VU Research Portal

Strategies for protein structure model generation

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

published in
Introduction to Structural Bioinformatics
2017

Link to publication in VU Research Portal

citation for published version (APA)
Abeln, S., Heringa, J., & Feenstra, K. A. (2017). Strategies for protein structure model generation. In *Introduction to Structural Bioinformatics*

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:
vuresearchportal.ub@vu.nl

Download date: 03. Mar. 2023

arXiv:1712.00425v1 [q-bio.BM] 1 Dec 2017

Structural Bioinformatics

Sanne Abeln **K. Anton Feenstra**

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 deals with approaches for protein three-dimensional structure prediction, starting out from a single input sequence with unknown structure, the ‘query’ or ‘target’ sequence. Both template based and template free modelling techniques are treated, and how resulting structural models may be selected and refined. We give a concrete flowchart for how to decide which modelling strategy is best suited in particular circumstances, and which steps need to be taken in each strategy. Notably, the ability to locate a suitable structural template by homology or fold recognition is crucial; without this models will be of low quality at best. With a template available, the quality of the query-template alignment crucially determines the model quality. We also discuss how other, courser, experimental data may be incorporated in the modelling process to alleviate the problem of missing template structures. Finally, we discuss measures to predict the quality of models generated.

Contents

7 Strategies for protein structure model generation 5
  7.1 Template based protein structure modelling 7
    7.1.1 Homology based Template Finding 7
    7.1.2 Fold recognition 8
    7.1.3 Generating the target-template alignment 10
    7.1.4 Generating a model 11
    7.1.5 Loop or missing substructure modelling 12
  7.2 Template-free protein structure modelling 12
    7.2.1 What if no suitable template exists? 12
    7.2.2 Generating models from structural fragments 12
    7.2.3 Fragment Assembly into decoys 14
    7.2.4 Constraints from co-evolution based contact prediction or experiments 14
  7.3 Selecting and refining models from structure prediction 15
    7.3.1 Model refinement 15
    7.3.2 Model quality assessment strategies 16
    7.3.3 Secondary Structure Prediction 17

Chapter 7

Strategies for protein structure model generation

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

Here we consider strategies for a typical protein structure prediction problem: we want to generate a structural model for a protein with a sequence, but without an experimentally determined structure. In the previous chapter on “Introduction to protein structure prediction” we introduced the problem of how to obtain the folded structure of the protein, given only an amino acid sequence. 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. We will in detail discuss first template-based and then template-free modelling. An overview of protein structure modelling, including both template-based and template-free modelling is given in Figure 7.2; see also Figure 7.1 for the the terminology used.

7.1 Template based protein structure modelling

7.1.1 Homology based Template Finding

Homology modelling is a type of template based modelling with a template that is homologous to the target protein. As mentioned before, homology modelling works well, because structure is more conserved than sequence. Typically, we will use a sequence-based homology detection method, such as BLAST, to search for homologous protein sequences in the full PDB dataset. If we find a sequence that has significant sequence similarity to the full length of our target sequence we have found a template. Of course it is possible that a template only covers part of the target sequence, see also the section on domains in Chapter 6 “Introduction to protein structure prediction”.

If a simple BLAST search against the PDB gives no good results we may need to start using alignment methods that can detect more distant homologues based on sequence comparison, such a PSI-BLAST or HMMs. PSI-BLAST uses hits from a previous iteration of BLAST to create a profile for our query sequence (Altschul et al., 1997). This allows successive iterations to give more weight to very conserved residues, allowing to find more distant homologues. Even more sensitive homology detection tools typically consider sequence profiles of both the query and the template sequence, and try to align / score these profiles against each other. Examples: Profile-Profile Alignment [[***TODO ref?***]] , Compass (Sadreyev and Grishin, 2003), HMM’s such as HHpred (Söding et al., 2005). Note that, to make full use of such methods it is important that the searching profiles are generated using extended sequence databases (so not just on the PDB). The evolutionary sequence profiles used in these methods can ensure that the most conserved, and therefore structurally most important, residues are more likely to be aligned accurately.

