The use of information in existing protein structure prediction methods

RBO, UMass

December 27, 2007

Contents

Abstract

This study will examine the sources of biological information and their use in protein structure prediction. The goal is to catalog previously explored and other possible information sources, explain why their use is helpful for protein structure prediction in the context of a specific prediction method, and hypothesize how the information could be used within the protein structure prediction framework under development in the Robotics and Biology Laboratory. This report contains two main parts: 1) An in-depth analysis of existing approaches, including a catalog of possible information sources, and 2) A detailed discussion of future opportunities for leveraging information in protein structure prediction.

Chapter 1

Background and terminology

In this chapter, we provide an introduction to the problem of protein structure prediction from an engineering perspective, and we establish terminology used throughout this report.

A protein is a chain of amino acids connected by peptide bonds. Hence, it is called a **polypeptide chain**, or a peptide. The term protein is sometimes used to indicate agglomerates of polypeptide chains, but in this report, we will use *protein* and *polypeptide chain* interchangably. Different permutations of 20 amino acids form different proteins. The number of amino acids also vary among different proteins.

The **primary structure**, or the **sequence** of a polypeptide chain is defined by an ordered list of amino acids. For example, ARN represents a polypeptide chain formed by the amino acids Alanine, Arginine and Asparagine. Every such chain has a direction or order. Hence, the chains ARN and NRA represent two different proteins.

The chemical bonds which constitute a protein are flexible to varying degrees. When provided sufficient force, it is possible to rotate a molecule about a bond, or to alter the bond length. This means that a polypeptide chain is capable of assuming many different 3-dimensional **conformations**. The set of all possible conformations of a protein constitutes the **conformation space** of that protein.

In its natural environment, the probability of a polypeptide chain assuming a certain conformation depends on a physical quantity called **free energy** associated with that conformation - The lower the free energy associated with a conformation, the more likely the protein is to have that conformation in its natural environment. The mapping from conformations to the associated free energy values is referred to as the **energy landscape**. Every point in the energy landscape corresponds to a unique conformation.

Protein structure prediction (PSP) refers to the problem of finding the conformations the protein is likely to adapt in its natural environment,

given its sequence. Hence, protein structure prediction involves finding the conformations for which the associated free energy is sufficiently close to the global minimum. Thus, protein structure prediction is an *optimization* or *search* problem. It involves exploration of the energy landscape in the search for the global minimum.

Exploration of the energy landscape involves movement from one point in the energy landscape to another. The energy landscape is riddled with many local minima. The region of the energy landscape surrounding a local minimum is called a **well**, or a **funnel**. If movement in the energy landscape corresponds to moving from a point in a well to a point closer to the minimum of the same well, then such movement is called **gradient descent**. As the energy landscape has so many local minima, no protein strucutre prediction algorithm can rely entirely on gradient descent. Any PSP algorithm with any chance of finding the global minimum should be capable of moving from one well to another.

Many proteins in nature fold spontaneously, whereas a few fold only due to the action of other **chaperone proteins**. To explain the fact that these proteins fold spontaneously and quickly, it has been suggested that the rough shape of the energy landscape is that of a funnel, with many shallow wells.

Chapter 2

A catalog of possible information sources

In this chapter, we list and describe different sources of information useful in protein structure prediction. Because one information from one source often influences the content and confidence in information from other sources, we do not attempt to develop a taxonomy of information sources on that basis.

Some sources of information can be used even before the exploration of the energy landscape begins. Below we list them:

Information from X-Ray Crystallography and NMR spectroscopy experiments

Description : X-Ray Crystallography experiments yield the diffraction patterns for crystallized proteins. From this primitive data, three dimensional electron density maps may be obtained. From this, the structure of the crystallized molecule may be deduced, with different resolution. [1]

Use in PSP : Low resolution models, or even primitive data obtained from X-Ray crystallography experiments serves to constrain the size of the volume of the energy landscape to be searched. There are disadvantages to relying on data from X-Ray crystallography in PSP: X-Ray Crystallography experiments are time and resource intensive. Also, a protein's conformation in its natural environment in the cell may differ from the conformation of the protein in a crystal.

