

An Overview of Protein Structure Prediction: From Homology to Ab Initio

Final Project For Bioc218, Computational Molecular Biology

Zhiyong Zhang

Abstract

The current status of the protein prediction methods, comparative modeling, threading or fold recognition, and Ab Initio prediction, is described. The accuracy, applicability and shortcomings, as well as possible improvements will be discussed.

Introduction:

The biological role of a protein is determined by its function, which is in turn largely determined by its structure. Thus there is enormous benefit in knowing the three dimensional structures of all the proteins. Although more and more structures are determined experimentally at an accelerated rate, it is simply not possible to determine all the protein structures from experiments. As more and more protein sequences are determined, there is pressing need for predicting protein structures computationally. Decades of intense research in this area brought about huge progress in our ability to predict protein structures from sequences only. The protein structure prediction methods can be broadly divided into three categories: 1) homology modeling, 2) threading or fold recognition, and 3) Ab Initio. Essentially, the classification reflects the degree to which different methods utilize the information content available from the known structure database. In the following, I will briefly discuss each kind of methods and their accuracy, applicability and shortcomings. Possible improvements to protein structure prediction are also discussed.

Comparative homology modeling:

So far protein prediction methods based on homology have been the most successful. Homology modeling is based on the notion that new proteins evolve gradually from existing ones by amino acid substitution, addition, and/or deletion and that the 3D structures and functions are often strongly conserved during this process. Many proteins thus share similar functions and structures and there are usually strong sequence similarities among the structurally similar proteins. Strong sequence similarity often indicates strong structure similarity, although the opposite is not necessarily true. Homology modeling tries to identify structures similar to the target protein through sequence comparison. The quality of homology modeling depends on whether there exists one or more protein structures in the protein structure databases that show significant sequence similarity to the target sequence.

There are usually four steps in homology based protein structure prediction methods: (1) identify one or more suitable structural templates from the known protein structure databases; (2) align the target sequence to the structural template; (3) build the backbone from the alignment, including the loop region and any region that is significantly different from the template; and (4) place the side-chains. The first two steps, identification of structural templates and alignment of the target sequence onto the parent structures, are usually related. Sequence comparison methods determine sequence similarity by aligning the sequences optimally. The aligned residuals of the structure templates are used to construct the structural model in the second step. The quality of the sequence comparison thus not only determines whether a suitable structural template can be found but also the quality of the alignment between the target sequence and the parent structure, which in turn determines the accuracy of the structural model. Of critical

importance is the ability for the sequence comparison to detect remote homologues and to correctly align the target sequence to and parent structure. In the following I discuss the various sequence comparison methods in relation to homology modeling and their range of applicability, accuracy and shortcomings.

For comparative modeling, local sequence comparison methods are usually used since the sequence similarity is most likely over segments of the two sequences. The local sequence comparison can either be pair wise or profile based. Pair wise comparisons, such as the widely used BLAST (Altschul, 1990) in the early days, can detect sequence similarities better than 30%. A number of tools have also been developed to detect weak homology relationships. Methods like profile (Gribskov, 1987) and HMM (Krogh, 1996) use a statistical profile of a protein family. To further increase the chance of detecting remote homologues, PSI-BLAST (Altschul, 1997) and SAM-T98 (Karplus, 1998) build the profile or HMM by searching the database iteratively until no new hits are found. Methods such as PSI-BLAST encode the information about a whole protein family for the target sequence in a model to increase the chance of detecting remote homologies. To further increase the detection sensitivity, the sequences in the structure database can also be encoded in profiles. This forms the basis of the profile-profile based comparison methods (Koehl, 2002). With low sequence identities (<20%), profile-profile methods clearly outperform the other two kinds of methods (Sauder, 2000): profile-profile methods identified more than 90% of homologous pairs, determined from structure-structure similarity comparison, with sequence identity better than 10% and an impressive 38% even for cases with sequence identities between 5% and 9%.

