling bottlenecks in de novo protein structure prediction. J Mol Biol 393: 249–260. doi: 10.1016/j.jmb.2009.07.063 PMID: 19646450
5. Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini J, et al. (2010) NMR structure determination for larger proteins using backbone-only data. Science 327: 1014–1018. doi: 10.1126/science.1183649 PMID: 20133520
6. Lange OF, Rossi P, Sgourakis NG, Song Y, Lee HW, Aramini JM, et al. (2012) Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. Proc Natl Acad Sci U S A 109: 10873–10878. doi: 10.1073/pnas.1203013109 PMID: 22733734
7. Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. Nat Biotechnol 30: 1072–1080. doi: 10.1038/nbt.2419 PMID: 23138306
8. Gobel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. Proteins 18: 309–317. PMID: 8208723
9. Hatrick K, Taylor WR (1994) Sequence conservation and correlation measures in protein structure prediction. Comput Chem 18: 245–249. PMID: 16649265
10. Neher E (1994) How frequent are correlated changes in families of protein sequences? Proc Natl Acad Sci U S A 91: 98–102. PMID: 8278414
11. Shindyalov IN, Kolchanov NA, Sander C (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng 7: 349–358. PMID: 8177884
12. Burger L, van Nimwegen E (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS Comput Biol 6: e1000633. doi: 10.1371/journal.pcbi.1000633 PMID: 20052271
13. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. (2011) Protein 3D structure computed from evolutionary sequence variation. PLoS One 6: e28766. doi: 10.1371/journal.pone.0028766 PMID: 22163331
14. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci U S A 108: E1293–1301. doi: 10.1073/pnas.1111471108 PMID: 22106262
15. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci U S A 106: 67–72. doi: 10.1073/pnas.0805923106 PMID: 19116270
16. Lapedes A GB, Jarzynski C (2012) Using sequence alignments to predict protein structure and stability with high accuracy. ArXiv e-prints.
17. Jones DT, Buchan DW, Cozzetto D, Pontil M (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28: 184–190. doi: 10.1093/bioinformatics/btr638 PMID: 22101153
18. Kamisetty H, Ovchinnikov S, Baker D (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc Natl Acad Sci U S A 110: 15674–15679. doi: 10.1073/pnas.1314045110 PMID: 24009338
19. Ekeberg M, Lovkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys Rev E Stat Nonlin Soft Matter Phys 87: 012707. PMID: 23410359
20. Seemayer S, Gruber M, Soding J (2014) CCMpred—fast and precise prediction of protein residue-residue contacts from correlated mutations. Bioinformatics 30: 3128–3130. doi: 10.1093/bioinformatics/btu500 PMID: 25064567

21. Skwark MJ, Raimondi D, Michel M, Elofsson A (2014) Improved contact predictions using the recognition of protein like contact patterns. PLoS Comput Biol 10: e1003889. doi: 10.1371/journal.pcbi.1003889 PMID: 25375897

22. Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H (2009) High-resolution protein complexes from integrating genomic information with molecular simulation. Proc Natl Acad Sci U S A 106: 22124–22129. doi: 10.1073/pnas.0912100106 PMID: 20018738

23. Ovchinnikov S, Kamisetty H, Baker D (2014) Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. Elife 3: e02030. doi: 10.7554/eLife.02030 PMID: 24842992

24. Hopf TA, Scharfe CP, Rodrigues JP, Green AG, Kohlbacher O, Sander C, et al. (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. Elife 3.

25. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS (2012) Three-dimensional structures of membrane proteins from genomic sequencing. Cell 149: 1607–1621. doi: 10.1016/j.cell.2012.04.012 PMID: 22579045

26. Hayat S, Sander C, Marks DS, Elofsson A (2015) All-atom 3D structure prediction of transmembrane beta-barrel proteins from sequences. Proc Natl Acad Sci U S A 112: 5413–5418. doi: 10.1073/pnas.1419956112 PMID: 25858953

27. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, et al. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 54: 905–921. PMID: 9757107

28. Brunger AT (2007) Version 1.2 of the Crystallography and NMR system. Nat Protoc 2: 2728–2733. PMID: 18007608

29. Jones DT (2001) Predicting novel protein folds by using FRAGFOLD. Proteins Suppl 5: 127–132.

30. Jones DT, Bryson K, Coleman A, McGuffin LJ, Sadowski MI, Sodhi JS, et al. (2005) Prediction of novel and analogous folds using fragment assembly and fold recognition. Proteins 61 Suppl 7: 143–151. PMID: 16187356

31. Kosciolek T, Jones DT (2014) De novo structure prediction of globular proteins aided by sequence variation-derived contacts. PLoS One 9: e92197. doi: 10.1371/journal.pone.0092197 PMID: 24637808

32. Rohl CA, Strauss CE, Misura KM, Baker D (2004) Protein structure prediction using Rosetta. Methods Enzymol 383: 66–93. PMID: 15063647

33. Michel M, Hayat S, Skwark MJ, Sander C, Marks DS, Elofsson A (2014) PconsFold: improved contact predictions improve protein models. Bioinformatics 30: i482–488. doi: 10.1093/bioinformatics/btu458 PMID: 25161237

34. Skwark MJ, Abdel-Rehim A, Elofsson A (2013) PconsC: combination of direct information methods and alignments improves contact prediction. Bioinformatics 29: 1815–1816. doi: 10.1093/bioinformatics/btt259 PMID: 23658418

35. Adhikari B, Bhattacharya D, Cao R, Cheng J (2015) CONFOLD: Residue-residue contact-guided ab initio protein folding. Proteins 83: 1436–1449. doi: 10.1002/prot.24829 PMID: 25974172

36. Lange OF, Baker D (2012) Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. Proteins 80: 884–895. PMID: 22423358

37. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, et al. (2002) The Protein Data Bank. Acta Crystallogr D Biol Crystallogr 58: 899–907. PMID: 12037327

38. Balakrishnan S, Kamisetty H, Carbonell JG, Lee SI, Langmead CJ (2011) Learning generative models for protein fold families. Proteins 79: 1061–1078. doi: 10.1002/prot.22934 PMID: 21268112

39. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402. PMID: 9254694

40. Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. Proteins 23: 566–579. PMID: 8749853

41. Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. Proteins 57: 702–710. PMID: 15476259

42. Schrodinger, LLC (2010) The PyMOL Molecular Graphics System, Version 1.3r1.

A.3 Supporting Information

SUPPLEMENT MATERIAL TO:

Combining evolutionary information and an iterative sampling strategy for accurate protein structure prediction

Tatjana Braun, Julia Koehler Leman, Oliver F. Lange

Table of Contents

Supporting Figures

Figure A: Computational expense for the initial RASREC run

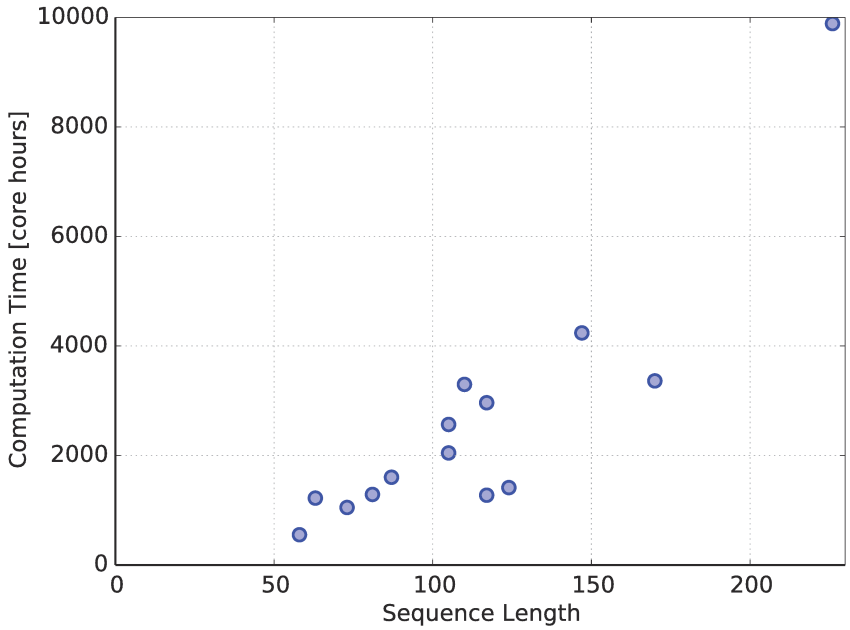

Figure A: Computational expense for the initial RASREC run. The computation time of the initial RASREC run for each of the targets of the EVFold benchmark set is plotted against the corresponding sequence length. The prediction has been carried out using a pool size of 500 on 2.6 GHz AMD Opteron processors. The computation time is dependent on several factors, including sequence length, fold complexity, and instructiveness of the restraints.

Figure B: Lowest-Energy ReRASREC-PLM Structures

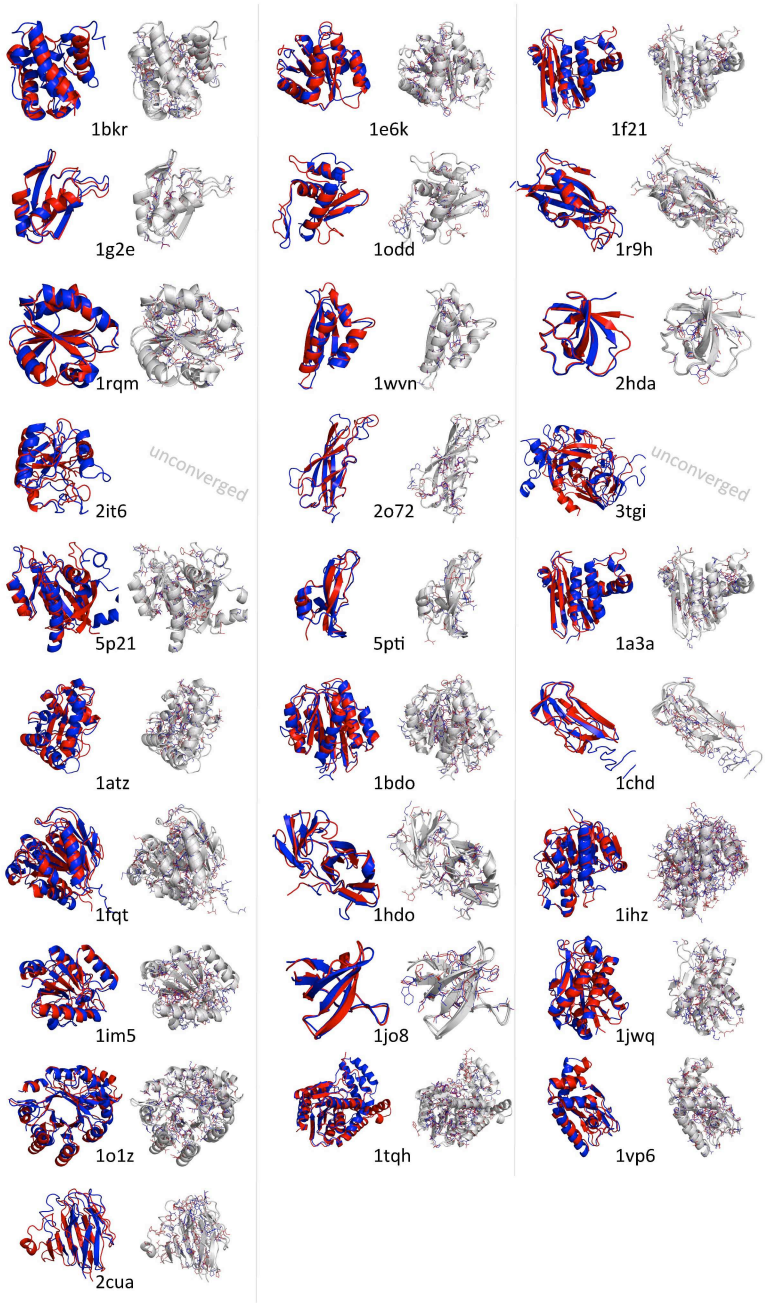

Figure B: Lowest-Energy ReRASREC-PLM Structures. The lowest-energy structures (red) are shown superimposed with the reference structure (blue). The same structures are shown with non-polar sidechains as lines in the right hand panel. If the RASREC calculation did not converge, the panel for sidechain details is omitted.