Homology with other proteins : Homologous proteins are proteins which share a common evolutionary ancestry.

Some proteins

- The degree of similarity between the query sequence and the sequences of other proteins:

[1] [followThisUp]

- The degree of similarity between the structure of the query protein and the structures of other proteins:
- Multiple sequence alignments: Multiple sequence alignments align three or more protein sequences.
- Information about the secondary structure of the target protein: Secondary structure of the target protein may be predicted by considering sequences and structure of other proteins.
- Information about the overall tertiary structure of the target protein: Homologous proteins share evolutionary ancestry. Protein structure is more conserved than the protein sequence, and homology among proteins is often inferred on the basis of sequence similarity. Hence, structures of homologous proteins can give us clues about the structure of the query protein.
- Domains in the homologous proteins.

  Regions of structural variability.

  Possible structures of short substrings of the query sequence.
- Homologous proteins:

  Co-conserved pairs of residues, which are likely to interact with each other.

  Co-conserved residues, which are likely to obligate a fold

  Consensus among alignments about alignment of residues.
- Folding routes hypothesized for a protein
- Predicted secondary structures:

2.1 Information gathered during the prediction process

- Information about the local region in conformation space

  Net forces experienced by all atoms in the molecule.

  Forces experienced by coarse collections of atoms

  Energy functions Below is a list of various information sources which are subsumed by energy functions:

  Packing density

  Stereochemistry

  Closeness to homologous structures

  Inferences about the surrounding region in conformation space

The velocity with which decoy energy changes in a given region.

Geometric constraints on the motion of a protein.

Closeness to proteins in the predicted SCOP or CATH family.

- Inferences about the non-local regions in conformation space

  Coarse structure of convex regions This is used by structure prediction methods which involve smoothing.

  The degree of familiarity and the promise of regions

  The probable fold family of the protein

Chapter 3

The use of information in existing protein structure prediction methods

3.1 Protein modelling using experimental data

Data from X-Ray crystallography and NMR experiments can be used to inform proteins structure prediction methods.

3.2 Threading or fold recognition

3.3 Ab Initio or de novo methods

3.4 Molecular Dynamics

3.5 Fragment replacement methods

3.6 Homology or Comparative Modelling

Homology modelling is the process of predicting a protein's structure by referring to the structures of homologous proteins. The protein whose structure is being modelled is called the *query sequence* or *target*. The reference proteins are called *templates*.

Structures of regions of the query sequence for which no template structure can be found, are modelled using a process called *loop modelling*. Such regions often correspond to the exposed loops of a protein. [1]

[1] A loop or turn is defined by the close approach of two C_α atoms which are not involved

[10] is an excellent reference for information on Comparative Modelling. Much of the information in this section is drawn from that article.

3.6.1 When can homology modelling be used?

Homology modelling is not used in cases where no template can be found to align with a significant portion of the target sequence. [2] Suppose that a protein chain 300 amino-acids long is being modelled. Suppose that one may find atemplate which aligns with a 64 amino acid segment. In this alignment, the proportion of amino acids which are present at identical positions in the target and the template segments is known as 'sequence identity'.

It is not applicable for membrane proteins.

As structures of proteins are more conserved than their sequences, detectable levels of sequence similarity usually implies structural similarity. [10], which was published in the year 2000, says: 'It has been estimated that approximately one third of all sequences are recognizably related to at least one known protein structure. Because the number of known protein sequences is approximately 500,000 (9), comparative modeling could in principle be applied to more than 150,000 proteins. This number can be compared to approximately 10,000 protein structures determined by experiment.', 'One half of these models are in the least accurate class, based on less than 30% sequence identity to known protein structures. The remaining 35 and 15% of the models are in the medium ($< 50\%$ sequence identity) and high ($> 50\%$ identity) accuracy classes.' and 'Depending on a genome, the probability of finding a related protein of known structure for a sequence randomly picked from a genome ranges from 20% to 50%'. The same paper also says: 'Assuming the current growth rate in the number of known protein structures, the structure of at least one member of most of the globular folds will be determined in less than 10 years.'

[10] says that homology modelling is currently not applicable to membrane proteins.