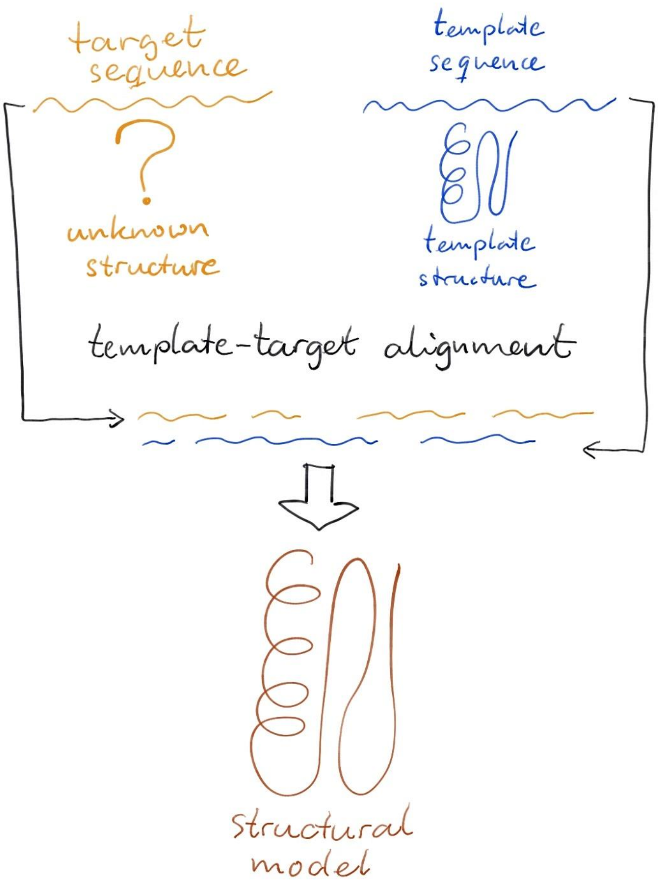

Figure 7.1: Terminology used in protein structure prediction. We start from our protein of interest (with no known structure): the target sequence. First step is find a matching protein: a template sequence with known structure; the template structure. We then create a template-target sequence alignment, and from this alignment create the structural model which is the solution structure for our target protein.

Once we have a good template, and an alignment of the target sequence with the template structure, we can build a structural model; this is called homology modelling in cases of clear homology between the target and the template.

7.1.2 Fold recognition

If no obvious homologs with a known structure can be found in the PDB, it becomes substantially more difficult to predict the structure for a sequence. We may then use another trick to find a template: remember that the template does not only have a sequence, but also a structure. Therefore, we may be able to use structural information in this search. Fold recognition

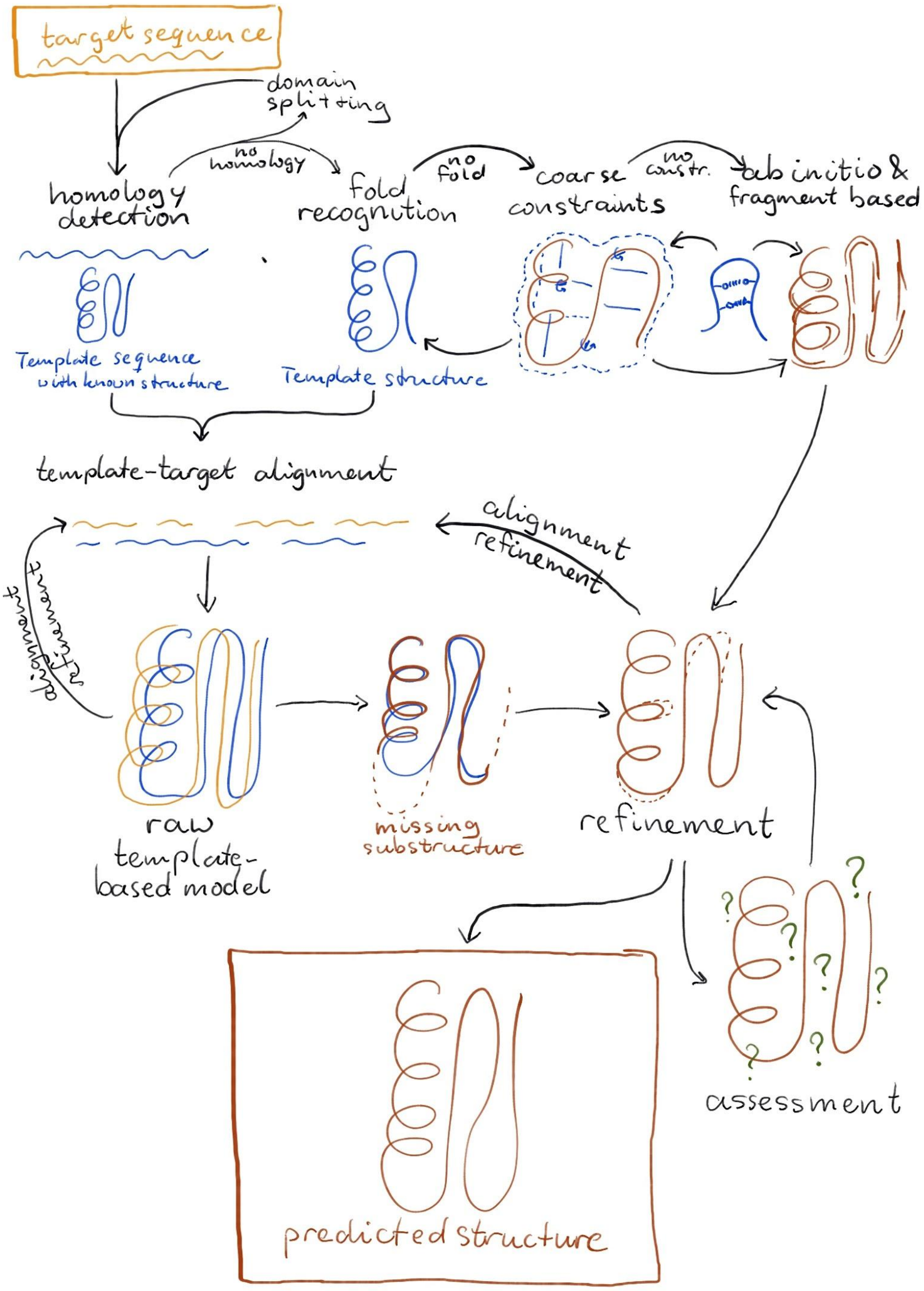
target sequence
domain splitting
homology detection
no homology
fold recognition
no fold
coarse constraints
no constr.
ab initio & fragment based
Template sequence with known structure
Template structure
template-target alignment
alignment refinement
alignment refinement
raw template-based model
missing substructure
refinement
assessment
predicted structure

