Introduction to Protein Structure Prediction

Abeln, Sanne; Heringa, Jaap; Feenstra, K. Anton

published in
Introduction to Structural Bioinformatics
2017

Structural Bioinformatics

K. Anton Feenstra **Sanne Abeln**

Centre for Integrative Bioinformatics (IBIVU), and
Department of Computer Science,
Vrije Universiteit, De Boelelaan 1081A, 1081 HV Amsterdam, Netherlands

December 4, 2017

Abstract

This chapter gives a graceful introduction to problem of protein three-dimensional structure prediction, and focuses on how to make structural sense out of a single input sequence with unknown structure, the 'query' or 'target' sequence. We give an overview of the different classes of modelling techniques, notably template-based and template free. We also discuss the way in which structural predictions are validated within the global community, and elaborate on the extent to which predicted structures may be trusted and used in practice. Finally we discuss whether the concept of a single fold pertaining to a protein structure is sustainable given recent insights. In short, we conclude that the general protein three-dimensional structure prediction problem remains unsolved, especially if we desire quantitative predictions. However, if a homologous structural template is available in the PDB model or reasonable to high accuracy may be generated.

Contents

Chapter 7

Introduction to Protein Structure Prediction

Sanne Abeln Jaap Heringa K. Anton Feenstra

Centre for Integrative Bioinformatics (IBIVU) and
Department of Computer Science,
Vrije Universiteit, De Boelelaan 1081A, 1081 HV Amsterdam, Netherlands

7.1 What is the protein structure prediction problem?

7.1.1 Predicting the structure for a protein sequence

This chapter revolves around a simple question: "given an amino acid sequence, what is the folded structure of the protein?" (Figure 7.1) Even though this seems like a simple question, the answer is far from straightforward. In fact, whether we can give an answer at all depends heavily on the sequence in question and available protein structures that can be used as modelling templates. While the number of structures deposited in the Protein Data Bank (PDB) (Berman et al., 2000) continues to rise rapidly [1], the number of sequenced genes rises much faster. The large and widening gap between protein structures and sequences makes structure prediction an important problem to solve. Fortunately, recently developed methods can use these large resources of sequence data to increase the quality of some predictions. Here, we will give an overview of current structure prediction methods, and describe some tools that provide insight into how reliable the structure predicted will be.

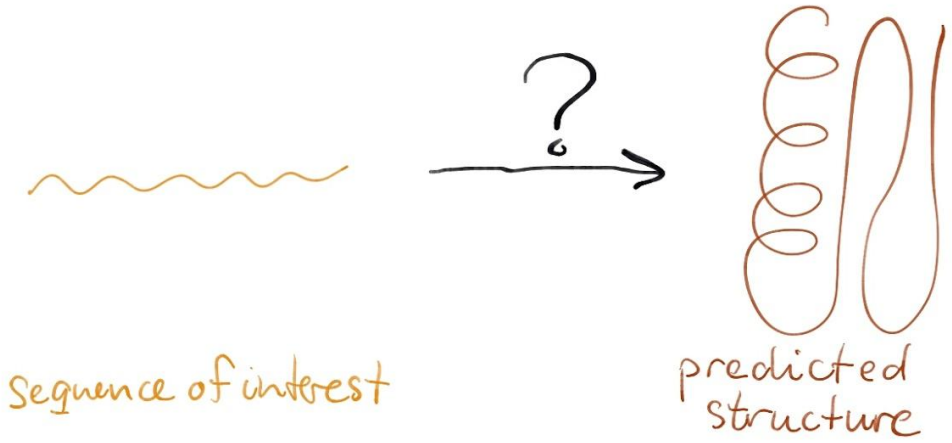

Figure 7.1: Structure prediction methods try to answer the question: given an amino acid sequence, what is the folded protein structure?

The typical problem is that we want to generate a structural model for a protein with a sequence, but without an experimentally determined structure. In this chapter, we will build up a workflow for tackling this problem, starting from the easy options that, if applicable, are likely to generate a good structural model, and gradually working up to the more hypothetical options whose results are much more uncertain.

Another very important remark is in place here: the modelling strategy should depend heavily on what we want to do with the structure. Do we