3

Figure C: Comparison of ReRASREC-DI and EVFold-DI

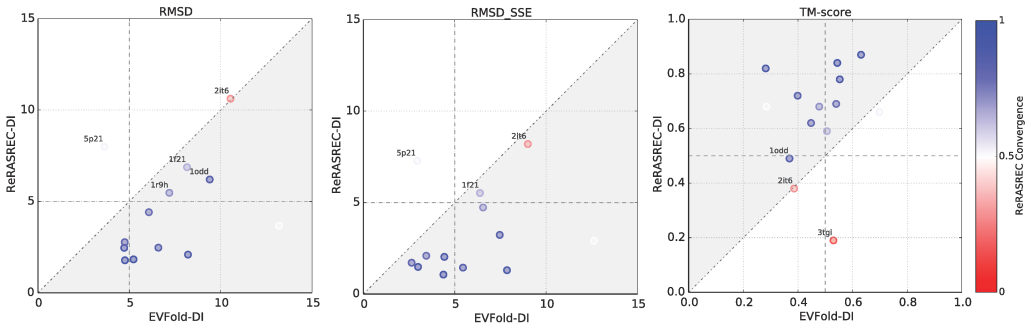

Figure C: Comparison of ReRASREC-DI and EVFold-DI. In case of ReRASREC-DI, the similarity measures are averaged over the 10 lowest-energy models, while for EVFold-DI the single top ranked model is evaluated. The color represents the fraction of converged residues in the 30 lowest-energy models of ReRASREC-DI. The gray areas indicate an improvement of ReRASREC-DI over EVFold-DI.

Figure D: Comparison of top-ranked ReRASREC structures and lowest-RMSD EVFold models

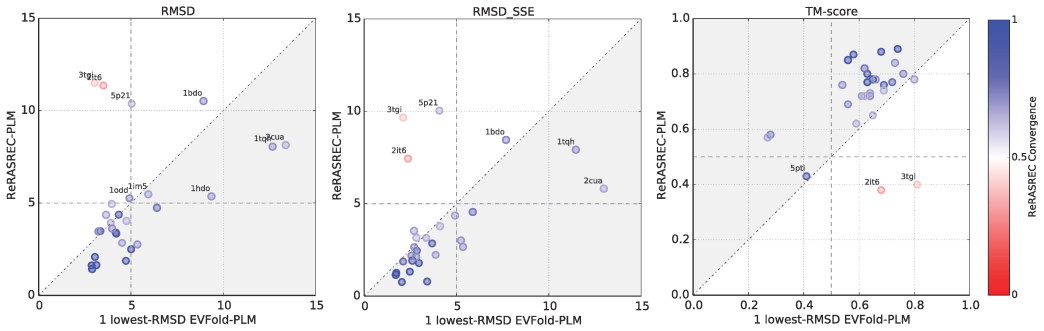

Figure D: Comparison of top-ranked ReRASREC models and lowest RMSD EVFold models. For each different metric, the single best performing EVFold structure (lowest RMSD) was selected among all 50 provided models and is compared to the average of this metric across the 10 lowest-energy models of a RASREC refinement run (ReRASREC-PLM)

5

Figure E: Analysis of convergence and performance

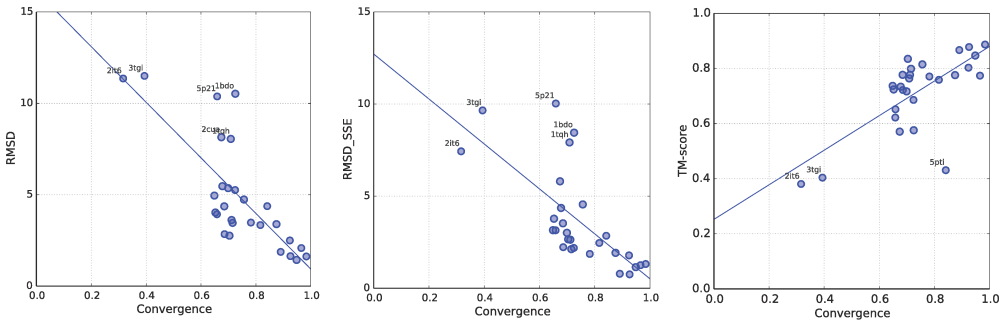

Figure E: Analysis of prediction performance and convergence. The different similarity measures (RMSD, RMSD_SSE, and TM-score) averaged over the 10 lowest energy models are compared to the fraction of converged residues in the 30 lowest energy models.

Figure F: Fraction of satisfied restraints in the native and top-ranked models

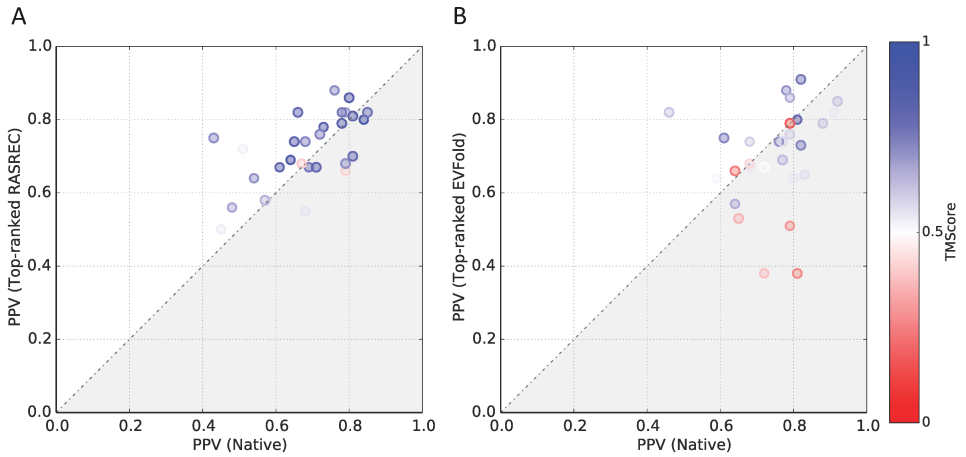

Figure F: Fraction of satisfied restraints in native structures and top-ranked models. PPVs were calculated for the restraint sets used for predicting the top-ranked models. Restraints were considered satisfied if the Cβ-Cβ distance is smaller or equal to 8Å. A) PPV values of restraints on top-ranked RASREC models (y-axis) and on the native structures (x-axis). For each target, L (=sequence length rounded down to the nearest multiple of 10) restraints were used. B) PPV values of Cβ-Cβ restraints on top-ranked EVFold models (y-axis) and on the native structures (x-axis). The restraints used for evaluation are the ones that have been used for generating the top-ranked EVFold model and the number therefore varies for each target between 10 and the sequence length.

Supporting Tables

Table A: TM-scores for EVFold, RASREC, PconsFold and FRAGFOLD

Benchmark set	Target	EVFold-PLM*	ReRASREC-PLM†	PconsFold (20k decoys)‡	FRAGFOLD§	Highest TM-score FAGFOLD||	Lowest-RMSD ReRASREC-PLM¶
EVFold	1bkr	0.30	0.62	**0.74**	N/A	<0.76	<0.69
	1e6k	0.71	**0.89**	0.82	N/A	<0.85	<0.91
	1f21	0.70	**0.76**	0.61	N/A	<0.58	<0.82
	1g2e	0.56	**0.88**	0.80	N/A	<0.80	<0.91
	1odd	0.51	**0.69**	0.59	N/A	<0.62	<0.76
	1r9h	0.57	**0.72**	0.65	N/A	<0.91	<0.81
	1rqm	0.54	0.80	**0.83**	N/A	<0.80	<0.85
	1wvn	0.54	**0.87**	0.60	N/A	<0.70	<0.89
	2hda	0.42	**0.77**	0.57	N/A	<0.85	<0.87
	2it6	**0.66**	0.38	0.54	N/A	<0.45	<0.42
	2o72	0.65	**0.77**	0.53	N/A	-	<0.80
	3tgi	**0.80**	0.40	0.54	N/A	<0.51	<0.42
	5p21	0.59	0.65	**0.67**	N/A	<0.88	<0.66
	5pti	0.38	0.43	**0.57**	N/A	<0.57	<0.55
	Mean	0.57	**0.69**	0.65	N/A	<0.71	<0.74
Pconsfold	1a3a	0.61	**0.74**	N/A	0.56	N/A	<0.77
	1atz	0.73	**0.84**	N/A	0.64	N/A	<0.88
	1bdo	0.25	**0.58**	N/A	0.44	N/A	<0.59
	1chd	0.76	**0.78**	N/A	0.65	N/A	<0.79
	1fqt	0.61	**0.78**	N/A	0.7	N/A	<0.88
	1hdo	0.61	**0.72**	N/A	0.56	N/A	<0.73
	1ihz	0.63	**0.78**	N/A	0.62	N/A	<0.81
	1im5	0.59	**0.73**	N/A	0.56	N/A	<0.78
	1jo8	0.49	**0.85**	N/A	0.84	N/A	<0.9
	1jwq	0.73	**0.80**	N/A	0.31	N/A	<0.84
	1o1z	0.17	**0.82**	N/A	0.61	N/A	<0.85
	1tqh	0.54	**0.76**	N/A	0.5	N/A	<0.77
	1vp6	0.36	**0.72**	N/A	0.52	N/A	<0.85
	2cua	0.27	**0.57**	N/A	0.56	N/A	<0.61
	Mean	0.52	**0.75**	N/A	0.58	N/A	<0.79

Table A: TM-scores for EVFold, RASREC, PconsFold and FRAGFOLD. Column 1-4 show the top ranked results for methods EVFold-PLM, ReRASREC-PLM, PconsFold and FRAGFOLD. Column 5 furthermore shows the highest TM-scores obtained with FRAGFOLD and Column 6 the average TM-score of the ensemble of 10 lowest-RMSD models obtained with ReRASREC-PLM. The numbers in these two columns are preceeded by '<' as they do not reflect the 'real' (best ranked) results, but the ones closest to the native and are therefore not comparable to the rest.

The highest TM-score for each target amongst the top-ranked results (Column 1-4) is highlighted.

* Single top ranked results using the webserver of EVFold with PLM-restraints

† Ensemble of 10 lowest-energy models of ReRASREC using PLM-restraints

‡ Top ranked models obtained with PconsFold. Restraints have been generated with PconsC. Values are taken from (Michel et al., 2014).

§ Highest TM-score amongst top 5 ranked models using FRAGFOLD (using all contacts). Restraints have been generated with PSICOV. Values are taken from (Kosciolek & Jones, 2014).

|| Highest TM-scores obtained with FRAGFOLD. Restraints have been generated with PSICOV. Values are taken from (Kosciolek & Jones, 2014).

¶ Ensemble of 10 lowest-RMSD structures obtained with ReRASREC-PLM.