The structure models are constructed from the residuals of the structure template that are aligned to the target sequence in the sequence comparison. The quality of this alignment thus is critical for the accuracy achievable. The aligned residues from sequence comparison are generally different from that from structure-structure comparison though, especially when the sequence identity is low. To assess the ability of the sequence comparison methods to align the sequences correctly, it is instructive to compare the sequence-sequence alignment to the structure-structure alignment of the same pair of proteins. To determine how well the different similarity search methods can detect remote homologies and assess their ability in correctly aligning the sequences, Sauder et al. (Sauder, 2000) compared various sequence alignment methods to the CE (Shindyalov, 1998) structure alignment of the SCOP (Murzin, 1995) protein structures. For sequence identities less than 30%, profile-based comparison methods, such as PSI-BLAST and profile-profile comparison, are all obviously better than the pair wise BLAST method. For example, at 10-15% sequence identity, BLAST aligns only 20% correctly while PSI-BLAST and profile-profile comparison can correctly align 40% and 48% respectively. This also indicates that there is still large room for improvement in correctly aligning the target sequence to the target structure.

One indication of the accuracy of comparative modeling is the sequence identity between the target and the template. It is believed that if two protein sequences have 50% or higher sequence identity, then the RMSD of the alignable portion between the two structures will normally be less than 1_ (Gerstein, 1998). In the so-called "twilight zone"

(Doolittle, 1986), with sequence identity between 20%~30%, 95% of the sequences with this level of identity have different structures though (Rost, 1999). When a structure template can indeed be found within the known protein structure databases in such cases, the backbone RMSD can be expected to be no better than 2_ (Chung, 1996). Structurally similar proteins can have low sequence identities in the 8~10% range (the midnight zone, Rost, 1997) and can still be identified with sensitive profile-profile based comparison, but the RMSD can be as large as 3~6_. The error largely comes from the misalignment from sequence comparison. At such low sequence identity, comparison method that can detect the remote homology as well as align the sequences close to the optimal from structure-structure alignment will be desirable.

Threading or fold recognition:

For evolutionally remotely related proteins, even if the sequence similarity is difficult to detect with sequence comparison methods, there could still be identifiable structural similarity. Structure alignments has been shown to be able to identify homologous protein pairs with sequence similarities less than 10%. (Gerstein, 1998; Brenner, 1998; Rost, 1997). When sequence comparison based methods are no longer sensitive enough to recognize the correct fold for the target sequence, fold recognition or threading can still be used to assign the correct fold to the target sequence.

Threading or fold recognition is the method by which a library of unique or representative structures is searched for structure analogs to the target sequence, and is based on the theory that there may be only a limited number of distinct protein folds. For example, in an early paper, Chothia postulated that the number of unique protein folds would be on the order of only about 1000 unique protein folds (Chothia, 1992). In another estimation, the number of distinct domains and folds were placed around 7000 (Orengo et. al., 1994). Even though the number of new structures solved has been increasing at an accelerated rate (close to 3000 structures solved in 2002), the proportion of new folds, as determined by the CE algorithm (<http://cl.sdsc.edu/ce.html>), to the total number of new structures solved in a given year decreased from an average of ca. 30% in the 80's steadily down to only ca. 8% in year 2001 (<http://www.rcsb.org/pdb/holdings.html>). It is reasonable to expect that as more and more protein structures are determined experimentally, we will be able to find close structure analogues in the databases of known structures for almost any protein sequence in the near future.