[1] https://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=100

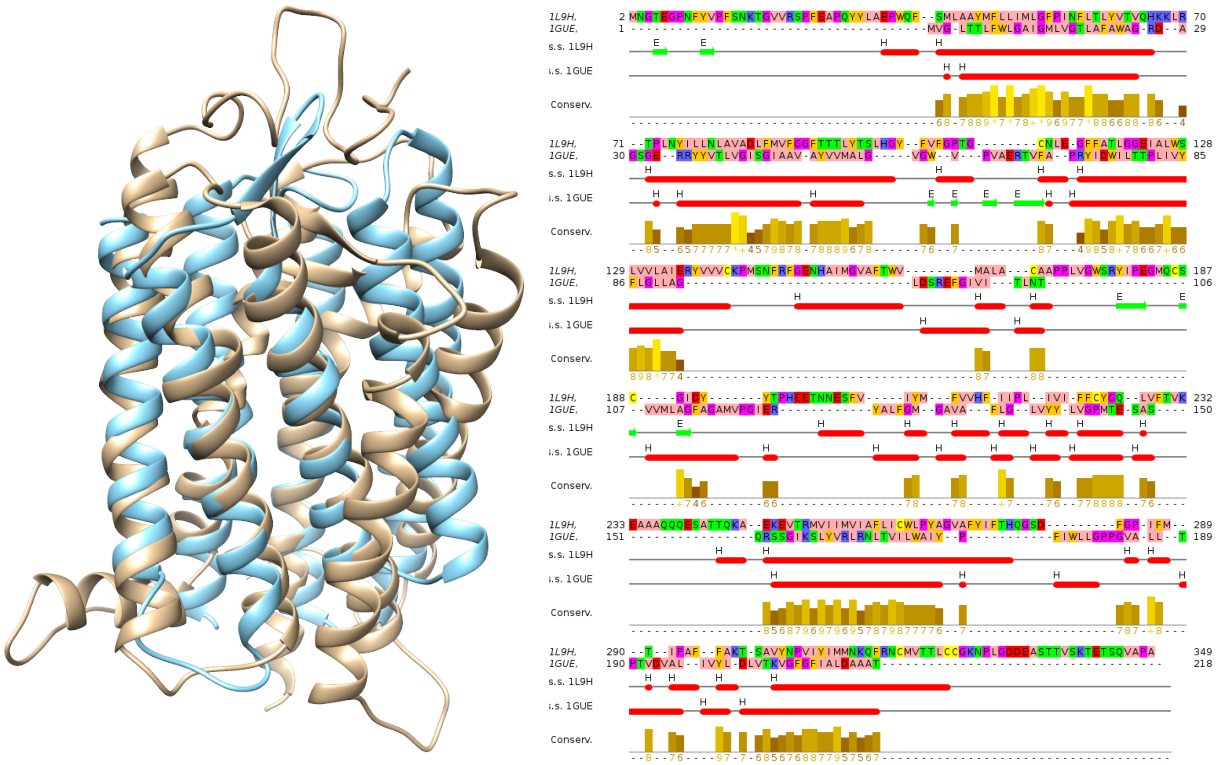

Figure 7.2: Protein structure more conserved than sequence. Here the output of a structural alignment is shown on the left, created using Chimera [2] (Pettersen et al., 2004). The structural alignment shows both proteins are highly similar; the RMSD is 2.3 over 144 aligned residues. Furthermore, the function of the two proteins, one from cattle (PDB:1L9H, light brown) and one from a archaeon (PDB:1GUE, light blue), is similar: both are light sensitive rhodopsins, used for vision and phototaxis, respectively. However, as can be seen in the sequence alignment on the right, the sequence identity is only 7%. This is lower than would be expected for any two random sequences. The alignment shown is based on the structural alignment on the left, and visualised using JalView (Waterhouse et al., 2009).

want to predict where the functional site of the protein is, whether a specific substrate binds, or if a certain residue may be exposed to the surface? These different questions imply a different degree of accuracy in the answer, and may lead to choices regarding technology and methods to carry out these predictions. It is important to keep in mind that one of the most important aspects of any scientific model is whether a research question may be answered with the model produced or not. Even if we do have an experimental structure available, some of these questions may not be straightforward to answer; we will come back to this issue later in the chapter.

7.1.2 Structure is more conserved than sequence

Almost all structure prediction relies on the fact that, for two homologous proteins, structure is more conserved than sequence (see Figure 7.2). The real power of this observation manifests itself when we turn this statement around: if two protein sequences are similar, these two proteins are likely to have a very similar structure. The latter statement has very important consequences. It means that if our sequence of interest is similar to a protein sequence with a known structure, we have a good starting point for a structural model. In such a scenario we use sequence similarity, suggesting an homologous relation between the proteins, to predict the structure. The vast majority of accurate structure prediction methods use structure conservation as an underlying principle; while methods that have been developed to deal with the more difficult modelling questions, exploit the sequence-structure-conservation relation in an advanced manner, as discussed towards the end of this chapter.

7.1.3 Terminology in structure prediction

Firstly, we should take care to lay down a good problem definition. Here we will generously borrow the nomenclature from the Critical Assessment of Protein Structures (CASP). CASP is a scientific competition, in which structure prediction groups and structure prediction servers compete to predict the structure for an unknown sequence, that has been running since 1995 (Moult et al., 1995). The sequence for which we will predict a structure is called the *target* sequence. If there is a suitable structure to build a model for our query sequence we call this structure a *template*, see also Figure 7.3. Using the structure of the template and using the *sequence alignment* between the template and the target sequence, we can create one or more structural *models*: the predicted structure, for a target sequence. In CASP structural models from different prediction methods, are compared to the experimentally determined solution or target structure.

7.1.4 Different classes of structure prediction methods

We can classify structure prediction strategies into two categories of difficulty: template-based modelling, and template-free modelling (see Figure 7.4). In the first case, it is possible to find a suitable template for the target sequence in the PDB, as a basis for the model, whereas for template-free modelling no such experimental structure is available. Note that it may not be trivial at all to find out in which of these two categories a structure prediction problem falls. Only if we can find a close homolog – based on sequence

[2]Molecular graphics and analyses were performed with the UCSF Chimera package. Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIGMS P41-GM103311).

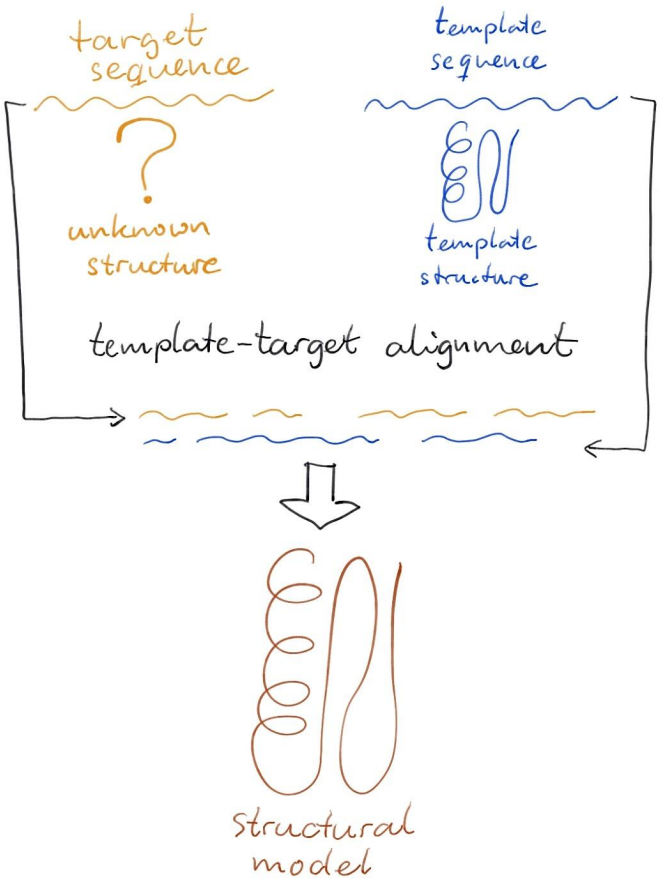

Figure 7.3: Terminology used in protein structure prediction. We start from our protein of interest (with no known structure): the target sequence. First step is find a matching protein: a template sequence with known structure; the template structure. We then create a template-target sequence alignment, and from this alignment create the structural model which is the solution structure for our target protein.