Table B: Accuracy of side-chain χ1 rotamers for EVFold-PLM and ReRASREC-PLM

Benchmark set	Target		Recovered χ1 Rotamers					
			ReRASREC-PLM		EVFold-PLM		EVFold-PLM (relaxed)	
		buried*	Recovered χ1 †	Fraction of rec. χ1 ‡	Recovered χ1 †	Fraction of rec. χ1 ‡	Recovered χ1 †	Fraction of rec. χ1 ‡
EVFold benchmark set	1bkr	42	28	0.67	14	0.33	16	0.38
	1e6k	49	39	0.80	27	0.55	25	0.51
	1f21	53	38	0.72	32	0.60	37	0.70
	1g2e	25	19	0.76	11	0.44	13	0.52
	1odd	27	19	0.70	13	0.48	13	0.48
	1r9h	36	28	0.78	12	0.33	16	0.44
	1rqm	42	25	0.60	22	0.52	21	0.50
	1wvn	19	14	0.74	6	0.32	7	0.37
	2hda	16	9	0.56	5	0.31	11	0.69
	2it6	48	24	0.50	21	0.44	26	0.54
	2o72	27	20	0.74	11	0.41	10	0.37
	3tgi	101	52	0.51	44	0.44	49	0.49
	5p21	71	45	0.63	31	0.44	38	0.54
	5pti	14	5	0.36	8	0.57	8	0.57
Pconsfold benchmark set	1a3a	56	37	0.66	23	0.41	27	0.48
	1atz	72	47	0.65	29	0.40	34	0.47
	1bdo	25	15	0.60	8	0.32	11	0.44
	1chd	74	50	0.68	32	0.43	37	0.50
	1fqt	44	32	0.73	16	0.36	21	0.48
	1hdo	84	46	0.55	30	0.36	38	0.45
	1ihz	51	35	0.69	26	0.51	25	0.49
	1im5	68	40	0.59	27	0.40	32	0.47
	1jo8	15	11	0.73	4	0.27	6	0.40
	1jwq	76	47	0.62	30	0.39	40	0.53
	1o1z	99	70	0.71	34	0.34	58	0.59
	1tqh	106	79	0.75	43	0.41	50	0.47
	1vp6	50	39	0.78	25	0.50	27	0.54
	2cua	46	25	0.54	15	0.33	20	0.43
Average	N/A	N/A	33.5	0.65	21.4	0.41	25.6	0.49

Table B: Accuracy of side-chain χ1 rotamers. Buried side chains in single top-ranked models are selected and their adopted rotamer assignments are compared to those in the reference crystal structure. In case of EVFold, the analysis has been carried out for the models as generated by CNS (EVFold-PLM) and after relaxing them with fixed backbone atoms within the Rosettal full-atom energy (EVFold-PLM (relaxed)). Glycine and Alanine are excluded from this analysis.

* Side chains that are buried in the reference structure (SASA < 40Å)

† Side chains that are buried in the reference structure and have the same χ1 rotamer assignment in the top-ranked models as in the reference structure

‡ Fraction of buried side chains with the the same χ1 rotamer assignment in the top-ranked models and the reference structure

Table C: Comparison between top ranked and lowest-RMSD structures

		RMSD								TM-score							
Bench-mark set	Target	ReRASREC-PLM				EVFold-PLM				ReRASREC-PLM				EVFold-PLM			
		top 10*	best 10†	top 1‡	best 1§	top 10*	best 10†	top 1‡	best 1§	top 10*	best 10†	top 1‡	best 1§	top 10*	best 10†	top 1‡	best 1§
EVFold bench-mark set	1bkr	3.93	3.06	4.11	2.93	9.52	4.77	13.79	3.91	0.62	0.69	0.60	0.72	0.41	0.54	0.30	0.59
	1e6k	1.62	1.37	1.57	1.27	3.44	3.10	3.34	2.87	0.89	0.92	0.90	0.93	0.70	0.72	0.71	0.74
	1f21	3.34	2.63	2.87	2.56	4.84	4.40	4.21	4.21	0.76	0.82	0.81	0.83	0.65	0.67	0.70	0.69
	1g2e	1.64	1.09	1.48	0.82	3.83	3.39	4.23	3.13	0.88	0.92	0.89	0.94	0.59	0.65	0.56	0.68
	1odd	5.26	3.27	5.46	2.93	6.15	5.28	6.14	4.92	0.69	0.76	0.67	0.79	0.50	0.53	0.51	0.56
	1r9h	2.84	2.05	2.51	1.85	5.93	4.86	4.87	4.52	0.72	0.82	0.76	0.84	0.52	0.56	0.57	0.61
	1rqm	2.50	1.77	2.26	1.63	7.23	5.73	5.91	5.01	0.80	0.86	0.82	0.86	0.52	0.57	0.54	0.63
	1wvn	1.87	1.35	2.01	1.29	5.83	5.26	5.87	4.73	0.87	0.89	0.85	0.90	0.51	0.52	0.54	0.58
	2hda	2.08	1.36	2.03	1.30	4.64	3.60	4.91	3.05	0.77	0.87	0.84	0.89	0.49	0.53	0.42	0.63
	2it6*	11.36	9.54	11.74	9.17	4.31	3.72	3.94	3.51	0.38	0.42	0.37	0.45	0.60	0.66	0.66	0.68
	2o72	3.48	2.84	2.86	2.62	4.13	3.95	4.14	3.35	0.77	0.80	0.82	0.82	0.66	0.67	0.65	0.72
	3tgi*	11.50	10.04	11.17	9.75	3.53	3.19	3.12	3.04	0.40	0.42	0.40	0.43	0.77	0.80	0.80	0.81
	5p21	10.38	8.78	10.70	7.88	7.04	5.50	6.58	5.04	0.65	0.66	0.65	0.67	0.57	0.61	0.59	0.65
	5pti	4.37	3.27	4.69	2.92	8.65	4.93	5.82	4.34	0.43	0.55	0.43	0.60	0.27	0.37	0.38	0.41
Pcons-Fold bench-mark set	1a3a	4.95	3.47	3.38	2.62	5.30	4.37	5.28	3.96	0.74	0.77	0.78	0.83	0.64	0.68	0.61	0.69
	1atz	2.76	2.21	2.44	2.15	9.93	6.52	5.35	5.21	0.84	0.89	0.87	0.89	0.61	0.71	0.73	0.73
	1bdo	10.52	8.17	11.73	7.54	17.67	10.66	11.39	8.93	0.58	0.59	0.56	0.58	0.25	0.25	0.25	0.28
	1chd	4.36	3.98	4.31	3.78	4.22	3.97	4.10	3.65	0.78	0.79	0.78	0.81	0.77	0.79	0.76	0.80
	1fqt	3.39	1.70	2.24	1.53	4.88	4.76	5.36	4.18	0.78	0.88	0.82	0.89	0.61	0.61	0.61	0.65
	1hdo	5.36	4.79	5.57	4.74	10.53	9.49	9.46	9.36	0.72	0.73	0.71	0.73	0.58	0.61	0.61	0.64
	1ihz	3.62	2.68	3.90	2.46	5.00	4.49	4.46	3.98	0.78	0.81	0.76	0.81	0.61	0.63	0.63	0.66
	1im5	5.47	3.93	5.69	3.59	8.03	6.54	7.30	5.94	0.73	0.78	0.70	0.81	0.58	0.62	0.59	0.64
	1jo8	1.43	0.95	1.54	0.84	4.55	3.19	3.56	2.90	0.85	0.90	0.85	0.91	0.47	0.54	0.49	0.56
	1jwq	3.46	2.86	3.50	2.71	3.54	3.41	3.57	3.24	0.80	0.84	0.79	0.85	0.73	0.74	0.73	0.76
	1o1z	4.74	4.24	4.95	4.21	15.25	8.69	25.41	6.41	0.82	0.85	0.80	0.84	0.39	0.55	0.17	0.62
	1tqh	8.05	5.20	7.98	4.33	15.82	13.44	12.67	12.67	0.76	0.78	0.78	0.79	0.48	0.53	0.54	0.54
	1vp6	4.03	2.03	2.75	1.93	12.58	9.04	10.95	4.76	0.72	0.85	0.82	0.86	0.35	0.44	0.36	0.62
	2cua	8.14	7.10	8.32	6.99	19.94	18.27	19.22	14.18	0.57	0.61	0.57	0.61	0.29	0.30	0.27	0.27
Average		**4.87**	**3.78**	**4.78**	**3.51**	**7.72**	**6.02**	**7.32**	**5.18**	**0.72**	**0.77**	**0.73**	**0.78**	**0.54**	**0.59**	**0.55**	**0.62**

Table C: Comparison between top ranked and lowest-RMSD structures.

* ensemble of 10 lowest-energy (ReRASREC-PLM)/ best ranked (EVFold-PLM) models

† ensemble of 10 lowest-RMSD models

‡ single lowest-energy (ReRASREC-PLM)/ best ranked (EVFold-PLM) models

§ single lowest-RMSD models

Table D: Restraint classification performance of RASREC

Benchmark Set	Target	#Restraints	TP	FP	TN	FN	TPR= TP/(TP+FN)	TNR= TN/(TN+FP)	PPV= TP/(TP+FP)	NPV= TN/(TN+FN)	ACC= (TP+TN)/(P+N)
EVFold Benchmark Set	**1bkr**	110	40	15	45	10	0.80	0.75	0.73	0.82	0.77
	1e6k	120	81	12	20	7	0.92	0.63	0.87	0.74	0.84
	1f21	140	79	15	28	18	0.81	0.65	0.84	0.61	0.76
	1g2e	80	63	1	12	4	0.94	0.92	0.98	0.75	0.94
	1odd	80	38	13	24	5	0.88	0.65	0.75	0.83	0.78
	1r9h	100	70	12	9	9	0.89	0.43	0.85	0.50	0.79
	1rqm	100	55	12	27	6	0.90	0.69	0.82	0.82	0.82
	1wvn	70	45	3	22	0	1.00	0.88	0.94	1.00	0.96
	2hda	50	36	5	6	3	0.92	0.55	0.88	0.67	0.84
	2it6	110	50	11	24	25	0.67	0.69	0.82	0.49	0.67
	2o72	110	80	17	9	4	0.95	0.35	0.82	0.69	0.81
	3tgi	220	125	21	25	49	0.72	0.54	0.86	0.34	0.68
	5p21	170	73	23	66	8	0.90	0.74	0.76	0.89	0.82
	5pti	60	30	11	9	10	0.75	0.45	0.73	0.47	0.65
Pconsfold Benchmark Set	**1a3a**	140	87	8	22	23	0.79	0.73	0.92	0.49	0.78
	1atz	180	134	12	22	12	0.92	0.65	0.92	0.65	0.87
	1bdo	80	38	20	19	3	0.93	0.49	0.66	0.86	0.71
	1chd	200	131	9	30	30	0.81	0.77	0.94	0.50	0.81
	1fqt	110	84	6	10	10	0.89	0.63	0.93	0.50	0.85
	1hdo	200	78	72	41	9	0.90	0.36	0.52	0.82	0.60
	1ihz	140	99	11	20	10	0.91	0.65	0.90	0.67	0.85
	1im5	180	116	20	30	14	0.89	0.60	0.85	0.68	0.81
	1jo8	50	40	3	7	0	1.00	0.70	0.93	1.00	0.94
	1jwq	170	100	26	33	11	0.90	0.56	0.79	0.75	0.78
	1o1z	230	138	16	51	25	0.85	0.76	0.90	0.67	0.82
	1tqh	240	149	29	48	14	0.91	0.62	0.84	0.77	0.82
	1vp6	130	85	21	23	1	0.99	0.52	0.80	0.96	0.83
	2cua	130	63	12	44	11	0.85	0.79	0.84	0.80	0.82
Average		-	-	-	-	-	0.88	0.63	0.84	0.70	0.80

Table D: Restraint classification performance of RASREC. Evaluation is carried out for the single top-ranked RASREC models. Restraint sets are the ones used for model generation. A restraint is defined as correct (P) if the Cβ-Cβ distance in the native structure is ≤ 8Å, otherwise as incorrect (N).
TP: Restraints satisfied in both model and native structure
FP: Restraints satisfied in model, but not in the native structure
TN: Restraints neither satisfied in model nor native structure
FN: Restraints satisfied in native structure, but not in model

Supporting Methods

Detailed instructions about how to recreate and analyze the results presented in the manuscript can be found in the protocol capture (provided as File S2 and in the current Rosetta release). The protocol capture contains all necessary flag files, command lines and scripts.