3.6.2 Steps involved in Homology Modelling:

Template selection and Fold Assignment

Fold Assignment seems to refer to the identification of the family of structures to which the protein belongs, whereas *Template selection* seems to mean the identification of the specific template or templates which will be used in homology modelling.

Template selection methods:

Using pairwise sequence comparisons : *FASTA* and *BLAST* may be used to accomplish this.

in an α helix or β sheet.

Using multiple sequence comparisons : This technique identifies a greater number of homologs than techniques which only use pairwise alignments. [10] says: 'The multiple sequence methods for fold identification are especially useful for finding significant structural relationships when the sequence identify between the target and the template drops below 25%.' [2]

Using multiple sequence comparisons with structure predictions : [7] used structural information predicted from the target sequence along with multiple sequence alignments. [3]

By cosntruction and comparison of profiles : "Profile" refers to the model of sequence changes expected within parts of sequences. They are a sort of 'average sequence'. [13] says: "HMMer and SAM employ Hidden Markov Models (HMMs) which incorporate position-specific gap penalties." "COMPASS aligns profiles against profiles, whereas HHSearch and PRC align HMMs." That study indicated that profile-profile methods, which compare profiles against a database of other profiles, outperform profile-sequence methods, which compare profiles against a database of other sequences in finding remote homologs. *PSI-BLAST* uses profiles, which may be constructed using other tools.

Using 'Protein threading' : Protein threading, also known as fold recognition or 3D-1D alignment, can be used as a search technique for identifying templates. The target sequence is sequentially compared with template proteins, an alignment is established, and the match is calculated using a scoring function which utilizes structural information. [10] says: 'These methods are especially useful when there are no sequences clearly related to the modeling target, and thus the search cannot benefit from the increased sensitivity of the sequence profile methods.' [4]

The choice of templates :

Often, several candidate template structures are identified by these approaches. Some methods can generate models from multiple templates, while others rely on a single template.

The choice of templates to be used is guided by several factors, such as phylogeny, the similarity of the query and template sequences, their functions, environments, the quality of the experimental structures and the plausibility of the resulting model. In case of automated homology modelling, one has to rely on alignment metrics. Blast E-values and FASTA sequence identities were used by a previous study, and they were both found to be 'fair'.[6]

Current Limitations :

[2] [followThisUp]
[3] [followThisUp]
[4] [followThisUp]

- Also see the limitations listed under the subsection *'Sequence Alignment'*.
- **Multiple templates**: In a study of some automated homology modelling programs, while operating at low sequence identity, the use of multiple templates did not generate a model which was reliably better when compared to one generated using only the best template.[6]
- **Inaccurate structure:** Inaccuracies in the template structure can also reduce the quality of the prediction. Also, if the template structure is not determined in a functionally meaningful environment, the quality of the model based on that template is liable to be inaccurate. [10]

Sequence alignment

Even though the process of template selection may yield alignments, [10] suggests the use of specialized methods for achieving optimum alignment, as the search methods are usually tuned for identifying remote relationships.

Importance of correct alignment : Incorrect alignment can easily misguide comparative modelling. So, finding the best alignment is important. [10] also says: 'Other factors such as template selection and alignment accuracy usually have a larger impact on the model accuracy, especially for models based on less than 40% sequence identity to the templates.' It has been shown that, given the structural alignment between a target and a template sequence, highly accurate models of the target protein sequence can be produced; a major stumbling block in homology-based structure prediction is the production of structurally accurate alignments given only sequence information. [16] [5] The choice of sequence alignment strategy, more than the choice of modelling program, is critically important for accurate modelling. [15, 6] [6] [6] found 3D-Coffee to be the best among the sequence alignment programs it tested.

Sequence alignment methods:

Using sequence information only: Examples: BLAST, FASTA and PSI-BLAST. Multiple alignments are generally more reliable than pairwise alignments. [10]

Using both sequence and structure information: One could use threading (3D-Coffee), build and align locale profiles (Staccato) or penalize indels in secondary structure locales (SAlign). These may involve multiple, rather than only two sequences. Note that the method used by SAlign seems to involve secondary structure prediction.