similarity – in the PDB we can be sure that a template based modelling strategy will suffice; this is also referred to as homology modelling. With a template, the constraints from the alignment between the model and the template sequence, in addition to the template structure, will give sufficient constraints to build a structural model for the target sequence. Even in this case, small missing substructures in the alignment, e.g. loops, may require a template-free modelling strategy.

If no close homologs are available in the PDB, we may need to use more advanced template finding strategies, such as remote homology detection or fold recognition methods.

If no suitable template is available, we will need to resort to a template-

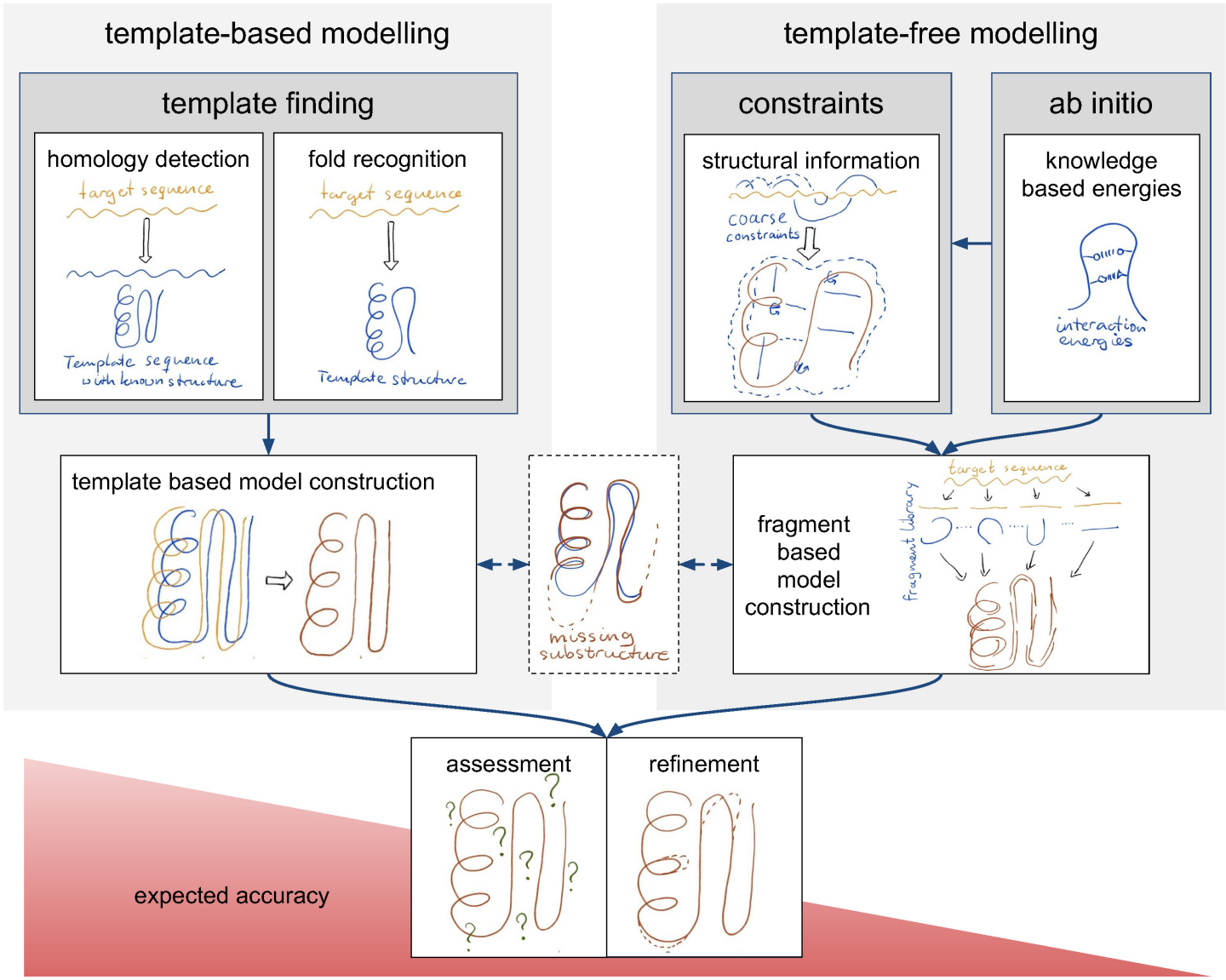

Figure 7.4: Overview of Structure Prediction. Template-based modelling: a template is found on the basis of homology between the template and the target. Fold recognition: no obvious homologous structure can be found in the PDB, we need fold recognition methods to find a suitable template. Template-free modelling: no suitable template for protein domains can be found. Without template, we need to use a combination of coarse constraints from experiment or co-evolution analysis, and ab initio prediction. Ab initio methods typically work with taking fragment templates from various proteins, and assemble these into a model or decoy structure. Expected model accuracy declines from left to right: good accuracy is expected if based on homology; in contrast, ab initio modelling should only be considered if no other options remain.

free modelling strategy. In the "ab initio" approach knowledge-based energy terms are used to generate structural models based on the sequence of the template alone. Small, suitable fragments, from various PDB structures are assembled to generate possible structural models. In some cases, we can find additional constraints, for example from experiments, such as NMR or cryoEM, or from contact prediction methods; in that case we have a much better chance of building a suitable model (Moult et al., 2016). In fact, we could consider such constraints an alternative for the constraints provided by homology.