Figure 7.2: Flowchart of protein three-dimensional structure prediction. It starts at the top left with the target protein sequence of interest, and ends with a predicted 3D structure at the bottom. Depending on the availability of a homologous template, a suitable fold, or coarse/experimental constraints, different options are available, with sharply decreasing expected model accuracy for each step. Homology-based models (far left) are most accurate, while ab-initio modelling (far right) is notoriously unreliable. Template-based (by homology or fold recognition) models require an alignment with the target sequence, from which the initial model will be built. Course constraints may sometimes be incorporated in this stage. The raw model may need to be completed by separate modelling of the missing substructures. Template-free modelling can benefit greatly if such constraints are available, if not the only sources left are fragment libraries and knowledge-based energy functions. The model, whether template-based or -free, is usually refined until a desired level of (estimated) quality is reached to produce the final predicted structure for our target protein sequence of interest.

methods look explicitly for a plausible match between our query sequence and a *structure* from a database. Typically, such methods use structural information of the putative template to determine a match between the target sequence and the putative template sequence and structure.

Note that, in contrast to homology-based template search, it is not strictly necessary for a target sequence and a template sequence to be homologous – they may have obtained similar structures through convergent evolution. Therefore some fold recognition methods are trained and optimised for detection of structural similarity, rather than homology. In fact, for very remote homologs we may not be able to differentiate between convergent and divergent evolution. For the purpose of structure prediction this may not be an important distinction, nevertheless divergent evolution, i.e., homology through a shared common ancestor, invokes the principle of structure being more conserved, giving more confidence in the final model.

Threading is an example of a fold recognition method (Jones et al., 1992). The query sequence is threaded through the proposed template structure. This means structural information of the template can be taken into account when score if the (threaded) alignment between the target and the potential template is a good fit. Typically threading works by scoring the pairs in the structure that make a contact, given the sequence composition. A sequence should also be aligned (threaded) onto the structure giving the best score. Scores are typically based on knowledge based potentials. For example, if two hydrophobic residues are in contact this would give a better score than a contact between a hydrophobic and polar amino acid. Note that threading does not necessarily search for homology. Threading remains a popular fold recognition methods, with several implementations available (Jones et al., 1992; Zhang, 2008; Song et al., 2013).

Another conceptually different fold recognition approach is to consider that amino-acid conservation rates may strongly differ between different structural environments. For example, one would expect residues on the surface to be less conserved, compared to those buried in the core. Similarly, residues in a α-helix or β-strand or typically more conserved than those in loop regions. In fact, the chance to form an insertion or deletion in a loop regions is seven times more likely than within (other) secondary structure elements. Since we know the structural environment of the residues for the potential template, we can use this to score an alignment between the target sequence and potential template sequence. FUGUE is a method that scores alignments using structural environment-specific substitution matrices and structure-dependent gap penalties (Shi et al., 2001).