[5] [checkCitation]
[6] [followThisUp]

Current limitations: [10] says: 'For closely related protein sequences with identity over 40%, the alignment is almost always correct.' and 'The alignment becomes difficult in the twilight zone of less than 30% sequence identity.' Sequence alignment quality decreases with decreasing sequence identity. [6] The chief inaccuracies in homology modeling, which worsen with lower sequence identity, derive from errors in the initial sequence alignment and from improper template selection. [15] [7] [8]

Model Generation: Conserved regions

[6], which tested various modelling programs did not find much variation in the accuracy of the models predicted. The important distinctions between the various modellers seem to lie in their flexibility to handle new data, and ease of automation.

Model Generation methods:

Rigid body assembly: The structurally conserved regions of the target are constructed from the corresponding regions in the template. Nest, Builder and Swiss-Model use this. [6]

Segment Matching or Coordinate Reconstruction: Using atoms which are conserved in the target and the template as guiding positions, one may identify and assemble short segments which fit these guiding positions. [10] says: 'The basis of modeling by coordinate reconstruction is the finding that most hexapeptide segments of protein structure can be clustered into approximately 100 structural classes.' Suitable segments are found either by searching a database, or through conformation space search. SegMod/ENCAD uses this. [6]

Satisfaction of spatial restraints: In another method, one or more target-template alignments are used to construct a set of geometrical criteria. Additional constraints may be added, for example: from experimental data. This serves as the basis of a global optimization procedure to arrive at the final prediction. [14, 9, 8] A popular software in spatial restraint-based modeling is MODELLER and a database called ModBase has been established for reliable models generated with it. [12]

Model Generation: Loop Modelling

Loops contribute to active and binding sites. So, accurate loop modelling is important for the use of the models generated in the study of protein-ligand interactions. [10]

[7] [checkCitation]
[8] [followThisUp]

Loop modelling methods:

Ab initio loop modelling: This involves conformation space search. [6] says: 'Modeller builds loops by optimizing a series of probability density functions describing backbone geometry based on amino-acid type, and then refined with an energetic minimization procedure. Nest uses the Loopy algorithm, which generates a set of conformations for each loop, with the best selected by a colony energy term that favors conformations low in energy but also close in structure to other conformations.'

Loop modelling using databases: 'Builder and SegMod/ENCAD fit fragments that best match local geometry and sequence, selected from a local database, and subsequently undergo energy minimization.' [6] This is similar to modelling by 'Segment Matching or Coordinate Reconstruction'.

Loop modelling using a combination of the above: 'Swiss-Model initially explores conformational space with constraint space programming and uses a force field-based scoring scheme to determine the best loop conformation. However, when this process fails or when loops are longer than 10 residues, a pre-determined loop library is utilized.' [6]

Comparison: [10] says: 'The database search approach to loop modeling is accurate and efficient when a specific set of loops is created to address the modeling of that class of loops, such as β-hairpins and the hypervariable regions in immunoglobulins.' A study of homology modelling in the low sequence identity region found far less variation than there was between the sequence alignment methods. Even though the margins were small, Modeller and Nest, both of which use ab-initio loop modelling, were found to outperform the rest. [6] [9]

Current limitations: Parts of the structure predicted by 'Loop modelling' are less accurate than the parts modelled by comparison to a template, especially when the loop is long.

A study of homology modelling in the low sequence identity region found that all the loop building methods tested suffer as loop length increases, and on average loops longer than six residues are modelled inaccurately. [6]

Model Generation: Sidechains

Method : Sidechains are placed as close as possible to the template sidechain. 'Rotamer' or torsion-angle libraries are used to pick a likely conformation. They are then optimized for maximum packing and minimum free energy. Formation of disulphide bridges can be considered a special case. [10]

[9] [followThisUp]

Current limitations: [10] says: 'Side-chain errors are critical if they occur in regions that are involved in protein function, such as active sites and ligand-binding sites.' The first two sidechain dihedral angles can usually be estimated within 30A for an accurate backbone structure; however, the later dihedral angles found in longer side chains such as lysine and arginine are notoriously difficult to predict. [5] [10]