Lastly, several steps may be taken to refine the model, and to select the most likely model, from several model building attempts. Note that some structure prediction methods, may also include variations of model refinement and model selection steps higher up in the modelling workflow.

7.1.5 Domains

So far we have implied that we may follow the above strategy for an entire protein, however, this generally is not the case. In fact many proteins consist of multiple domains. If this is the case, it is wise to also run one or multiple disorder prediction methods on the target sequence; Any large regions (> 25 residues) predicted to be disordered should be left out for further structure prediction and template finding.

Most structure prediction methods only work well at the domain level. This means that a sequence first needs to be split in multiple domains, before we can start to make models. However, domain splitting is often ambiguous both given the sequence and the structure, while combining models built from various domains is far from trivial. In practice, this means that multiple templates might be necessary for a single target sequence and that it is difficult to resolve the orientation of the modelled domains with respect to each other.

Predicting the orientation for several domains is currently an unsolved problem, unless there is a suitable, homologous, template available – with the domains in the same orientation. In some cases, coarse constraints on the domain orientations such as data from small-angle scattering experiments, or distance restraints from NMR, chemical cross-links or co-evolution may help to put different homology models in the correct orientation.

Typically, it only makes sense to generate a model, be it template-based or template-free, for a single domain. In fact, in CASP model predictions are assessed per structural domain, separately. Therefore, it is essential to split the target sequence into its constituent domains – which is a non-trivial task, particularly if no homologous templates are available for each of the domains.

7.2 Assessing the quality of structure prediction methods

Generally, as with any prediction problem, we can assess the quality of a prediction if we have a true answer to the question. Here, truth will be represented by an experimentally determined protein structure (of high quality). Fortunately, there are now (November 2017) over 120,000 deposited protein structures in the PDB[3]. However, simply assessing how well a method performs over this set is problematic. The methods have been trained on this data set; that means that there may be a strong bias in these methods, to predict good models for sequences that are within their dataset, and therefore homologs of those. In order to truly assess a method, a completely independent data set is required.

7.2.1 Critical Assessment of protein Structure Prediction

Every other year CASP, a Critical Assessment of Protein Structure Prediction, provides such an independent validation benchmark. CASP is a blind test or competition: experimentalists provide sequences for which they know the structures will be solved imminently; modelling groups and servers try to predict the structure (Moult et al., 1995). Once the structure is solved, the models can be evaluated using the solution structure of the target (see also Figure 7.5).

CASP was started because the protein structure prediction problem was claimed to have been solved several times. The problem was, that algorithms were trained on databases that contained the structures that were evaluated in benchmarking tests. CASP overcomes this problem.

Note that the very first step in any practical structure prediction approach, should be to inspect the results from the latest CASP round (Moult et al., 2016) via the CASP website[4] to see what the state of the art methods are, and what their expected performance is.

7.2.2 Root-Mean-Square Deviation (RMSD):

If we want to asses the quality of a method, we need to measure the quality of the predictions made by the method. Hence, one would like to structurally compare atomic coordinates of the model and of the solution structure, and quantify the (dis)similarity.

The problem of comparing a model to to a solution structure, is less difficult than the comparison between two homologous protein structures. This is because the alignment is trivial: the model has the same sequence as

[3] structures in the PDB: https://www.rcsb.org/pdb/statistics/holdings.do
[4] CASP website: http://predictioncenter.org/

the solution structure; we know which residues, and atoms should correspond in the two structures.

The easiest way to compare structures, is the calculate the Root-Mean-Square Deviation (RMSD) after a structural superpositioning (Marti-Renom et al., 2009). The superpositioning is required, because two arbitrary structures will typically not be positioned at coordinates suitable for comparison; first a translation and rotation needs to be applied to one of the two structures, to minimise the RMSD; the resulting RMSD after superpositioning can be used as a dissimilarity measure.

The Root-Mean-Square Deviation (RMSD) calculates the squared difference between two sets of atoms, and can be defined as follows:

$$\begin{aligned} RMSD(v,w) &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|v_i - w_i\|^2} \\ &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_{ix}-w_{ix})^2 + (v_{iz}-w_{iz})^2 + (v_{iz}-w_{iz})^2} \end{aligned}$$

Here, v_i is the position vector of i^{th} atom of structure v; w_i is the position vector of i^{th} atom of structure w; and n is the total number of aligned atoms.

The RMSD takes the average over all aligned pairs. In protein structures typically one representative atom per residue is chosen, such as Cα or Cβ.

7.2.3 GDT – Global Distance Test

If a model gets a loop very wrong, it tends to stick out and can be positioned very distant from the true structure, even though the remaining structure may be reasonably accurate. This partial outlier weighs heavily on the average distance calculated. Hence the RMSD is over sensitive to such outliers.

The global distance test total score (GDT_TS) is a more robust structural similarity measure that is well defined given an alignment between two structures. The key idea is to count the number of residues that can maximally be fitted within a certain distance cutoff, see also Figure 7.5. The GDT score will therefore produce a percentage. In the formula below, the final score is the average over four different distance cutoffs (1, 2, 4, 8 Å).

$$GDT\_TS = \frac{1}{4}\sum_{v=1,2,4,8\text{Å}}\frac{G(v)}{t} \tag{7.1}$$

Here, $G(v)$ is the number of aligned residues within given RMSD cutoff v (in Ångstrom – $10^{-10}m$) and t is the total number of aligned residues. A related score called GDT_HA was introduced in CASP some time ago (Read