7.1.3 Generating the target-template alignment

Once a suitable template has been found, one can start building a structural model. Typical model building methods will need the following inputs:

(1) the target sequence, (2) the template structure, (3) sequence alignment between target and template and (4) any additional known constraints. The output will be a structural model, based on the constraints defined by the template structure and the sequence alignment.

Here, it is important to note that the methods that recognize good potential templates, are not necessarily the methods that will produce the most accurate alignments between the target and template sequence. As the final model will heavily depend upon the alignment used, it is important to consider different methods, including for example structure and profile based methods, multiple sequence alignment programs or methods that can include structural information of the template in constructing the alignment. Examples of good sequence-based alignment methods, which can also exploit evolutionary signals and profiles, are *Praline* (Simossis et al., 2005) and *T-coffee* (Notredame et al., 2000). The T-coffee suite also includes the structure-aware alignment method *3d-coffee* (O'Sullivan et al., 2004). Some aligners are context-aware, taking into account for example secondary structure (Simossis and Heringa, 2005) or trans-membrane (TM) regions explicitly (Pirovano et al., 2008, 2010; Floden et al., 2016), which are useful extension as TM proteins are severely underrepresented in the PDB. Of particular interest to the general user are automated template detection tools such as HHpred (Söding et al., 2005). Bawono et al. (2017) give a good and up-to-date overview of multiple sequence alignment methodology, including profile-based and hidden-Markov-based methods.

Moreover, once a first model is created, it may be wise to interactively adapt the alignment, based on the resulting model, which might lead to an updated model. This procedure may also be carried out in an interative fashion.

7.1.4 Generating a model

Here we consider the *MODELLER* software to generate template based alignments (Sali and Blundell, 1993), which is one of several alternative approaches to construct homology models (Schwede et al., 2003; Zhang, 2008; Song et al., 2013). Firstly, the known template 3D structures should be aligned with the target sequence. Secondly, spatial features, such as Cα-Cα distances, hydrogen bonds, and mainchain and sidechain dihedral angles, are transferred from the template(s) to the target. *MODELLER* uses "knowledge based" constraints. The constraints are based on the template distances, the alignment, but also on knowledge- based energy functions (probability distribution). The constraints are optimised using molecular dynamics with simulated annealing. Finally, a 3D model can be generated by satisfying all the restraints as well as possible.

7.1.5 Loop or missing substructure modelling

We now have a model for all the residues that were aligned well between the template and the target. The remaining substructure(s), that are not covered by the template, will show as gaps in the alignment between target sequence and template. 'Loop modelling' is used to determine the structure for these missing parts. Loop models are typically based on fragment libraries, knowledge based potentials and constraints from the aligned structure. This problem is in fact closely related to the template-free modelling procedure, as we need to generate a structure without a readily available template (template-free approaches are discussed in more detail in the next section). In CASP11, consistent refinement overall as well as for loop regions was achieved; the limiting factor for effective refinement was concluded to be the energie functions used, in particular missing physicochemical effects and balance of energy terms (Lee et al., 2016).

7.2 Template-free protein structure modelling

7.2.1 What if no suitable template exists?

If on the other hand, no suitable template is available for our target protein of interest, we will need to follow a 'template-free' modelling strategy. Without a direct suitable template, we need an "ab initio" strategy that can suggest possible structural models based on the sequence of the template alone. In this case, we need to resort to fragment based approaches. Here small, suitable fragments, from various PDB structure, are assembled to generate possible structural models. As the fragments are typically matched using sequence similarity, one may even consider this as template based modelling at a smaller scale. However, since the sequence match is based on a limited number of residues, this would not generally imply a homologous relation between the fragment template and the target sequence. It is also important to note that that this type of "ab initio" modelling is still "knowledge based": the structural models are generated from small substructures present in the PDB, assessed by energy scoring functions generated by mining the PDB. In other words, these model are not based on physical principles and physico-chemical properties alone. This also means, that such models are likely to share any of the biases that are present in the PDB, such as lack of trans-membrane proteins and absence of disordered regions.