'The rotameric states of side chains and their internal packing arrangement also present difficulties in homology modeling, even in targets for which the backbone structure is relatively easy to predict. This is partly due to the fact that many side chains in crystal structures are not in their "optimal" rotameric state as a result of energetic factors in the hydrophobic core and in the packing of the individual molecules in a protein crystal. One method of addressing this problem requires searching a rotameric library to identify locally low-energy combinations of packing states.' [5] [11]

Model Evaluation

Imagine that you were doing comparative modelling. Suppose that you are unsure about the choice of template or alignment. In such a case, you could build a model from all those choices, and compare them in the end to decide on the final prediction. Models may be evaluated using energy functions and stereochemistry evaluators. [10]

3.6.3 Advantages of comparative modelling

Homology is the most accurate computational structure prediction method. [11] As protein structures are more conserved than protein sequences, detectable levels of sequence similarity usually imply significant structural similarity. A consortium has been formed, which attempts to find experimental structures for all classes of folds. [1]

3.6.4 Availability of various programs:

The authors of [6], who benchmarked three sequence-structure alignment programs: 3D-Coffee, Staccato and SAlign and five homology modelling programs: Builder, Nest, Modeller, SegMod/ENCAD and Swiss-Model, installed and used all those programs locally, except Swiss-Model. [10] contains a list of programs and servers useful in comparative modelling.

[10] [checkCitation]
[11] [checkCitation]

Chapter 4

Future opportunities for leveraging information

4.1 Homology Modelling

Sequence alignment, loop modelling and side-chain packing are the weak points of current homology modelling techniques.

4.1.1 Potential for use in the Robotics and Biology Laboratory

Novel ways of conformation space search are being explored at the Robotics and Biology lab. Model Based Search attempts to balance exploration (the process of finding information) and exploitation (the process of using information) in order to search in the most relevant regions of the conformation space. [4] Homology is one of the sources of information which could be used in this process.

Producing better alignments : It was suggested that the structure predicted by Model Based Search be used in template identification. However, that approach might not bring information which was not already present in the de-novo-predicted structure. Instead, a sequence homologue's structure is likely to bring fresh information for further search. So, one might choose the template simply based on its Blast score, and then use the structure predicted de-novo to find the optimum alignment with the template. This is a promising venue for research, because, given the structural alignment, highly accurate models of the target protein sequence can be produced. [16]

Better modelling : Energy functions, used in several de-novo structure prediction algorithms, have been found to be inaccurate for some proteins. For some proteins, some of the current conformation-space search algorithms have

been shown to miss the regions of the energy landscape with the lowest energy conformation. [1] Using homology information could solve these problems.

- If the query sequence has close homologs (with greater than 50% match) with known structures, then current homology modelling methods are probably adequate.
- If only templates with 30 - 50% match can be found, one could use existing homology modellers to predict the structure of the aligned portion, and, holding it fixed, one could use the techniques developed in the lab to search the portion of the conformation space spanned by the loop regions.
- If only templates with 0-30% match can be found, one could use the distance between any given conformation and the template in conjunction with the energy function values for the surrounding conformations in order to guide search. This is already being attempted at the lab. Proper alignment is critical in this zone.

[6] found that Ab initio loop modelling techniques outperform other techniques used by homology modellers by a slight margin. Ab initio loop modelling can be viewed as an instance of conformation space search. So, model based search can be very useful here.

The 30 - 50% appears to be a very attractive place to start and test our hypothesis, which claims that model based search can do loop modelling better. This zone is a safer target than the twilight zone, because the strength of Model Based Search of the conformation space is proven, whereas the superiority, over existing methods, of using a de-novo prediction for acheiving optimum template selection and alignment is not. Also, the superiority of using weak alignments in the twilight zone with Model Based Search over using Model Based Search alone is not proven.

Another hypothesis, which claims that the structure predicted by model based search can be used for better template identification and alignment, can be tested then.