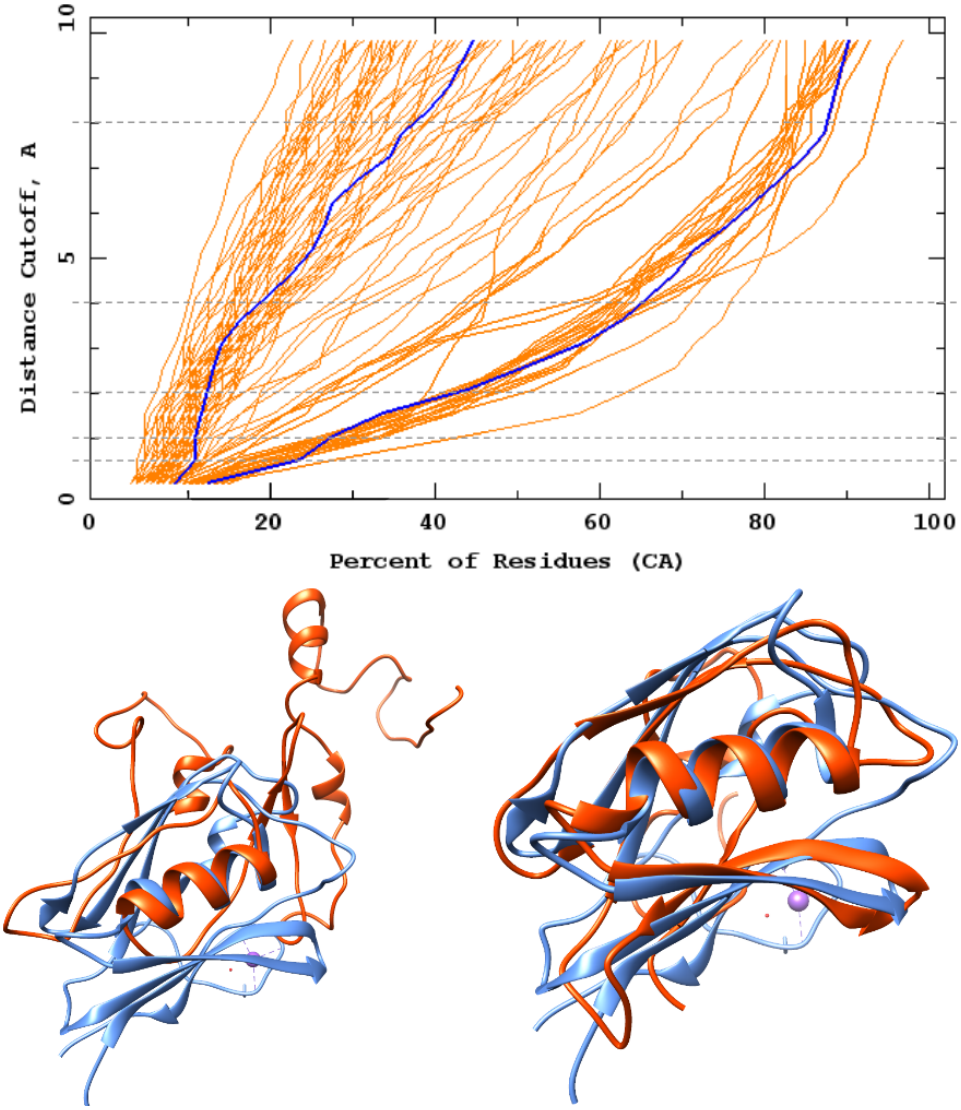

Figure 7.5: Example of structural comparison for the target T0886-D2 and two models submitted to CASP12. The top panel shows individual traces for all models generated for this target; the distance cutoff (vertical axis, in Å) is plotted against the fraction of residues (horizontal axis, in %) that can be aligned within this cutoff. The traces were obtained from predictioncenter.org/casp12. The dotted lines indicate the thresholds used in the GDT_TS (1, 2, 4, 8 Å) and GDT_HA (0.5, 1, 2, 4 Å) scores. Two models are highlighted in blue: a bad model (TS236, GDT_TS=18.90) on the left, and a good model (TS173; GDT_TS=51.97) on the right. Both model structures are also shown in the panels below in red, superposed onto the solution crystal structure in blue (PDB:5FHY). Structural superposition created using LGA at proteinmodel.org/AS2TS/LGA/ (Zemla, 2003), 3D visualisation using Chimera 1.11.2 (Pettersen et al., 2004).

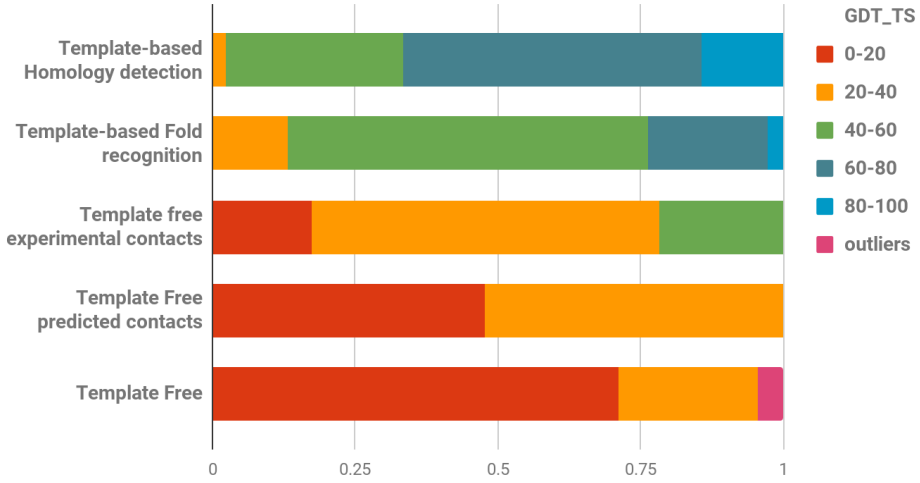

Figure 7.6: Distribution of GDT_TS scores for the different model categories in CASP11 for template-based (Modi et al., 2016), template-free with contact information (Kinch et al., 2016a) and template-free (Kinch et al., 2016b). The legend coloring corresponds to the GDT_TS scores, the bars indicate the fraction of models in each GDT_TS range for the six categories (GDT_TS scores for (Modi et al., 2016) were estimated from the reported GDT_HA scores using their Figure 4A). "Outliers" targets have unusually high GDT_TS due to being very short (∼ 50 residue) with extended structures. Targets selected for server prediction (top bar) were considered easier than those for human prediction (second from top), average sequence identity was 26% vs. 20%, respectively. It is clear that overall prediction accuracy sharply declines going down this list of categories. For template-free modelling, the quality of contact information used is crucial. Experimental information (from chemical cross linking or simulated NMR) can give reasonable models. Predicted contacts do not guarantee that an acceptable model can be obtained, but without even predicted contacts, more than two-thirds of models are at most 20% correct.