7.2.2 Generating models from structural fragments

Here we follow the keys ideas in the *Rosetta* based ab initio modelling suite (Simons et al., 1999). Quark, another fragment-based approach provides a similar performing alternative(Xu and Zhang, 2012). The overall approach

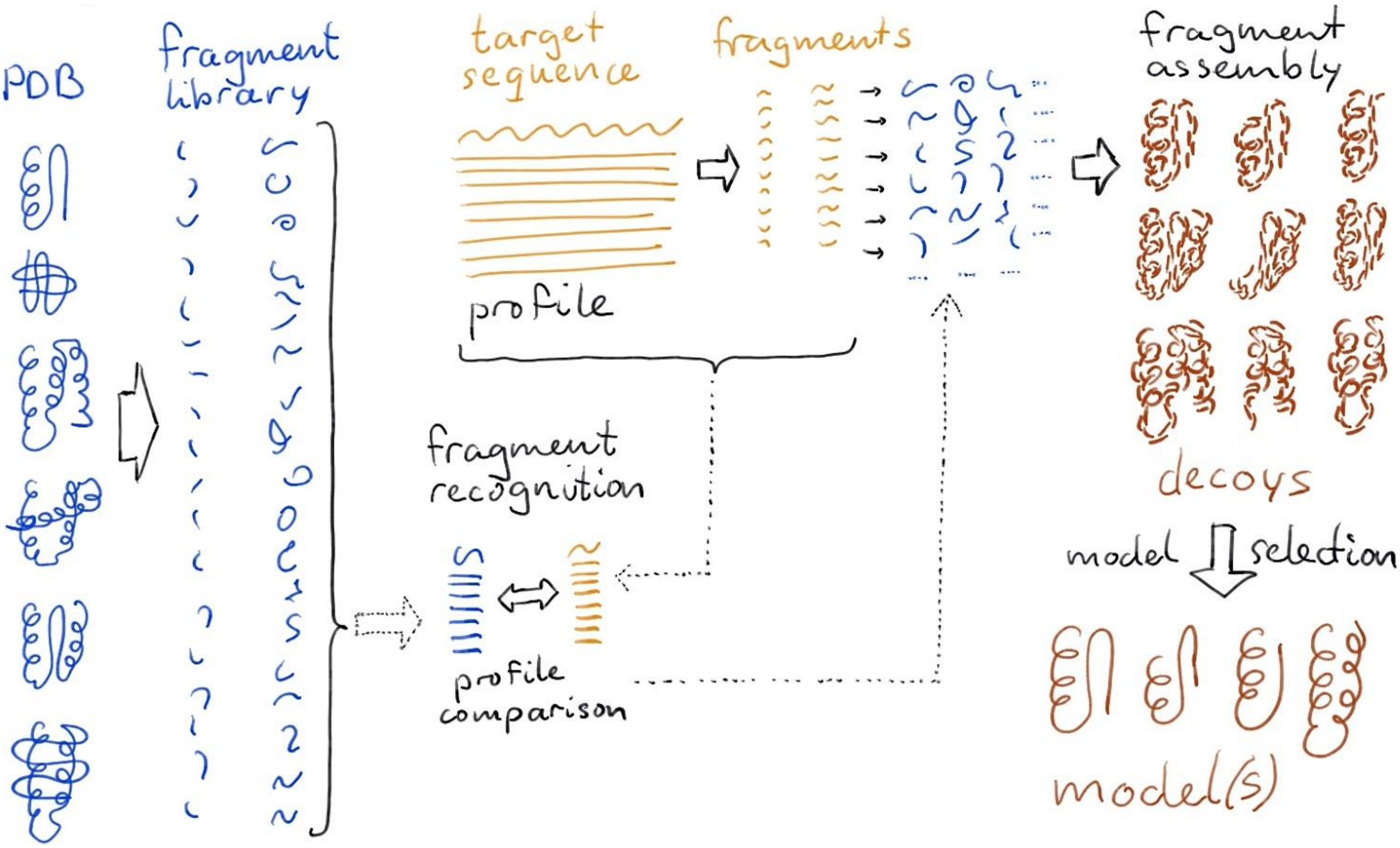

Figure 7.3: Overview of the fragment-based modelling strategy. A library of structure fragments was created once from the PDB; all small 3-residues and larger 9-residue fragments are collected and clustered. A target sequence of interest is also separated into 3- and 9-residue sequence fragments. For each of these, a profile-profile search is performed to find matching fragments from the fragment library; typically for each target fragment, multiple hits with different structure are retrieved. This collection of fragments of alternate structure are then assembled through a Monte Carlo algorithm into a large set of possible structures, called 'decoys'. Using knowledge-based potentials and overall statistics, from the decoy set, a final selection of model structures is made.

is to split the target sequence into 3 and 9-residue overlapping sequence fragments, i.e., sliding windows, and find matching structural fragments from PDB.

A fragment library is generated by taking 3 and 9-residue fragments from the PDB and clustering these together into groups of similar structure. For each fragment in the database sequence profiles are created. These profiles subsequently are used to search for suitable fragments for our query/target sequence, as we only have sequence information (see Figure 7.3 on the left-hand side).