Method A: Contact Prediction and Restraint File Generation

The contact predictions used in this manuscript have been generated with the EVFold webserver (Marks et al., 2011) (available at http://evfold.org/evfold-web/newprediction.do) using standard parameters. The results can be downloaded in form of a compressed folder, which is subdivided into several subdirectories. The all-by-all residue pairing scores are stored in `{jobname}_{scoringmethod}.txt` in the subfolder `ev_couplings`. In case of PLM as scoring method, the file is named `{jobname}_PLM.txt`.

From this file, the L top-ranked residue pairing scores having a minimum distance of 5 residues are extracted and translated into Rosetta specific distance restraints with a sigmoidal potential.

Exemplary excerpt from the generated distance restraint file:

```
AtomPair   CA   97   CB  117  SIGMOID 8.00 1.00 #ContactMap: 0.82
AtomPair   CB   18   CB   47  SIGMOID 8.00 1.00 #ContactMap: 0.78
AtomPair   CB   89   CB  113  SIGMOID 8.00 1.00 #ContactMap: 0.77
```

Method B: Structure Prediciton with the RASREC protocol

Fragment Selection

We have run the fragment picker for all targets with the following command:
`make_fragments.pl -nohoms`

The flag `-nohoms` leads to exclusion of fragments from homologous proteins. This flag should be omitted when not used for benchmarking.

Starting a RASREC run

The structures are generated with the RASREC protocol (Lange & Baker, 2012) of the molecular modeling suite Rosetta (using commit #aa72710 from March 2014) with a pool size of 500. A detailed list of all commands used for a RASREC calculation can be found below.

```
mpiexec -np <CORES> minirosetta.mpi.linuxgccrelease -out:file:silent decoys.out
@flags_denovo @flags_rasrec @flags_iterative -run:archive -out:nstruct <CORES-
3>
```

Flag Files, Patches and Broker File

Command-line flags and patches are separated into a number of different files. All parameters used to generate the data in our manuscript are listed below.

For executing RASREC, bold elements need to be replaced with actual input files.

flags_denovo

```
-run:protocol broker

#fragment files
-frag3 <frags.3mers>
-frag9 <frags.9mers>
#input fasta sequence
-in:file:fasta <sequence.fasta>

-out:file:silent_print_all_score_headers
-increase_cycles 2.000000

#jumping
-templates::topology_rank_cutoff 0.8
-jumps:ramp_chainbreaks
-jumps:sep_switch_accelerate 0.8
-abinitio:skip_convergence_check
-jumps:overlap_chainbreak

#energy fixes
-rsd_wt_helix 0.5
-rsd_wt_loop 0.5
-rg_reweight 0.5

#for loop closing
-overwrite_filter_scorefxn score3
-detect_disulf false

#loop-closing filter in SlidingWindow
-fast_loops:overwrite_filter_scorefxn score3

-abrelax:fail_unclosed

#specify logfile output level
-unmute memory_usage
-out:levels core.chemical:error
-out:levels core.io.pdb:error
-out:levels protocols.jobdist:error

#load flags in flags_nmr_patches
@flags_nmr_patches
```

flags_fullatom

```
-relax:fast
-relax:ramady
-abinitio:close_loops
-loops:idealize_before_loop_close
-loops:idealize_after_loop_close
-abinitio::clear_pose_cache
-short_frag_cycles 1
-scored_frag_cycles 1
```

```
-non_ideal_loop_closing
-alternative_closure_protocol
-fast_loops:window_accept_ratio .01
-fast_loops:nr_scored_sampling_passes 4
-fast_loops:min_breakout_good_loops 5
-fast_loops:min_breakout_fast_loops 80
-fast_loops:min_fast_loops 3
-fast_loops:min_good_loops 0
-fast_loops:nr_scored_fragments 20
-fast_loops:vdw_delta 0.5
-fast_loops:give_up 1000

flags_iterative

-iterative:enumerate:skip_half
#RASREC pool size
-iterative:pool_size 500
#Acceptance ratio for different RASREC stages
-iterative:accept_ratio 0.1 0.1 0.1  0.1 0.1 0.1 0.1 0.1
-jumps::max_strand_gap_allowed 10
-jumps:contact_score 0.2
-iterative:rmsf_nstruct 50
#Output levels for log file
-out:level 400
-out:levels all:warning
-out:levels protocols.jd2.MPIArchiveJobDistributor:info
-out:levels protocols.jd2.Archive:debug
-out:levels protocols.iterative:info
-out:levels core.util.prof:info

#obsolete
-iterative:evaluate_only_on_slaves

#scoring functions for fullatom and centroid stages
-iterative:fa_score talaris2013
-iterative:cen_score score3

#Stages:
# (1)  SS-RANDOM
# (2)  MIX
# (3)  BETA-TOP
# (4)  RESAM
# (5)  NOE-BETA-TOP
# (6)  NOE-RESAM
# (7)  CEN2FULL
# (8)  FULL-REFINE

-iterative:max_nstruct 0 0 0 0 -1 -1 0 0
-iterative:min_diversity 0 0 0 2.0 3.0 2.0 2.0 1.5
-iterative:fullatom

-iterative:safety_hatch_scorecut 0.1
-iterative::super_quick_relax_patch super_quick_relax.patch

#this is the relative weight the noesy-cst will have for filtering
#the relative weight provided in the following is multiplied with the overall
#weight for atom_pair_constraint in the patches
#given by -iterative:cen_score_patch and -iterative:fa_score_patch
-iterative:cenpool_noesy_cst_weight 1
-iterative:fapool_noesy_cst_weight 1

#exit as soon as queue is drained
```

```
-jd2:mpi_nowait_for_remaining_jobs
-jd2:mpi_timeout_factor 0

-iterative:flags_fullatom flags_fullatom

#important to obtain intermediate structures for proto-fold resampling (aka
stage2 resampling)
-abinitio:debug
-abinitio:debug_structures

-archive:completion_notify_frequency 125

flags_rasrec

#File containing information about distance restraints
-broker:setup setup_init.tpb

#Only needed for evaluation purposes (in case native structure is known)
-in:file:native <native.pdb>
-evaluation:rmsd NATIVE _full <native.rigid>



flags_nmr_patches

#patches used for abinitio stages
-abinitio::stage2_patch nmr_patch
-abinitio::stage3a_patch nmr_patch
-abinitio::stage3b_patch nmr_patch
-abinitio::stage4_patch nmr_patch

#for fullatom-relax
-score::patch nmr_relax_patch

# for loop closing
-fast_loops:patch_filter_scorefxn nmr_patch
-patch_filter_scorefxn nmr_patch

-iterative:fa_score_patch  nmr_pool_patch
-iterative:cen_score_patch nmr_pool_patch


nmr_patch
atom_pair_constraint = 5.0
rdc    = 5.0

nmr_pool_patch
chainbreak = 1
linear_chainbreak = 1.33
overlap_chainbreak = 1
atom_pair_constraint = 10
rdc    = 10

nmr_relax_patch
atom_pair_constraint = 0.1
rdc    = 0.1

setup_init.tpb

CLAIMER ConstraintClaimer
file <restraints.cst>
FULLATOM
```

```
CENTROID
SKIP_REDUNDANT 0
FILTER_WEIGHT  1.00
FILTER_NAME restraints_SIGMOID
END_CLAIMER
```

Method C: Refinement with Rasrec

If the convergence of the initial RASREC run was not sufficiently high (fraction of converged residues < 90%), a second RASREC run was carried out. This run reuses restraints from both predicted contact map and the previous results.

Repick Restraints

The refinement run uses restraints from both predicted contact map and the previous results. For this purpose, two different restraint files have been generated:

`target_converged_distances.cst`
This restraint file contains all short-range (<=8 Å) distances with a standard deviation < 1 Å in the 30 lowest-energy RASREC models. These converged distances will be enforced during ReRASREC using a strict bounded potential. To reflect the average distance *d* in the converged region, the lower bound is set to (*d*-1) and the upper bound to (*d*+1).

`target_filtered_contactmaps.cst`
This file contains additional residue-residue pairings from the predicted contact map that affect unconverged regions (residue-residue distance, SD > 1Å in 30 low-energy structures) and do not totally disagree (i.e. are short range with an average distance d $\leq$ 8Å) with the preliminary structures. The restraints are scored with a wide bounded potential with lower bound and upper bound set to 1.5Å and 8Å respectively. Random pairs of these restraints are combined into ambiguous restraints (see below).

Setup RASREC run

The flags and patches used for the refinement RASREC run are identical to the ones listed in Method B: Structure Prediciton with the RASREC protocol. The two RASREC runs only differ in the restraints used for structural guiding. The restraint files are added to a RASREC run in the broker file `setup_init.tpb`.

Both restraint files that have been generated above will be added to the broker file for the refinement run as follows

```
CLAIMER ConstraintClaimer
file target_converged_distances.cst
FULLATOM
CENTROID
SKIP_REDUNDANT 0
FILTER_WEIGHT  1.00
FILTER_NAME converged_distances
END_CLAIMERls
```

```
CLAIMER ConstraintClaimer
file target_filtered_contactmaps.cst
FULLATOM
CENTROID
COMBINE_RATIO 2   #make the restriants ambiguous
SKIP_REDUNDANT 0
FILTER_WEIGHT  1.00
FILTER_NAME filtered_contactmaps
END_CLAIMER
```

Please note, that for the filtered_contactmap.cst restraints the following line is added:
`COMBINE_RATIO 2`
This line transforms the restraints to ambiguous ones.

References

1. Michel M, Hayat S, Skwark MJ, Sander C, Marks DS, Elofsson A (2014) PconsFold: improved contact predictions improve protein models. Bioinformatics 30: i482-8.

2. Kosciolek T, Jones DT (2014) De novo structure prediction of globular proteins aided by sequence variation-derived contacts. PLoS One 9: e92197.

3. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. (2011) Protein 3D structure computed from evolutionary sequence variation. PLoS One 6: e28766.

4. Lange OF, Baker D (2012) Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. Proteins 80: 884-95.

B. Embedded Manuscript 1

B.1 Copy Permissions

Manuscript under preparation.

B.2 Full Article

Automatic Protein Structure Modelling into Cryo-EM density maps using EMfasa

Tatjana Braun[1*], Zhe Wang[1], Gunnar F. Schröder[1,2*]

[1] Institute of Complex Systems (ICS-6), Forschungszentrum Jülich, 52425 Jülich, Germany
[2] Physics Department, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany

* correspondence should be addressed to Tatjana Braun (tatjana.braun@julumni.fz-juelich.de) and Gunnar F. Schröder (gu.schroeder@fz-juelich.de)

Building atomic models of proteins with cryo-EM density maps at near-atomic resolution (3–5 Å) is very challenging and prone to errors. To address this problem, we present EMfasa, a fully automated protocol for *de novo* model-building into cryo-EM density maps. EMfasa couples backbone tracing with sequence non-specific fragment assembly to rapidly generate highly accurate all-alanine structures that are subsequently completed to full-atom models via automated side-chain assignment.