and Chavali, 2007) using stricter distance cutoffs (0.5, 1, 2, 4 Å), to cater for targets in the template-based modelling category where very high accuracies can be realized:

For a typical "difficult" CASP target no model even comes close to the experimentally solved structure; typical results would be similar to the left-most model in Figure 7.5. If we have a look at the latest CASP results one will see that a performance of GDT_TS< 20% is not an exception. In other words, the protein structure prediction problem has NOT yet been solved, especially not if one considers targets without a good template structure.

7.2.4 How difficult is it to predict?

Overall, if one can find a good template, the quality of the predicted model will be relatively good. CASP results show that for homology modelling

based on close homologues, it is possible to obtain models similar to the experimentally determined structure (Moult et al., 2016). The modelled structure will typically have a good accuracy for the regions that can be well aligned between the target and template (using the sequences). The top two bars in Figure 7.6 shows that one may expect the majority of such models to be accurate for $> 50\%$ of their residues. Gaps in an alignment will typically lie in loop regions of a structure and are more difficult to model. So, if we are interested in a large loop region that is not present in our template, we still may not be able to answer our scientific question with the resulting model structure (Moult et al., 2016).

If no acceptable template can be found, the chances of successfully answering our scientific question will become very low. As a last resort, ab initio modelling can provide us with structural models. Typically, ab initio methods use very small templates from various proteins (see Figure 7.4). The state of the art is that on average one may expect to find one structure that looks somewhat like the solution structure for the target among the top five or ten models (Moult et al., 2016). However, be aware that the best model is typically not recognised as being the best through the scores of the prediction program. In Figure 7.6 one sees that very clearly in the bottom few bars: without template, even with predicted contacts, one may have less than 20% of the structure correct in the majority of models; even in the best cases at most 40% of the residues are modelled accurately.

7.2.5 For which gene sequences can we predict a three-dimensional structure?

If and only if there is a structure of a homologous protein present in the PDB, it is possible to generate a structural model of reasonable accuracy. Based on this notion, we can estimate for which (fraction of) gene sequences it is possible to predict a structure. This way it has been estimated that for a about 44% of residues in Eukaryotic gene sequences, we cannot yet make a homology model, and 15% of these residues lie within a gene for which we can not make a homology model for a single domain (Perdigão et al., 2015). Especially membrane proteins are underrepresented in the PDB, due to the experimental difficulty of determining these structures. Note that these residues, may also lie in natively disordered regions (see also Section 7.3).

Similarly, it is possible to predict the range of protein structures present in an organism, based on the gene sequences their completed genome. This reveals that there is a subset of protein structures, that is present in nearly all organisms, for example TIM-barrels or Rossmann-folds (Abeln and Deane, 2005; Edwards et al., 2013). Nevertheless, there is also a group of structures that is extremely lineage specific. It is to be expected that for this type of protein structures, many new structures remain to be discovered. This also

implies that it will remain difficult to find suitable templates for homology modelling for these lineage specific protein families.

7.2.6 How accurate do we need to be?

We already mentioned that we may approach the modelling of a protein structure of interest differently, depending on the biological question we want to ask,e.g. which residues are likely to be crucial for the functioning of the protein. Sometimes an answer to the research question may be possible in a simpler way, without full-scale prediction of the protein structure, e.g. by direct prediction of the impact of certain mutations or of protein-protein interaction sites. Examples of fully-automated webservers that do just that, are HOPE – (Venselaar et al., 2010) and SeRenDIP (Hou et al., 2017). In some cases, a rough homology model inspires the understanding of experimental results, spurring forward the project and eventually ending with crystal structures highlighting the protein function (in this case, protein-protein interactions) of interest (e.g., De Vries-van Leeuwen et al., 2013). Also, specifically for enzymes, such as for example cytochromes P450, modelling of the protein structure should be done combination with that of the ligand (de Graaf et al., 2005).

In CASP11, three functional aspects were explictly scored, selected on being able to qualitatively evaluate them: multimeric state, (small) ligand binding, and mutation impact. Targets were selected that in solved crystal structure were dimeric, or had a ligand bound, or where from the crystallographers or in literature interest was expressed for evaluating mutants (Huwe et al., 2016).

For prediction of dimer structures, only in two cases out of ten a dimer model with reasonable accuracy could be generated for the majority of monomer model structures (Huwe et al., 2016). In the critical assessment of prediction of protein interaction (CAPRI) between 30-80% of models were of 'acceptable' or 'medium' quality for easy dimer targets, while for harder targets (difficult dimers, multimers and heteromers), this fraction dropped to below 10% (Lensink et al., 2016). Encouragingly, it was seen that also structure models of lower quality could sometimes lead to acceptable or even medium quality models of the bound proteins (Lensink et al., 2016).

For ligand binding, it was found that the accuracy of even the best models ($\sim$ 2Å) are not good enough for accurate ligand docking; the best ligands were around 5Å RMSD (Huwe et al., 2016). Something similar was found for mutation impact prediction; for most targets, model accuracy did not correlate with accuracy of impact prediction (Huwe et al., 2016). Apparently, either homology models are not yet accurate enough for this purpose, or methods are tuned to particular characteristics of crystal structures.