The target sequence also will be split into 3 and 9-residue overlapping sequence fragments. These target sequence fragments are then matched with the structural PDB based fragment library using profile-profile matching. Note that this procedure will generate multiple fragments for each fragment window in the sequence. The fragment windows on the sequence typically also overlap (see Figure 7.3 middle panel).

7.2.3 Fragment Assembly into decoys

The above scenario leads to many possible structural fragments to cover a sequence position. To find the best structural model is a combinatorial problem in terms of fragment combinations, see Figure 7.3 top right. A Monte Carlo algorithm is used to search through different fragment combinations. Good combinations are those that give a low energy. Each MC run will produce a different model, since it is a stochastic algorithm. Note that to be able to optimise a model, we need a scoring function. A knowledge based energy function is used, including the number of neighbours, given amino acid type, residue pair interactions, backbone hydrogen bonding, strand arrangement, helix packing, radius of gyration, Van der Waals repulsion. These are all terms that are relatively cheap to compute when a new combination of fragments is tried. The structural model is optimised by slowly lowering the MC temperature – this is also called simulated annealing. The generated models are called ‘decoys’.

Thereafter, decoys are refined using additional Monte Carlo cycles, and a more fine-grained energy function including: backbone torsion angles, Lennard-Jones interactions, main chain and side-chain hydrogen bonding, solvation energy, rotamers and a comparison to unfolded state.

Finally, the most difficult task is to select, from all the refined decoys, a structure that is a suitable model for the target sequence, see Figure 7.3 bottom right. Again, using a more detailed, knowledge-based energy function, decoys can be scored to assess how ‘protein-like’ they are. Such a selection procedure may get rid of very wrong models. However, selecting the best model, without any additional information (from for example experiments or co-evolution-based contact-prediction), is likely to lead to poor results (see Chapter 6 ’Introduction to Structure Prediction’).

7.2.4 Constraints from co-evolution based contact prediction or experiments

As already mentioned, valuable additions to the modelling process are coarse constraints from experimental data or contact prediction. Experimental data from NMR and chemical cross-linking can yield distance restraints that are particularly useful in the template-free modelling to narrow down the conformational space to be searched; still average accuracy of models produced remains extremely limited (Kinch et al., 2016). Other sources of information are contours or surfaces that can be obtained from cryo-EM, or small angle scattering experiments, either with electrons, neutrons, or x-ray radiation. However, since these techniques are employed mostly for elucidating larger macromolecular complexes, they are considered out of scope for the current chapter.

Of more general applicability may be methods for predicting intra-protein residue contacts; the main approach currently is based on some form of co-evolution information obtained by direct-coupling methods from 'deep' alignments (Marks et al., 2011; Jones et al., 2012; Morcos et al., 2011). Depth here signifies the amount of sequence variation present in the alignment in relation to the length of the protein (the longer the protein, the more variation is needed). Ovchinnikov et al. (2017) expresses this as the *effective protein length*: $Nf = N80\%ID/\sqrt{l}$ where l is the protein length, and $N80\%ID$ the number of cluster at 80% sequence identity. They showed that Nf can be greatly enhanced by the use of metagenomic sequencing data, and that this leads to a marked improvement in model quality, and estimate that this would triple the number of protein families for which the correct fold might be predicted (Ovchinnikov et al., 2017). Wuyun et al. (2016) investigated 'consensus'-based methods, which combine both direct-coupling and machine-learning approaches, and find that the machine-learning methods are less sensitive to alignment depth and target difficulty, which are crucial factors for success for the direct-coupling methods.

7.3 Selecting and refining models from structure prediction

Once we have created (several) models, we need to assess which model is the best one. Typically this can be done by scoring models on several properties using model quality assessment programs and visual inspection with respect to "protein like" features. Moreover, if any additional knowledge about the structure or function of our target protein is available, this may also help to assess the quality of the model(s). In addition, one may in some cases want to improve a model, or parts of it; this is called model refinement.

7.3.1 Model refinement

For many years in CASP, model refinement was a no-go area; the rule of thumb was: build our homology model and do not touch it! An impressive example of the failure of refinement methods was shown by the David Shaw group, who concluded that "simulations initiated from homology models drift away from the native structure" (Raval et al., 2012). Since CASP10 in 2014 (Nugent et al., 2014) and continuing in the latest CASP11, there is reason for moderate optimism. General refinement strategies report small but significant improvements of 3-5% over 70% of models (Modi and Dunbrack, 2016). Interestingly, and in stark contrast to the earlier results by (Raval et al., 2012), the average improvement of GDT_HA using simulation-based refinement now also is about 3.8, with an improvement (more than 0.5) for 26 models (Feig and Mirjalili, 2016). For five models, the scores became worse (by more than 0.5), and another five showed no significant change. Particularly, for very good initial models (GDT_HA$>$ 65), models were made worse. Moreover, they also convincingly showed that both more and longer simulations consistently improved these results; note however that protocol details such as using Cα restraints, are thought to be the limiting factor (Feig and Mirjalili, 2016), as already used previously (e.g., Keizers et al., 2005; Feenstra et al., 2006), and replicated by others (e.g., Cheng et al., 2017). Most successful refinement appears to come from correctly placing β-sheet or coil regions at the termini (Modi and Dunbrack, 2016).