Until recently, resolutions of cryo-electron microscopy (cryo-EM) density maps were limited to the medium- (5–10 Å) and low-resolution range (10–60 Å). Nowadays, thanks to a new generation of electron detectors and powerful image processing routines, large macromolecules can be obtained at near-atomic or even atomic resolution using cryo-EM (1, 2).

While structural interpretation of maps in the medium- and low-resolution range is limited to secondary structure assignment, topology determination, and a rigid or flexible fit of known atomic structures of individual proteins (e.g. obtained by X-ray crystallography or NMR) to obtain atomistic representations of macromolecular assemblies, high-resolution density maps allow to directly build atomic models. Cryo-EM is therefore becoming an important tool to fill the ever-increasing gap between known protein sequences and resolved protein structures.

For existing methods it is highly challenging and often impossible to automatically build accurate models from density maps with resolutions between 3.5 – 5 Å: In 2015, Wang et al. have described a *de novo* modeling approach for cryo-EM maps that matches and assembles sequence-based local backbone conformations, so called fragments, into the target density map within the molecular modeling suite Rosetta (3). The routine *map_to_model* of the Phenix software suite (4) tries to automatically build an atomic model by combining several procedures, including the standard RESOLVE model-building (5), chain tracing and map sharpening. EM-Fold (6, 7) assembles predicted secondary structural elements into the corresponding regions of the density map and subsequently uses Rosetta (8) to build loops and side-chains.

Here we present EMfasa, a *de novo* model-building approach for cryo-EM maps in the near-atomic resolution range (better than 5 Å) that combines backbone tracing, sequence non-specific fragment assembly, and automated side-chain assignment (Figure 1). Analogously, this method can be expected to be useful for the interpretation of electron density maps from X-ray crystallography. Our method scales approximately linearly with the size of the protein and is therefore particularly well suited for large systems. The high speed allows for sampling a large number of backbone conformations as well as sequence-to-backbone assignments. EMfasa aims at rapidly building a first model, which subsequently can be further refined using computationally more expensive methods (e.g. Rosetta or MDFF (9)).

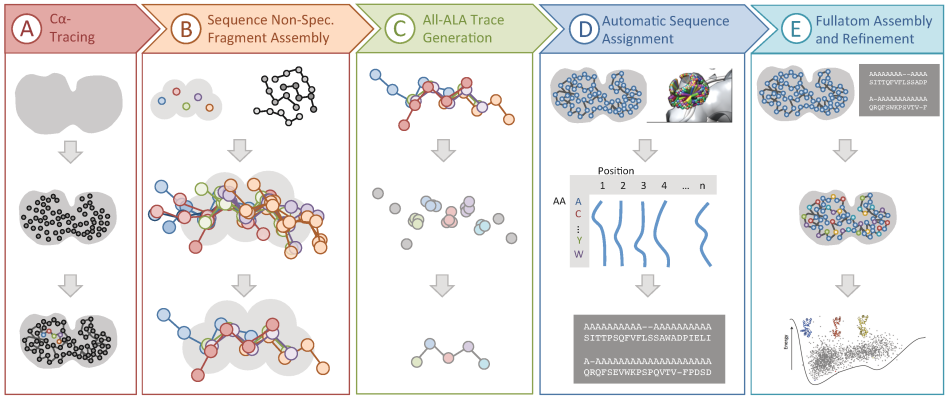

Figure 1: Overview of the EMfasa protocol. Our protocol consists of five steps: (A) generation of a Cα-trace using a combinatorial optimization algorithm, (B) matching of sequence non-specific local backbone conformations into the density map and identification of matching subsets of these fragments, (C) all-alanine structure generation by assembly of consistent fragment subsets via clustering, (D) automated sequence assignment by rotamer matching, (E) full-atom assembly and refinement.

Given a segmented protein density map at 3 to 4.8 Å resolution and the primary sequence, we start by generating a Cα-trace that roughly describes the location of and the connections between the different residues of our protein of interest. These locations are in the next step used to locally fit sequence non-specific 7-mer fragments of known protein structures into the density map. For each residue position, the best placement to build a subset of matching fragments is chosen using a scoring function considering several terms, including i.a. the fit of a fragment to the density and the overlap to neighboring fragments. The clustered and averaged atom positions of this subset are subsequently used to build a complete all-alanine model. Using this all-alanine model and a rotamer library, a profile is generated reflecting the fit of each of the 20 amino acids at each backbone position. This profile is in the next step used to assign the primary sequence of the protein to the backbone structure using a profile-sequence alignment. Finally, the assigned residues are mapped onto the protein backbone and the model is completed and followed by an all-atom refinement guided by experimental density data. Further details for each step are described in the Online Methods.

Our protocol was tested on six experimental density maps (Table 1) with reported resolutions between 3.3 Å and 4.8 Å. The targets range in size from 149 (brome mosaic virus) to 665 residues (nicastrin) and include different secondary structure elements and fold types. In order to contrast our method with other recent approaches, we ran the Rosetta *de novo* modeling approach for all targets as described in the Online Methods.

Target	EMDB / Res.[a]	PDB entry / Size[b]	Fold	EMfasa						
				top Ca-trace[1]		top all-alanine structure[2]	Lowest-RMSD full-atom model[3]		Lowest-RMSD full-atom model (Top10 CCmap)[4]	
				TMscore/ RMSD (#aligned residues)[c]	TMscore	TMscore/ RMSD (#aligned residues)[c]	Ca-RMSD (total)	Ca-RMSD (#corr. assigned res/ # residues)[d]	Ca-RMSD (total)	Ca-RMSD (#corr. assigned res/ # residues)[d]
Tobacco Mosaic Virus	2842 / 3.4	4udv / 153	α	0.83 / 1.53 (139)	0.54	0.94 / 0.8 (147)	1.9	0.7 (123/153)	2.1	0.7 (129/153)
Brome Mosaic Virus	6000 / 3.8	3j7l / 149	β	0.83 / 1.7 (137)	0.54	0.9 / 1.3 (143)	2.8	1.0 (97/149)	2.8	1.0 (97/149)
T20SProteasome (α Subunit)	5623 / 3.3	3j9i / 224	α/β	0.88 / 1.68 (211)	0.64	0.91 / 1.1 (210)	2.4	0.9 (153/224)	2.8	0.9 (132/224)
	6219 / 4.8	1pma / 221		0.84 / 2.27 (211)	0.52	0.77 / 2.0 (190)	5.6	1.4 (23/221)	7.8	1.3 (10/221)
T20SProteasome (β subunit)	5623 / 3.3	3j9i / 203	α/β	0.83 / 1.52 (181)	0.43	0.93 / 1.5 (198)	2.7	0.8 (149/203)	3.6	0.8(112/203)
γ-Secretase (Nicastrin)	3061 / 3.4	5a63 / 665	α/β	0.87 / 2.9 (623)	0.61	0.9 / 2.1 (640)	4.8	1.1 (240/665)	5.6	1.1 (196/665)

Table 1: Performance of EMfasa for 6 experimental density maps. [a]Reported resolution. [b]Number of residues in the corresponding PDB entry. [c] Structure alignment as generated by TMalign. The TMscore is normalized by the length of the respective native model. [d] Residues that are within 2.0 Å of their native residue partner are considered to be correctly assigned and used for the RMSD calculation. The numbers in brackets indicate the number of correctly assigned residues (used for the RMSD calculation) and the total number of residues in the full-atom model. [1] Cα-trace with highest TMscore. [2] all-alanine structure with highest TMscore. [3] Full-atom model with lowest Cα-RMSD in the entire full-atom model pool 4 Full-atom model with lowest Ca-RMSD amongst the 10 structures with highest map correlation (CCmap) in the model pool.

Our protocol was able to automatically generate Cα-traces with the correct topology for all six test systems, resulting in an average Cα-RMSD < 2 Å for at least 90 percent of the residues for the best Cα-trace of each protein structure, cf. Table 1. For each target, 100 Cα-traces were generated as described in the Online Methods and scored according to their free map correlation (C_{free}) (10). On average, five of the top-10 scoring Cα-traces were of correct topology, see Table S1. Since the directionality is generally not known at this step, each topologically unique Cα-trace needs to be considered in both directions in the subsequent protocol steps.

In this work, a Cα-trace with correct topology was selected for each target and used to generate several all-alanine structures by fragment-assembly and clustering. While the resulting all-alanine structures do not necessarily contain the correct amount of residues (in low-density regions, residues might have been missed), the majority of residues were placed highly accurately, cf. Table 1: the Cα-RMSDs over at least 85 percent of the protein structure (alignment of the top-TM-score all-alanine model to the native structure using TM-align (11)) are between 0.8 and 2.1 Å. In addition to accurate Cα-atoms, the all-alanine structures contain explicit atomic coordinates for all the other main-chain atoms (carbon, nitrogen, oxygen) and the Cβ atoms of the alanine side-chains, see Figure 2, Figure S1 and Figure S2. Accurate Cβ positions are essential for the

generation of sequence profiles that correctly reflect the fit of the 20 standard amino-acids to the different residue positions.

For four targets (TMV-3.4Å, BMV-3.8Å, 20S(α)-3.3Å, 20S(β)-3.3Å), EMfasa combined the all-alanine structures, the corresponding profiles and resulting profile-sequence alignments to full-atom models whereof the lowest-RMSD models have on average 70 percent correctly assigned and placed residues (residue-pair distance between model and native structure < 2 Å), see Table 1 and Figure S3.

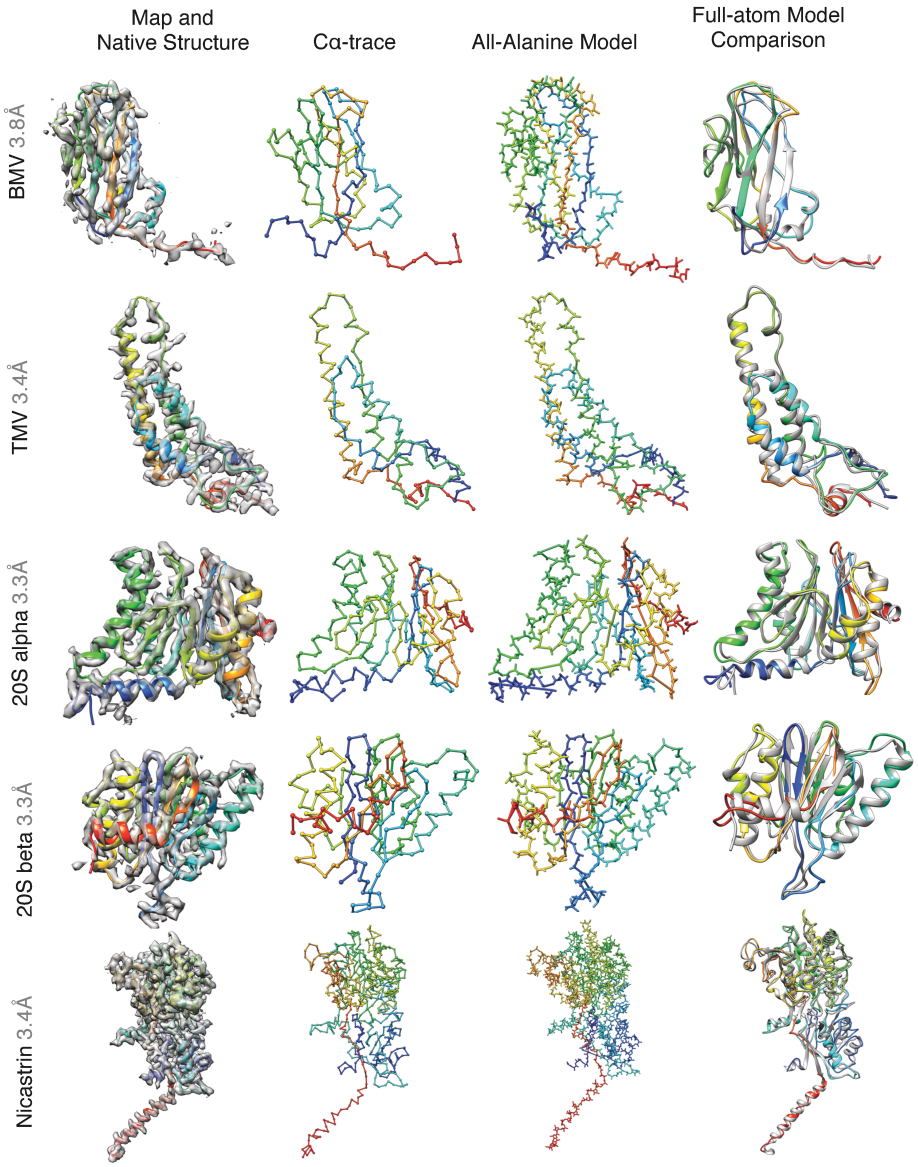

Figure 2: EMfasa results for five representative test systems. The native structure and the experimental density map are shown in column 1) The Cα-trace and the corresponding all-alanine model is shown in columns 2) and 3), respectively. Column 4) shows the lowest-RMSD EMfasa model (rainbow) of the 10 models with highest correlation to the density map (CCmap) compared to the native structure (gray).