7.3 Is there such a concept as a single native fold?

Before we conclude, we should consider a more physical description of protein structure. In fact, protein folding from a physical point of view is a very interesting process: given a sequence, a protein tends to fold always, and exactly into the same functional structure. In material design, it is extremely difficult to mimic such high specificity. The apparent observation of folding specificity also leads to the question, is there such a concept as a single native fold? Or, more pragmatically, is sequence-to-structure truly a one-to-one relation?

In fact, if one wants to start making quantitative predictions, such as the stability of a protein fold, or the binding strength between two proteins in terms of free energy, it is much more helpful to think in ensembles of structural configurations for a protein sequence (e.g. May et al., 2014; Pucci et al., 2017). The probability to find a protein in a specific ensemble of structural configurations will depend on conditions such as the presence or absence of binding partners, the pressure, the pH or the temperature (e.g. van Dijk et al., 2015, 2016). There are a few specific cases, common cases, for which even the functional or biologically relevant structural ensembles do not resemble a single globular folded structure.

7.3.1 Disordered proteins

Not all proteins fold into single configurations, some proteins stay natively unfolded, i.e. they can take up a a large variety of more extended, and very different configurations (Uversky et al., 2000; Mészáros et al., 2007). Some disordered regions contain elements that do form stable structures upon binding. The regions that remain disordered are thought to be important to prevent aggregation within the cell (Abeln and Frenkel, 2008). Missing residues in X-ray structures are typically removed for crystallization; for this reason disorder prediction methods have been developed. Disordered regions are relatively easy to predict in protein sequences just like secondary structures; broadly speaking, prediction can be based on the large amount of charged/polar (hydrophilic) amino acids in combination with the presence of amino acids that disrupt the secondary structure (proline and glycine) in these regions (Oldfield et al., 2005; Wang et al., 2016). We know sequences of many proteins contain large disordered segments (33% of eukaryotic, 2% archaeal, and 4% bacterial proteins).

7.3.2 Allostery and functional structural ensembles

It is important to realize that one protein, typically, does not correspond to one defined three-dimensional structure. Disordered regions or proteins are one particularly salient case, but also proteins which fold into specific three-

dimensional configurations, may exist in multiple functional states each with a specific structure. The biological question of interest dictates which state is relevant. Most proteins have only been crystallized in one particular state, and often it is not known to which biological condition this crystal structure may correspond. One may have cases where a homology model of the relevant state may be preferred over a crystal structure of a different or unknown state (e.g., de Graaf et al., 2005).

7.3.3 Amyloid fibrils

Lastly, we should consider a competing state of folded proteins: the aggregated state, where multiple peptide chains clog together in fibrillar structures or amorphous aggregates. Amyloid fibres are formed by β-strands formed between different protein or peptide (small protein) chains. Fibril formation is associated with various neurodegenerative diseases, such as Alzheimer's, Creutzfeldt-Jakob and Parkinson's (Chiti and Dobson, 2006). In fact, the fibrillar state is more favorable than the state of separately folded structures for several protein types. The general cellular toxicity of such aggregates, puts evolutionary pressure on avoiding structural characteristics on the surface of proteins; hence it is extremely rare to observe solvent accessible β-strand edges or large hydrophobic surface patches (Richardson and Richardson, 2002; Abeln and Frenkel, 2011). The propensity proteins have to form Amyloid fibrils is relatively easy to predict (Graña-Montes et al., 2017). However, reference databases are still small so it is difficult to verify such methods.

7.4 Acknowledgements

We thank Nicola Bonzanni, Kamil K. Belau, Ashley Gallagher, Jochem Bijlard for insightful discussions and critical proofreading of early versions.

Bibliography

Abeln, S. and Deane, C. M. (2005). Fold usage on genomes and protein fold evolution. *Proteins*, 60(4):690–700.

Abeln, S. and Frenkel, D. (2008). Disordered flanks prevent peptide aggregation. *PLoS Comput. Biol.*, 4(12):e1000241.

Abeln, S. and Frenkel, D. (2011). Accounting for protein-solvent contacts facilitates design of nonaggregating lattice proteins. *Biophys. J.*, 100(3):693–700.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–242.

Chiti, F. and Dobson, C. M. (2006). Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.*, 75:333–366.

de Graaf, C., Vermeulen, N. P. E., and Feenstra, K. A. (2005). Cytochrome P450 in Silico: An Integrative Modeling Approach. *Journal of Medicinal Chemistry*, 48(8):2725–2755.

De Vries-van Leeuwen, I. J., da Costa Pereira, D., Flach, K. D., Piersma, S. R., Haase, C., Bier, D., Yalcin, Z., Michalides, R., Feenstra, K. A., Jiménez, C. R., de Greef, T. F. A., Brunsveld, L., Ottmann, C., Zwart, W., and de Boer, A. H. (2013). Interaction of 14-3-3 proteins with the estrogen receptor alpha F domain provides a drug target interface. *Proceedings of the National Academy of Sciences of the United States of America*, 110(22):8894–9.

Edwards, H., Abeln, S., and Deane, C. M. (2013). Exploring Fold Space Preferences of New-born and Ancient Protein Superfamilies. *PLoS computational biology*, 9(11):e1003325.

Graña-Montes, R., Pujols-Pujol, J., Gómez-Picanyol, C., and Ventura, S. (2017). Prediction of Protein Aggregation and Amyloid Formation. In *From Protein Structure to Function with Bioinformatics*, pages 205–263. Springer Netherlands, Dordrecht.

Hou, Q., De Geest, P., Vranken, W., Heringa, J., and Feenstra, K. (2017). Seeing the trees through the forest: Sequencebased homo- and heteromeric protein-protein interaction sites prediction using random forest. *Bioinformatics*, 33(10).

Huwe, P. J., Xu, Q., Shapovalov, M. V., Modi, V., Andrake, M. D., and Dunbrack, R. L. (2016). Biological function derived from predicted structures in CASP11. *Proteins: Structure, Function, and Bioinformatics*, 84(S1):370–391.

Kinch, L. N., Li, W., Monastyrskyy, B., Kryshtafovych, A., and Grishin, N. V. (2016a). Assessment of CASP11 contact-assisted predictions. *Proteins: Structure, Function, and Bioinformatics*, 84(S1):164–180.