(The robetta paper must be checked out next.) [2]

4.1.2 Implementation notes

A study of homology modelling in the low sequence identity region found that Blast E-values are adequate to identify templates for use in homology modelling. The same study found that the program 3D-Coffee the best at sequence alignment, when used with the homology modeller, Modeller.[6] So, Blast E-values are adequate for template identification. But, for sequence alignment, using 3D-Coffee is likely to yeild good results.

Currently, the lab's attempts at using homology information involve the use of multiple templates. A study of homology modelling in the low sequence

[1] [checkCitation]
[2] [followThisUp]

identity region found that using multiple templates did not generate a model which was reliably better when compared to one generated using only the best template.[6] Another article has suggested that, for effective modelling using multiple templates, they must belong to the same structural family. [3] Hence, one must conduct experiments to find which approach works best with the protein structure prediction framework at the lab.

Comparative modelling using the satisfaction of spatial constraints provides a nice model for the way in which Model Based Search can use homology information while predicting the structure of sequences with 30 - 50% identity with templates. The scoring function used by Rosetta can perhaps be modified to heavily penalize violation of specified restraints.

4.2 Stereochemistry

[10] mentions several programs which may be used to evaluate the stereochemistry of models. During Model Based Search, decoys [3] in various regions of the conformation space are evaluated. Stereochemistry could aid in this evaluation. This could be especially true in cases where the energy function is inaccurate.

4.3 Energy functions

Current efforts in RBO exclusively use Rosetta's energy function. The energy function used by Rosetta is known to be faulty in some cases. The lab may want to experiment and determine the utility of other energy functions.

[3] A decoy is a 3D model of a protein, which does not constitute the final prediction.

Bibliography

[1] Williamson AR. Creating a structural genomics consortium. *Nature Structural Biology*, page 953, 2000.

[2] David Baker and Andrej Sali. Protein Structure Prediction and Structural Genomics. *Science*, 294(5540):93–96, 2001.

[3] P.A. Bates, L.A. Kelley, R.M. MacCallum, and M.J. Sternberg. Enhancement of protein modeling by human intervention in applying the automatic programs 3d-jigsaw and 3d-pssm. *Proteins*, Suppl 5, 2001.

[4] T J Brunette and Oliver Brock. Improving protein structure prediction with model-based search. *Bioinformatics*, 21(1):66–74, 2005.

[5] Contributors. Homology modelling. *Wikipedia*, 2007.

[6] James A. Dalton and Richard M. Jackson. An evaluation of automated homology modelling methods at low target template sequence similarity. *Bioinformatics*, 23(15):1901–1908, August 2007.

[7] D. Fischer and D. Eisenberg. Assigning folds to the proteins encoded by the genome of mycoplasma genitalium. *Proc Natl Acad Sci U S A*, 94(22):11929–11934, October 1997.

[8] Michael Y. Galperin and Eugene V. Koonin. *Frontiers in Computational Genomics*. Horizon Scientific Press, 2003.

[9] B. John and A. Sali. Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res*, 31(14):3982–92, 2003.

[10] M. A. Martí-Renom, A. C. Stuart, A. Fiser, R. Sánchez, F. Melo, and A. Sali. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, 29:291–325, 2000.

[11] John Moult. A decade of casp: progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology*, 15(3):285–289, 2005.

[12] U. Pieper, N. Eswar, H. Braberg, M.S. Madhusudhan, F.P. Davis, A.C. Stuart, N. Mirkovic, A. Rossi, M.A. Marti-Renom, A. Fiser, B. Webb, D. Greenblatt, C.C. Huang, T.E. Ferrin, and A. Sali. Modbase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res*, 32(1):D217–22, 2004.

[13] Adam James J. Reid, Corin Yeats, and Christine Anne A. Orengo. Methods of remote homology detection can be combined to increase coverage by 10 *Bioinformatics*, August 2007.

[14] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, 234(3):779–815, December 1993.

[15] C. Venclovas and M. Margelevicius. Comparative modeling in casp6 using consensus approach to template selection, sequence-structure alignment, and structure assessment. *Proteins*, 61 Suppl 7:99–105, 2005.

[16] Y. Zhang and J. Skolnick. The protein structure prediction problem could be solved using the current pdb library. *Proc Natl Acad Sci U S A*, 102(4):1029–1034, January 2005.