For nicastrin (665 residues), 240 residues were within 2 Å of their native residue-partner, resulting in a total Cα-RMSD of 4.8 Å- that being quite accurate

for a protein of that size.

However, EMfasa had problems to assign the sequence of 20S(α)-4.8Å: only 23 residues were assigned correctly. In a density map of resolution < 4.5 Å, the side-chain information is apparently not good enough for building informative profiles.

Using the correlation of the model density with the EM map (CCmap) for final model selection, we obtained at least one model with on average 59 percent correctly assigned and placed residues amongst the 10 best scoring ones for the test systems with a resolution better than 4 Å. As we were in general not able to identify the lowest-RMSD model using the map correlation, it might be of interest to investigate other criteria to further improve the model selection.

We compared our protocol to the Rosetta *de novo* modeling protocol by running two iterations of the *denovo_density* tool (3) for each of the test systems, see Table S2. In 3 cases (TMV-3.4 Å, 20S(α)-3.3 Å, 20S(β)-3.3 Å), the performance of Rosetta is comparable to the lowest-RMSD models in the EMfasa full-atom model pool: The number of correctly assigned residues and the corresponding Cα-RMSDs are highly similar. In 2 cases (BMV-3.8Å, nicastrin-3.4Å), our protocol clearly outperformed the Rosetta *denovo_density* tool: for both targets Rosetta only placed less than half of the residues of which the majority was wrong. This suggests that our protocol has an advantage for proteins with an all-β topology and for very large structures. In case of the 20S(α)-4.8 Å resolution map, Rosetta only placed 14 residues in the final consensus trace, some of them being wrong (distance to the native residue > 2Å). In that resolution range, both EMfasa and Rosetta do not succeed in building accurate all-atom models. While Rosetta only places a small amount of residues, EMfasa was however able to generate a fairly accurate all-alanine structure (Cα-RMSD of 1.9 Å for 82% of the residues).

EMfasa uses a new approach to rapidly build models into cryo-EM density maps at near-atomic resolution. The protocol builds three types of models: Cα -traces, all-alanine structures, and full-atom models. It can therefore assist and greatly reduce human efforts during protein structure determination at various stages. The pool of full-atom models, combined with accurate all-atom force fields, can be used as basis for structure refinement and determination studies. In cases where the density resolution is not high enough to provide sufficient sequence information for side-chain assignment, both the all-alanine structures and Cα-traces can still provide important information about the protein structure and topology, for example for identifying homologous structures. EMfasa treats the sequence assignment as a global optimization problem (by using a global profile-sequence alignment), which we expect to yield an advantage over methods that try to identify residues locally in the density map, especially when the side-chain density is comparatively weak. For unknown proteins with near-atomic density-maps, the all-alanine structures can furthermore be utilized to build sequence profiles to search a protein sequence database for matches. Finally, since EMfasa uses only a single set of sequence non-specific fragments, it scales approximately linearly with the number of amino acids unlike approaches based on sequence-

specific fragments (3) and is therefore a good tool to model large protein structures (e.g. nicastrin).

The EMfasa protocol and a detailed step-by-step tutorial will be available at https://simtk.org/projects/emfasa/ .

Methods

Methods and any associated references are available in the online version of the paper.

Acknowledgements

We thank Nir Kalisman for helpful discussions. The authors gratefully acknowledge the computing time granted by the JARA-HPC Vergabegremium and VSR commission on the supercomputer JURECA at Forschungszentrum Jülich.

Author Contributions

All authors performed the research. T.B. wrote the EMfasa code. T.B and G.F.S wrote the manuscript.

Online Methods

Map Preparation

All maps have been masked (radius: 4 Å) with the corresponding deposited PDB model to obtain density maps for the monomeric subunit. For real application cases, the monomeric subunit needs to be manually or automatically segmented.

"Dust" was removed using UCSF Chimera (12). The monomeric subunits are subsequently locally normalized with *e2proc3d.py*, part of EMAN2 (13), and for the bead placement additionally modified by Gaussian filtering and emphases of ridges, i.e. skeletonization, with UCSF Chimera.

Generation of Cα-Traces

The generation of Cα -traces can be broken down into two individual steps: initially, pseudo Cα atoms are placed into the prepared density map of the monomeric subunit. Subsequently, these are connected with the combinatorial optimization heuristic Lin-Kernighan (14), an algorithm to solve the Traveling Salesman Problem (TSP).

Bead Placement

The tool *dxbeadgen*, part of DireX (15), is used to randomly place twice as many beads (pseudo Cα atoms) as there are residues in the protein chain. To achieve an even distribution of the beads in the map, the bead positions are subsequently refined with DireX using a low weight on the density map and with repulsive forces between the beads.

Trace Generation

For finding the shortest connection between all the beads, we use the LKH program (16), an implementation of a modified Lin-Kernighan heuristic. The input for the LKH program is prepared using Pathwalker (17) whose new version takes into account the density along connections and favors connections that lie within strong density regions. Gaussian noise (sigma=0.5 Å) was used on the bead positions. Using this setup, LKH is used to generate 10 traces with potentially different connectivities. These traces are used to generate a histogram of connections that subsequently is used to construct a cost matrix. The cost matrix is fed to the LKH program to yield traces with consensus connectivity. In total, 10-100 consensus traces are generated. All consensus traces are refined with DireX and the number of beads is reduced by a factor of two to obtain the original number of proteins in the density map. During refinement, distance restraints of 3.8 Å are applied between neighboring Cα atoms and 6 Å between 1-3 pairs to mimic angle restraints.

Fragment Assembly

To generate an all-alanine structure with precise and explicit atomic coordinates, we are matching local backbone conformations of known protein structures into the density maps at the bead locations stored in the backbone trace.

Construction of a Sequence-Nonspecific Fragment Library

The initial set of structures from which the elements of our fragment library were assembled, was selected using the protein sequence culling server PISCES (18). With it, we constructed a non-redundant set of protein structures including 4718 distinctive protein chains that all were determined by Xray crystallography and have a percentage sequence cutoff of 20 percent, a minimum resolution of 1.8 Å, and a R-factor cutoff of 0.25. By requiring structures resolved by Xray crystallography with the minimum resolution and R-factor stated above, we make sure that only well resolved fragments will be present in our final fragment library. From each of the structures in the non-redundant set of protein structures, continuous subsets of length 7 were extracted. To obtain sequence-unspecific fragments, all residues of the extracted subsets were mutated to alanine with Scwrl4 (19). To obtain fragments with ideal bond lengths, bond angles, and torsion angles, the Rosetta application *idealize.cc* was furthermore used. In the next step, the idealized all-alanine fragments were clustered into 100 groups using ClusCo (20) with the maximum-linkage algorithm. The core fragments of the final clusters describe the final fragment library.

Fragment Fitting

The fitting of the individual fragments into the density map is carried out using UCSF Chimera (12). In a first step, the input density map is cropped to a cube with an edge length of 20 Å centering around the current backbone position. This is done to ensure that the fragment will be fitted locally, i.e. in close proximity to the backbone atom of interest. In the next step, the fragment is moved to the center of the density cube by positioning it's central Cα-atom on the selected backbone position. Subsequently, 30 global search operations with the Chimera fitmap command are carried out: Each search operation generates a random initial placement of the fragment within 1 Å of the starting position and follows it by a local optimization. For the local optimization, a density is map with the user-specified resolution is generated from the coordinates of the fragment and a map-in-map fitting is carried out. The map is generated by describing each atom as a Gaussian distribution of width proportional to the defined resolution and amplitude proportional to the atomic number. As metric for the map-in-map fitting, correlation about zero was chosen. For each fragment, the 5 top-scoring placements are stored, which, assuming a fragment library of size 100, results in 500 fragment placements per backbone position.

Monte Carlo Sampling

To find the set of fragments that are mutually compatible, Monte Carlo Simulated Annealing sampling was carried out. To reduce the search space, only the 30 top-scoring fragment placements per bead are considered. At the beginning of the procedure, each bead position is randomly assigned to one of the corresponding 30 fragment placements. During each step of the MC-SA, the assigned fragment at a random bead position is exchanged and the move is either accepted or rejected based on the Metropolis criterion using the score function described below. To address the fact that no good fragments might have been found at a position, 'zero fragments' can be allowed as well. Simulated annealing was carried out by slowly reducing the temperature.
The compatibility of a set of fragments is calculated using a scoring function

consisting of 4 different terms. The term $score_{corr}$ reflects the fit of a fragment to the density and is based on the correlation about zero between the generated density map for the coordinates of the fragment atoms and the density map, used as metric during the fitting procedure with UCSF Chimera. $Score_{overlap}$ describes how well two neighboring fragments overlap, i.e. describe the same atoms. It is defined as the minimum RMSD over at least 2 Cα-atoms between two neighboring fragments. Fragments are considered to be neighbors, when they are assigned to two consecutive beads in the Cα-trace. $Score_{clash}$ evaluates whether there is contact (minimum distance between two Cα-atoms < 2Å) between two fragments that are far apart in sequence, i.e. assigned to beads that are at least 8 positions apart from each other, and $score_{dir}$ evaluates whether the fragment is orientated in the same way as the provided backbone at that location. A visual interpretation of the different score terms is shown in Figure S5. MC-SA is carried out several times to identify several alternative set of fragments with roughly equivalent scores.

All-Alanine Structure Generation

To build a complete all-alanine model, the residues of each compatible fragment set, identified using MC-SA as described before, are clustered based on their Cα coordinates using the density based clustering algorithm DBScan (21). The clustering is carried out in two iterative steps with increasing clustering radius. The clusters are in the next step connected based on the intra-fragment connections of the participating residues, resulting in a structure with several all-Alanine fragments. To increase the size of the all-alanine fragments and to eventually generate a full model, several fragmented all-alanine structures from different MC-SA runs are assembled.
To achieve realistic local geometries, the final all-alanine structures are refined to the density map with PHENIX (4).

Automatic Side-chain Assignment

For being able to build the final model of our protein of interest, the side-chains need to be assigned to the positions in the all-alanine structure. The automatic side-chain assignment is carried out in two steps: Initially, a profile reflecting the fit of each amino acid at each position in the all-alanine backbone is generated. This profile is in the next step aligned to the protein sequence by using a dynamic programming algorithm, resulting in the final assignment.