7.3.2 Model quality assessment strategies

It may be generally helpful to compare models generated by different prediction methods; if models from different methods look alike (more precisely if the pair has a low RMSD and/or high GDT_TS) they are more likely to be correct. Similarly, templates found by different template finding strategies, found for example both by homology sequence searches and fold recognition methods are more likely to yield good modelling results (Moult et al., 2016; Kryshtafovych et al., 2016). Such a consensus template is generally more reliable than the predictions from individual methods – especially if the individual scores are barely significant. Lastly, one can consider biological context to select good models.

Whether a model is built using homology modelling, fold recognition and modelling or *ab initio* prediction, all models can be given to a Model Quality Assessment Program (MQAP) for model validation. A validation program provides a score predicting how reliable the model is. These scores typically take into account to what extent a model resembles a “true” protein structure. The best performing validation programs take a large set of predicted models, and indicate which out of these is expected to be the most reliable.

Validation scoring may be based on similar ideas as validation for experimental structures or may be specific to structure prediction. For example, it can be checked if the amount of secondary structure, e.g. helix and strand vs. loop, has a similar ratio as in known protein structures; if a model for a sequence of 200 amino acids does not contain a single helix or β-strand, the model does not resemble true protein structures, and is therefore very unlikely to be the true structural solution for the sequence. A similar type of check may be done for the amount of buried hydrophobic groups and globularity of the protein.

Different models may also be compared to each other. One trick that is commonly used, is that if multiple prediction methods create structurally similar models, these models are more likely to be correct. Hence, a good prediction strategy is to use several prediction methods, and pick out the most consistent solution. A pitfall here is that if all models are based on the same, or very similar, templates, they will look similar but this may not indicate the likelihood of them being correct.

7.3.3 Secondary Structure Prediction

Secondary structure prediction is relatively accurate (see e.g. the review by Pirovano and Heringa, 2010). This problem is in fact much easier to solve than three-dimensional structure prediction, as is shown in Chapter [[***TODO ref to other chapter in book – presumably***]] . The accuracy of assigning strand, helix or loops to a certain residue can go up to 80% with the most reliable methods. Typically such methods use (hydrophobic) periodicity in the sequence combined with phi and psi angle preferences of certain amino acid types to come to accurate predictions. The real challenge lies in assembling the secondary structure element in a correct topology. Nevertheless, secondary structure prediction may be used to assess the quality of a model built with a (tertiary) structure prediction method. Many (automated) methods also incorporate secondary structure information during alignment (Simossis and Heringa, 2005), homology detection (Söding et al., 2005; Shi et al., 2001) and contact prediction (Terashi and Kihara, 2017; Wang et al., 2017).

Bibliography

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–402.

Bawono, P., Dijkstra, M., Pirovano, W., Feenstra, A., Abeln, S., and Heringa, J. (2017). Multiple Sequence Alignment. In *Methods in Molecular Biology – Bioinformatics – Volume I: Data, Sequence Analysis, and Evolution*, pages 167–189. Humana Press, New York, NY.

Cheng, Q., Joung, I., and Lee, J. (2017). A Simple and Efficient Protein Structure Refinement Method. *Journal of Chemical Theory and Computation*, 13(10):5146–5162.

Feenstra, K. A., Hofstetter, K., Bosch, R., Schmid, A., Commandeur, J. N. M., and Vermeulen, N. P. E. (2006). Enantioselective substrate binding in a monooxygenase protein model by molecular dynamics and docking. *Biophysical journal*, 91(9):3206–16.

Feig, M. and Mirjalili, V. (2016). Protein structure refinement via molecular-dynamics simulations: What works and what does not? *Proteins: Structure, Function, and Bioinformatics*, 84(S1):282–292.

Floden, E. W., Tommaso, P. D., Chatzou, M., Magis, C., Notredame, C., and Chang, J.-M. (2016). PSI/TM-Coffee: a web server for fast and accurate multiple sequence alignments of regular and transmembrane proteins using homology extension on reduced databases. *Nucleic acids research*, 44(W1):W339–43.