Kinch, L. N., Li, W., Monastyrskyy, B., Kryshtafovych, A., and Grishin, N. V. (2016b). Evaluation of free modeling targets in CASP11 and ROLL. *Proteins: Structure, Function, and Bioinformatics*, 84(S1):51–66.

Lensink, M. F., Velankar, S., Kryshtafovych, A., Huang, S.-Y., Schneidman-Duhovny, D., Sali, A., Segura, J., Fernandez-Fuentes, N., Viswanath, S., Elber, R., Grudinin, S., Popov, P., Neveu, E., Lee, H., Baek, M., Park, S., Heo, L., Rie Lee, G., Seok, C., Qin, S., Zhou, H.-X., Ritchie, D. W., Maigret, B., Devignes, M.-D., Ghoorah, A., Torchala, M., Chaleil, R. A., Bates, P. A., Ben-Zeev, E., Eisenstein, M., Negi, S. S., Weng, Z., Vreven, T., Pierce, B. G., Borrman, T. M., Yu, J., Ochsenbein, F., Guerois, R., Vangone, A., Rodrigues, J. P., van Zundert, G., Nellen, M., Xue, L., Karaca, E., Melquiond, A. S., Visscher, K., Kastritis, P. L., Bonvin, A. M., Xu, X., Qiu, L., Yan, C., Li, J., Ma, Z., Cheng, J., Zou, X., Shen, Y., Peterson, L. X., Kim, H.-R., Roy, A., Han, X., Esquivel-Rodriguez, J., Kihara, D., Yu, X., Bruce, N. J., Fuller, J. C., Wade, R. C., Anishchenko, I., Kundrotas, P. J., Vakser, I. A., Imai, K., Yamada, K., Oda, T., Nakamura, T., Tomii, K., Pallara, C., Romero-Durana, M., Jiménez-García, B., Moal, I. H., Férnandez-Recio, J., Joung, J. Y., Kim, J. Y., Joo, K., Lee, J., Kozakov, D., Vajda, S., Mottarella, S., Hall, D. R., Beglov, D., Mamonov, A., Xia, B., Bohnuud, T., Del Carpio, C. A., Ichiishi, E., Marze, N., Kuroda, D., Roy Burman, S. S., Gray, J. J., Chermak, E., Cavallo, L., Oliva, R., Tovchigrechko, A., and Wodak, S. J. (2016). Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment. *Proteins: Structure, Function, and Bioinformatics*, 84(S1):323–348.

Marti-Renom, M. A., Capriotti, E., Shindyalov, I. N., and Bourne, P. E. (2009). Structure Comparison and Alignment. In Gu, J. and Bourne, P. E., editors, *Structural Bioinformatics, 2nd Edition*, pages 397–418. John Wiley & Sons, Inc.

May, A., Pool, R., van Dijk, E., Bijlard, J., Abeln, S., Heringa, J., and Feenstra, K. A. (2014). Coarse-grained versus atomistic simulations: realistic interaction free energies for real proteins. *Bioinformatics (Oxford, England)*, 30(3):326–334.

Mészáros, B., Tompa, P., Simon, I., and Dosztányi, Z. (2007). Molecular principles of the interactions of disordered proteins. *J Mol Biol*, 372(2):549–561.

Modi, V., Xu, Q., Adhikari, S., and Dunbrack, R. L. (2016). Assessment of template-based modeling of protein structure in CASP11. *Proteins: Structure, Function, and Bioinformatics*, 84(S1):200–220.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2016). Critical assessment of methods of protein structure prediction: Progress

and new directions in round XI. *Proteins: Structure, Function and Bioinformatics*, 84(S1):4–14.

Moult, J., Pedersen, J. T., Judson, R., and Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Genetics*, 23(3):ii–iv.

Oldfield, C. J., Cheng, Y., Cortese, M. S., Brown, C. J., Uversky, V. N., and Dunker, A. K. (2005). Comparing and combining predictors of mostly disordered proteins. *Biochemistry*, 44(6):1989–2000.

Perdigão, N., Heinrich, J., Stolte, C., Sabir, K. S., Buckley, M. J., Tabor, B., Signal, B., Gloss, B. S., Hammang, C. J., Rost, B., Schafferhans, A., and O'Donoghue, S. I. (2015). Unexpected features of the dark proteome. *Proceedings of the National Academy of Sciences*, 112(52):15898–15903.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). UCSF Chimera – A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612.

Pucci, F., Kwasigroch, J. M., and Rooman, M. (2017). SCooP: an accurate and fast predictor of protein stability curves as a function of temperature. *Bioinformatics*.

Read, R. J. and Chavali, G. (2007). Assessment of CASP7 predictions in the high accuracy template-based modeling category. *Proteins: Structure, Function, and Bioinformatics*, 69(S8):27–37.

Richardson, J. S. and Richardson, D. C. (2002). Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc Natl Acad Sci U S A*, 99(5):2754–2759.

Uversky, V. N., Gillespie, J. R., and Fink, A. L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins*, 41(3):415–427.

van Dijk, E., Hoogeveen, A., and Abeln, S. (2015). The Hydrophobic Temperature Dependence of Amino Acids Directly Calculated from Protein Structures. *PLOS Computational Biology*, 11(5):e1004277.

van Dijk, E., Varilly, P., Knowles, T. P. J., Frenkel, D., and Abeln, S. (2016). Consistent Treatment of Hydrophobicity in Protein Lattice Models Accounts for Cold Denaturation. *Physical Review Letters*, 116(7):078101.

Venselaar, H., te Beek, T. A., Kuipers, R. K., Hekkelman, M. L., and Vriend, G. (2010). Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinformatics*, 11(1):548.

Wang, S., Ma, J., and Xu, J. (2016). AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. *Bioinformatics (Oxford, England)*, 32(17):i672–i679.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191.

Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures. *Nucleic acids research*, 31(13):3370–4.