Construction of Position-Specific Profile

The position-specific profile for the all-alanine backbone is based on the fit of each amino-acid to the density at each of its residue positions. To determine the fit of each amino acid, a rotamer library (Dunbrack backbone-dependent rotamer library (22) or common-atom values from the Richardson backbone-independent rotamer library (23)) is used and the correlation of each rotamer to the input density is calculated. These correlation values are in the next step used to build the profile. In a first step, the best matching rotamer for each of the 20 amino-acids is selected by maximizing the values of the interpolated map at each

atom position. In the next step, the residue-density compatibility for each of the selected rotamers is calculated by zoning the density map around the atoms of the 20 best-fitting rotamers and by calculating the correlation between the zoned map and the calculated map of the rotamer. This results in a matrix that lists a correlation value for each amino-acid at each position in the consensus backbone trace. These values are in the next step transformed to standard (Z) scores using the following equation

$$Z_{aa,i} = \frac{X_{aa,i} - M_{aa}}{SD_{aa}}$$

where $X_{aa,i}$ is the correlation of amino acid aa at backbone position i, and M_{aa} and SD_{aa} the mean and standard deviation of the correlation of that amino acid at each position in the backbone, respectively. The final profile P is calculated as follows:

$$P_{aa,i} = e^{\frac{3}{2}Z_{aa,i}}$$

Profile-Sequence Alignment

To globally align the protein sequence to the profile, the Needleman-Wunsch algorithm (24) with affine gap costs, as described by Gotoh (25). The backtracking routine has been implemented according to (26). The score of aligning residue aa to position i of the protein backbone is stored in the sequence profile $P_{aa,i}$. Affine gap penalties discriminate between gap opening and gap extension and are described in the form $g_{gapOpen} + l \times g_{gapExtend}$. $g_{gapOpen}$ refers to the cost required to open a gap and $g_{gapExtend}$ the cost to extend the length of an existing gap by 1.

Application for structures with close N- and C- termini

For structures whose N- and C- termini are close in 3D space, automatic and correct detection during the Cα-tracing is difficult and, without the use of additional information, might be incorrect. EMFasa can use a profile (generated from an all-alanine structure) to generate alignments for all possible sequence iterations (1-N, 2-N+1, …, N – N-1) and store the sequence resulting in the best-scoring alignment in a FASTA file. This sequence file can in the next step be used to adapt and correct the termini of the all-alanine structure.

Final Model Assembly and Refinement

The profile-sequence alignment, obtained as described before, is in the next step used to build the final model of our protein. To do so, the automodel class of Modeller (27) is used with additional absolute position restraints on each Cα of the all-alanine structure with a standard deviation of 2 Å. The model is in the next step refined to the density using the real space refinement of PHENIX

consisting of global minimization, local rotamer fitting, morphing and simulated annealing.

Pool Generation and Selection of Final Models

A pool of protein structures is generated by using <*u*> topologically different Cα-traces in both directions (A), combining several resulting MC-SA trajectories for each Cα-trace to <*v*> different all-alanine structures (B), calculating <*w*> different alignments via varying gap penalties for each all-alanine structure (C), and generating <*x*> real space refinements for each model assembled based on one alignment (D). This results in a total pool of $2 \cdot u \cdot v \cdot w \cdot x$ different protein structures. The largest variability is obtained by using steps (A) through (C). In this work, we combined 16 MC-SA trajectories to up to 25 different all-alanine structures and calculated one profile for each. For each of the resulting profiles, 21 different alignments were generated (Combinations of GapOpen and GapExtend penalties ranging from -12 to -2 in steps of 2). This results in a maximum of 525 structures, given one Cα-trace in one direction. The structures are not necessarily unique – different profiles and varying trace penalties can result in identical alignments.

The final structures are scored based on their correlation to the density map (cc_map, calculated during the PHENIX real space refinement), cf. **Figure S4**. The correlation score helps to separate models that are based on Cα-traces in the wrong direction from the ones that are based on the correct direction and is able to identify structures at the lower end of the Cα-RMSD range.

Total Runtime

The generation of a Cα-trace takes approximately between 5 minutes (for small proteins like BMV) and 30 minutes for larger structures (nicastrin) on a single core. As the different traces are independent, they can be calculated in parallel. For a fragment library of size 100 and 30 search iterations for each fragment, the fragment fitting takes 10 minutes for each backbone position on a single core. The fittings for each backbone position can be run in parallel. The Monte-Carlo Simulated Annealing takes 10 minutes for small proteins like BMV and up to 2 hours for structures as large as nicastrin on a single core (the run time can be influenced by specifying varying parameters, e.g. iterations per temperature cycle, starting temperature, and temperature reduction factor). For each backbone position, the profile generation takes only several minutes and as they are independent of their neighbors, can be calculated in parallel. The final model assembly and real space refinement take few minutes per model on a single core.

Metrics used for Performance Evaluation

To evaluate the all-alanine structures, TMalign (11) was used to generate a structure alignment between the all-alanine structure (not necessarily containing the correct amount of amino acids) and the corresponding native model. The corresponding TM-scores are normalized by the length of the respective native protein structure and the RMSD values are calculated on the aligned residues in the structure alignment.

Model Generation with Rosetta

Fragments have been generated using the Robetta webserver (28). To mimic real application cases, homologous protein structures were excluded. The models were generated as described in the tutorial available at https://faculty.washington.edu/dimaio/wordpress/software/. The number of translations to search was set to approx. 2 times the number of residues in the map for nicastrin (as suggested in the tutorial) and approx.10 times for the other proteins in the test set. Only partial models based on 2 iterations were generated. The results of the first iteration were used as starting model for the second iteration. RosettaCM was not used to complete the models.

References

1. Bai X-C, McMullan G, & Scheres SH (2015) How cryo-EM is revolutionizing structural biology. *Trends in Biochemical Sciences* 40(1):49-57.
2. Kühlbrandt W (2014) The resolution revolution. *Science* 343(6178):1443-1444.
3. Wang RY-R, *et al.* (2015) De novo protein structure determination from near-atomic-resolution cryo-EM maps. *Nature methods* 12(4):335-338.
4. Adams PD, *et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta crystallographica. Section D, Biological crystallography* 66(Pt 2):213-221.
5. Terwilliger T (2004) SOLVE and RESOLVE: automated structure solution, density modification and model building. *Journal of synchrotron radiation* 11(1):49-52.
6. Lindert S, *et al.* (2012) Ab initio protein modeling into CryoEM density maps using EM‐Fold. *Biopolymers* 97(9):669-677.
7. Lindert S, *et al.* (2009) EM-fold: De novo folding of α-helical proteins guided by intermediate-resolution electron microscopy density maps. *Structure* 17(7):990-1003.

8. Leaver-Fay A, *et al.* (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in enzymology* 487:545.
9. Trabuco LG, Villa E, Mitra K, Frank J, & Schulten K (2008) Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* 16(5):673-683.
10. Falkner B & Schröder GF (2013) Cross-validation in cryo-EM–based structural modeling. *Proceedings of the National Academy of Sciences* 110(22):8930-8935.
11. Zhang Y & Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research* 33(7):2302-2309.
12. Pettersen EF, *et al.* (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry* 25(13):1605-1612.
13. Tang G, *et al.* (2007) EMAN2: an extensible image processing suite for electron microscopy. *Journal of structural biology* 157(1):38-46.
14. Lin S & Kernighan BW (1973) An effective heuristic algorithm for the traveling-salesman problem. *Operations research* 21(2):498-516.
15. Schroder GF, Brunger AT, & Levitt M (2007) Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. *Structure* 15(12):1630-1641.
16. Helsgaun K (2000) An effective implementation of the Lin–Kernighan traveling salesman heuristic. *European Journal of Operational Research* 126(1):106-130.
17. Baker MR, Rees I, Ludtke SJ, Chiu W, & Baker ML (2012) Constructing and validating initial Cα models from subnanometer resolution density maps with pathwalking. *Structure* 20(3):450-463.
18. Wang G & Dunbrack RL, Jr. (2003) PISCES: a protein sequence culling server. *Bioinformatics* 19(12):1589-1591.
19. Krivov GG, Shapovalov MV, & Dunbrack RL (2009) Improved prediction of protein side‐chain conformations with SCWRL4. *Proteins: Structure, Function, and Bioinformatics* 77(4):778-795.
20. Jamroz M & Kolinski A (2013) ClusCo: clustering and comparison of protein models. *BMC bioinformatics* 14:62.
21. Ester M, Kriegel H-P, Sander J, & Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*, pp 226-231.
22. Dunbrack RL (2002) Rotamer Libraries in the 21 st Century. *Current opinion in structural biology* 12(4):431-440.
23. Lovell SC, Word JM, Richardson JS, & Richardson DC (2000) The penultimate rotamer library. *Proteins: Structure, Function, and Bioinformatics* 40(3):389-408.
24. Needleman SB & Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology* 48(3):443-453.
25. Gotoh O (1982) An improved algorithm for matching biological sequences. *Journal of molecular biology* 162(3):705-708.

26. Blazewicz J, Frohmberg W, Kierzynka M, Pesch E, & Wojciechowski P (2011) Protein alignment algorithms with an efficient backtracking routine on multiple GPUs. *BMC bioinformatics* 12(1):1.
27. Sali A & Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology* 234(3):779-815.
28. Kim DE, Chivian D, & Baker D (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic acids research* 32(suppl 2):W526-W531.

B.3 Supporting Information

Supporting Information

Automatic Protein Structure Modelling into Cryo-EM Density Maps using EMfasa

Tatjana Braun[1*], Zhe Wang[1], Gunnar F. Schröder[1,2*]

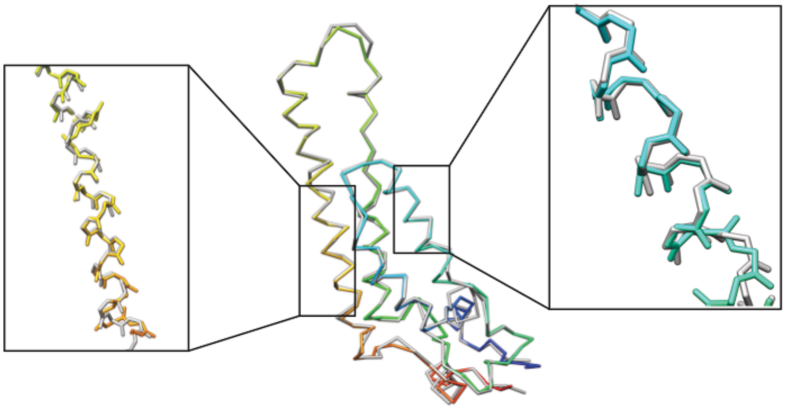

Figure S1: all-Alanine model for TMV. The native structure of TMV (gray) and the generated all-Alanine model (rainbow) are shown in ribbon representation. The close-ups show the atoms and bonds of the main chain (excluding side-chains) of both structures.