Jones, D. T., Buchan, D. W. A., Cozzetto, D., and Pontil, M. (2012). PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2):184–190.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, 358(6381):86–89.

Keizers, P. H. J., de Graaf, C., de Kanter, F. J. J., Oostenbrink, C., Feenstra, K. A., Commandeur, J. N. M., and Vermeulen, N. P. E. (2005). Metabolic Regio- and Stereoselectivity of Cytochrome P450 2D6 towards 3,4-Methylenedioxy-N-alkylamphetamines: in Silico Predictions and Experimental Validation. *Journal of Medicinal Chemistry*, 48(19):6117–6127.

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

Kryshtafovych, A., Barbato, A., Monastyrskyy, B., Fidelis, K., Schwede, T., and Tramontano, A. (2016). Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. *Proteins: Structure, Function, and Bioinformatics*, 84(S1):349–369.

Lee, G. R., Heo, L., and Seok, C. (2016). Effective protein model structure refinement by loop modeling and overall relaxation. *Proteins: Structure, Function, and Bioinformatics*, 84(S1):293–301.

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. a., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. *PloS one*, 6(12):e28766.

Modi, V. and Dunbrack, R. L. (2016). Assessment of refinement of template-based models in CASP11. *Proteins: Structure, Function, and Bioinformatics*, 84(S1):260–281.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America*, 108(49):E1293–301.

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.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1):205–217.

Nugent, T., Cozzetto, D., and Jones, D. T. (2014). Evaluation of predictions in the CASP10 model refinement category. *Proteins: Structure, Function, and Bioinformatics*, 82(S2):98–111.

O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G., and Notredame, C. (2004). 3DCoffee: Combining Protein Sequences and Structures within Multiple Sequence Alignments. *Journal of Molecular Biology*, 340(2):385–395.

Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G. A., Kim, D. E., Kamisetty, H., Kyrpides, N. C., and Baker, D. (2017). Protein structure determination using metagenome sequence data. *Science*, 355(6322):294–298.

Pirovano, W., Abeln, S., Feenstra, K. A., and Heringa, J. (2010). Multiple alignment of transmembrane protein sequences. In *Structural Bioinformatics of Membrane Proteins*, pages 103–122. Springer Vienna, Vienna.

Pirovano, W., Feenstra, K. A., and Heringa, J. (2008). PRALINETM: a strategy for improved multiple alignment of transmembrane proteins. *Bioinformatics*, 24(4):492–497.

Pirovano, W. and Heringa, J. (2010). Protein Secondary Structure Prediction. In Carugo, O. and Eisenhaber, F., editors, *Data Mining Techniques for the Life Sciences*, pages 327–348. Humana Press.

Raval, A., Piana, S., Eastwood, M. P., Dror, R. O., and Shaw, D. E. (2012). Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics*, 80(8):2071–2079.

Sadreyev, R. and Grishin, N. (2003). COMPASS: A Tool for Comparison of Multiple Protein Alignments with Assessment of Statistical Significance. *Journal of Molecular Biology*, 326(1):317–336.

Sali, A. and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3):779–815.

Schwede, T., Kopp, J., Guex, N., and Peitsch, M. C. (2003). SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Research*, 31(13):3381–3385.

Shi, J., Blundell, T. L., and Mizuguchi, K. (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol*, 310(1):243–257.

Simons, K. T., Bonneau, R., Ruczinski, I., and Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins*, Suppl 3:171–176.

Simossis, V. A. and Heringa, J. (2005). PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Research*, 33(Web Server):W289–W294.

Simossis, V. A., Kleinjung, J., and Heringa, J. (2005). Homology-extended sequence alignment. *Nucleic Acids Research*, 33(3):816–824.

Söding, J., Biegert, A., and Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research*, 33(Web Server issue):W244–8.

Song, Y., Dimaio, F., Wang, R. Y. R., Kim, D., Miles, C., Brunette, T., Thompson, J., and Baker, D. (2013). High-resolution comparative modeling with RosettaCM. *Structure*, 21(10):1735–1742.

Terashi, G. and Kihara, D. (2017). Protein structure model refinement in CASP12 using short and long molecular dynamics simulations in implicit solvent. *Proteins: Structure, Function and Bioinformatics.*

Wang, S., Sun, S., and Xu, J. (2017). Analysis of deep learning methods for blind protein contact prediction in CASP12.

Wuyun, Q., Zheng, W., Peng, Z., and Yang, J. (2016). A large-scale comparative assessment of methods for residueresidue contact prediction. *Briefings in Bioinformatics*, page bbw106.

Xu, D. and Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function and Bioinformatics*, 80(7):1715–1735.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, 9(1):40.