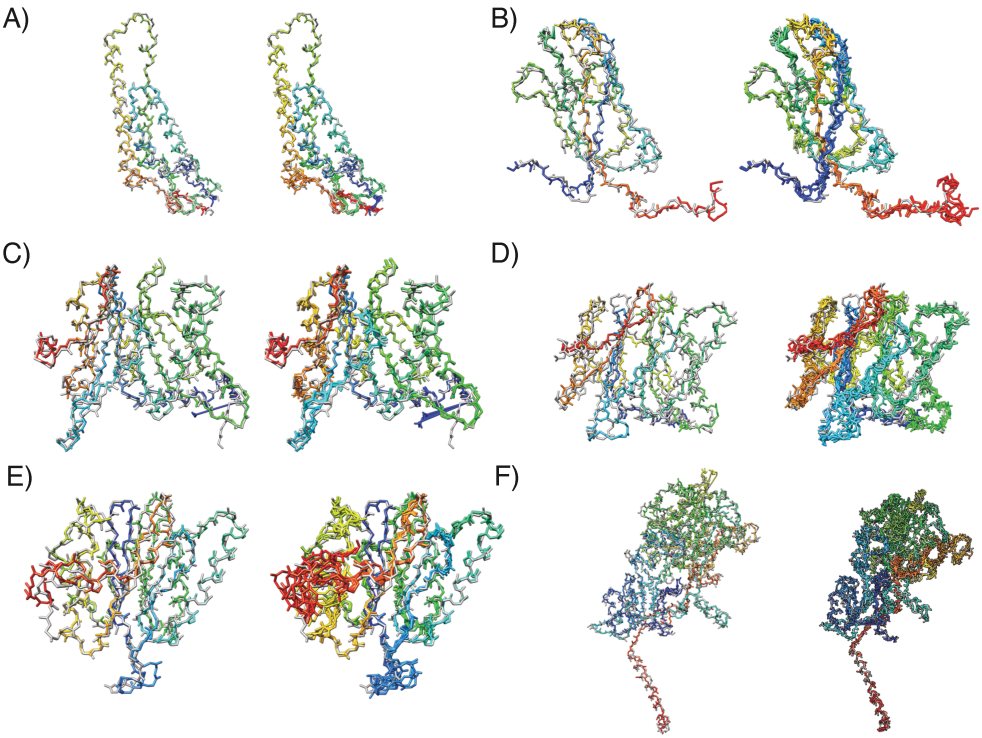

Figure S2: all-Alanine structures for all 6 test systems. The all-alanine structures are shown in rainbow color and the native protein backbone in gray for A) TMV B) BMV C) 20S-alpha (3.3Å) D) 20S-alpha (4.8Å) E) 20S-beta (3.3Å) and F) nicastrin. For each target, the left panel shows one all-Alanine structure while the right one shows five different ones to give an idea about the variability of the different all-alanine structures.

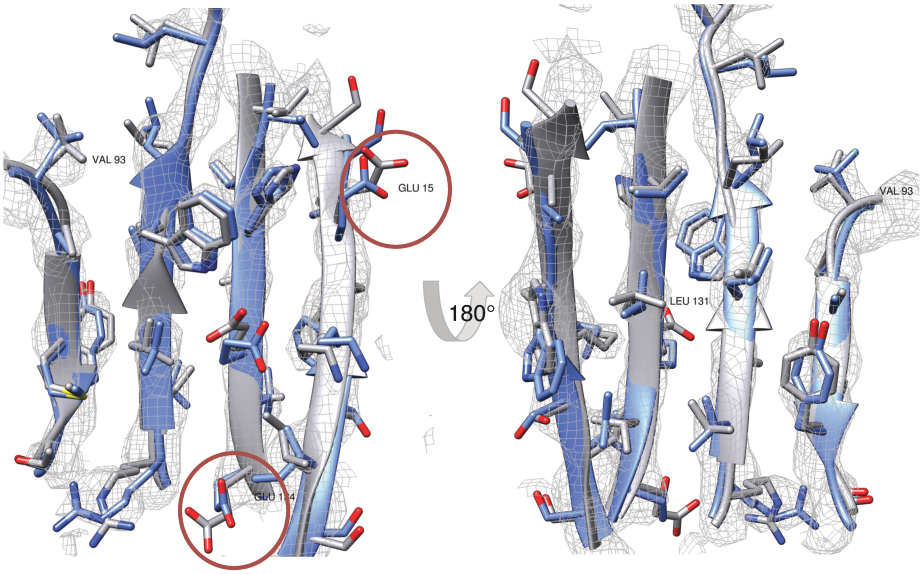

Figure S3: Close-up of a BMV model. This close-up shows the good agreement between the sheets of one of the generated EMFasa models (blue) and the native protein structure (gray). The red circles highlight regions with low density where EMfasa was still able to accurately assign the side-chains.

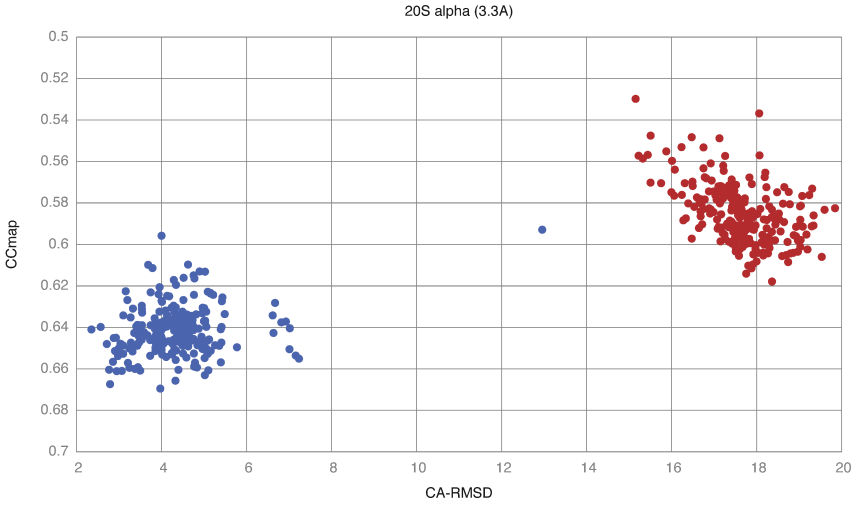

Figure S4: **EMfasa results for 20S-alpha, 3.3Å**. CCmap vs. Calpha-RMSD plot for full-atom models obtained using a topologically correct Cα-trace in the correct (blue) and inverse (inverse) direction.

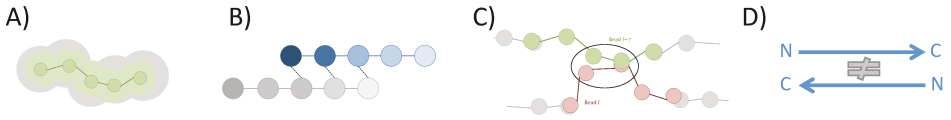

Figure S5: Score-terms during MC-SA. The compatibility of a set of fragments is calculated using a scoring function consisting of 4 different terms: A) $score_{corr}$ B) $score_{overlap}$ C) $score_{clash}$ and D) $score_{dir}$.

Target	EMDB / Res.[a]	EMfasa	
		Ca-trace	
		TM-score > 0.7 in 10 best-scoring traces (Cfree)	TM-score > 0.7 in total set of traces
Tobacco Mosaic Virus	2842 / 3.4	2/10 (6/10)*	17/100
Brome Mosaic Virus	6000 / 3.8	8/10	40/100
T20SProteasome (alpha Subunit)	5623 / 3.3	7/10	72/100
	6219 / 4.8	1/10	4/100
T20SProteasome (beta Subunit)	5623 / 3.3	4/10	39/100
y-Secretase (Nicastrin)	3061 / 3.4	2/10	35/100

Table S1: Calpha-traces generated for all six test systems. The residue-residue alignment between the Calpha-traces and the respective native structures was calculated with TM-align. We consider traces with TM-scores > 0.7 to have the same topology as the native structure. *) Number of traces with correct topology determined manually. The termini of TMV are very close together and the tracing algorithm is not always able to determine them correctly.

Target	EMDB / Res.[a]	PDB entry / Size[b]	Fold	Rosetta			
				Round 1		Round 2	
				Ca-RMSD (Total Residues) [1]	Ca-RMSD (Corr. Assignment / Total Res) [2]	Ca-RMSD (Total Residues) [1]	Ca-RMSD (Corr. Assignment / Total Res) [2]
Tobacco Mosaic Virus	2842 / 3.3	4udv / 153	α	1.2 (122)	0.8 (121/122)	1.3 (123)	0.7 (117/123)
Brome Mosaic Virus	6000 / 3.8	3j7l / 149	β	9.3 (60)	0.8 (30/60)	14.1 (78)	0.8 (32/78)
T20SProteasome (alpha Subunit)	5623 / 3.3	3j9i / 224	α/β	6.0 (164)	1.1 (153/164)	5.3 (187)	0.7 (164/187)
	6219 / 4.8	1pma / 221		6.5 (14)	1.2 (11/14)	6.5 (14)	1.0 (9/14)
T20SProteasome (beta Subunit)	5623 / 3.3	3j9i / 203	α/β	7.7 (153)	0.6 (136/153)	7.2 (171)	0.72 (153/171)
y-Secretase (Nicastrin)	3061 / 3.4	5a63 / 665	α/β	2.9 (26)	0.7 (21/26)	-	-

Table S2: Rosetta results for all six test systems. The *denovo_density* tool of the Rosetta software suite was run for two iterations. The model obtained after the first round was used as starting structure for the second round. For nicastrin, the number of translation searches was set to approx. 2x the number of residues in the sequence (as suggested in the tutorial). For the other 5 test systems, the number of translations to search was set to approx. 10x their sequence length. The results presented here might be improved by further increasing the number of translation searches and the total number of iterations, greatly increasing the runtime. [1] Calpha-RMSD of the common residues based on structural alignment with TMalign. The total number of residues in the model is shown in brackets. [2] Ca-RMSD of the correctly assigned/placed residues (=residues are placed within 2Å of the native residue).

C. Embedded Publication 2

Please find the full article including the supplemental information on PNAS Online using the following reference:

Braun T, Vos MR, Kalisman N, Sherman NE, Rachel R, Wirth R, Schröder GF, Egelman EH. *Archaeal flagellin combines a bacterial type IV pilin domain with an Ig-like domain.* Proc Natl Acad Sci U S A. 2016 Sep 13;113(37):10352-7.
PMID: 27578865, PMCID: PMC5027424, DOI: 10.1073/pnas.1607756113

D. Tables

Table D.1 Fragment library validation set

PDB ID	Length (Residues)	Resolution (Å)	SCOP class
1AIU	105	2.00	α/β
1NAT	124	2.45	α/β
2RN2	155	1.48	α/β
1ILW	180	2.05	α/β
1VL1	232	1.55	α/β
1XWY	264	2.00	α/β
2HVM	273	1.80	α/β
1OBR	326	2.3	α/β
1VFF	423	2.5	α/β
1SMD	496	1.6	α/β
1WM3	72	1.2	$\alpha+\beta$
1CEW	107	2.0	$\alpha+\beta$
1EKG	127	1.8	$\alpha+\beta$
1Z2U	150	1.1	$\alpha+\beta$
1SQW	188	1.90	$\alpha+\beta$
1XKR	203	1.75	$\alpha+\beta$
1W66	232	1.08	$\alpha+\beta$
1RL0	255	1.4	$\alpha+\beta$
2YVT	260	1.60	$\alpha+\beta$
1MSK	331	1.8	$\alpha+\beta$
1AYE	401	1.8	$\alpha+\beta$
1B4V	504	1.5	$\alpha+\beta$
1CSP	67	2.45	all β
1BMG	98	2.5	all β
1XD6	112	2.0	all β
1NEP	130	1.7	all β
1CZT	160	1.87	all β
1T9F	187	2.00	all β
2AYH	214	1.6	all β
1P6F	241	2.2	all β
1SEF	274	2.05	all β
1WL7	312	1.9	all β
1OKQ	394	2.80	all β
1ENH	54	2.1	all α
2J9V	99	2.0	all α
2MHR	118	1.3	all α
1JWF	147	2.10	all α
1SFE	180	2.10	all α
1SDI	213	1.65	all α
1VIN	268	2.0	all α
1V5C	386	2.0	all α

41 proteins comprising the validation set, taken from [185]. Proteins are single-doman, single chain, and belong to distinct PFAM families.