Threading or fold recognition involves similar steps as comparative modeling. The difference is in the fold identification step. First of all, a structure library needs to be defined. The library can include whole chains, domains, or even conserved protein cores. Once the library is defined, the target sequence will be fitted to each library entry and an energy function is used to evaluate the fit between the target sequence and the library entries to determine the best possible templates. Depending on the algorithms to align the target sequence with the folds and the energy functions to determine the best fits, the threading methods can roughly be divided into four classes. (Jones, 2001) (1) The earliest threading methods used the environment of each residue in the structure as the

energy function and dynamical programming to evaluate the fit and the alignment (Bowie, 1991). (2) Instead of using overly simplified residual environment as the energy function, statistically derived pair wise interaction potentials (Sippl, 1990) between residue pairs or atom pairs can be used to evaluate the best possible fits between the target sequence and library folds (Jones, 1992). In this method, for efficient optimal alignment between the target sequence and the folds, the potential for residual i is obtained by summing over all the pair wise potentials involving i , and then “double dynamical programming” (Taylor, 1989; Jones, 1998) method can be used. (3) The third kind of methods does not use any explicit energy function at all. Instead, secondary structures and accessibility of each residue are predicted first and the target sequence and library folds are encoded into strings for the purpose of sequence-structure alignment. (4) Finally, sequence similarity and threading can be combined for fold recognition. For large-scale genome wise protein structure prediction, sequence similarity can be first used for the initial alignments and the alignments can be evaluated by threading methods (Jones, 1999).

The threading methods are limited by the high computational cost since each entry in the whole library of thousands of possible folds needs to be aligned in all possible ways to select the fold(s). Another major bottleneck is the energy function used for the evaluation of the alignment. As these functions are drastically simplified for efficient evaluation, it is not reasonable to expect to be able to find the correct folds in all cases with a single form of energy function. Nevertheless, with the current functions, it is possible to reduce the thousands of possible folds to only a few. Similar to the comparative modeling case, for sequence similarities at protein family level, threading can produce alignments that are accurate to 1 to 3 %, or in the case with low sequence similarity at the super-family level, alignment at the range of 3 to 6 % can still be expected. As more protein structures are determined and sequence comparison methods improve, more and more target sequences fold assignment can be achieved by comparative modeling though.

Worth mentioning is the threading program PROSPECT (Xu, 2001), which performed best in its category in the CASP4 competition. What is unique to PROSPECT is that it is designed to find the globally optimal sequence-structure alignment for the given form of energy function (Xu, 2000). The divide-and-conquer algorithm is used to speed up the calculation by explicitly avoiding the conformation search space that is shown not to contain the optimal alignment (Xu, 1998). In several cases that have sequence identity as low as 17%, perfect sequence-structure alignment is still achieved for the alignable portions between the target and template structures. Even in cases that no fold templates exist for the target sequence, important features of the structure are still recognized through threading the target sequence to the structures.

Ab Initio methods:

When no suitable structure templates can be found, Ab Initio methods can be used to predict the protein structure from the sequence information only. Common to all Ab Initio methods are: 1) Suitably defined protein representation and corresponding protein conformation space in that representation; 2) Energy functions compatible with the

protein representation; 3) Efficient and reliable algorithms to search the conformational space to minimize the energy function. The conformations that minimize the energy function are taken to be the structures that the protein is likely to adopt at native conditions. The folding of the protein sequence is ultimately dictated by the physical forces acting on the atoms of the protein and thus the most accurate way of formulating the protein folding or structure prediction problem is in terms of all-atom model subject to the physical forces. Unfortunately the complexity of such a representation makes the solution simply impossible with today's computational capacity. For practical reasons, most Ab Initio prediction methods use reduced representations of the protein to limit the conformational space to manageable size and use empirical energy functions that capture the most important interactions that drive the folding of the protein sequence toward the native structures. Currently, many Ab Initio methods can predict large contiguous segments of the protein to accuracy within 6_ of RMSD and there are several reviews that highlight the success and failure of the current Ab Initio methods. (Hardin, 2002 and references therein). The ROSETTA Ab Initio method performed better than the other Ab Initio methods in the recent CASP4 meeting and there are extensive literature (Bonneau, 2001; Simons, 2001; Bonneau, 2001) covering this method so we concentrate on a brief discussion of method used in ROSETTA. The ROSETTA method also illustrates many features and techniques that are common to the majority of the Ab Initio methods based on reduced representation of the protein and empirical potentials. Discussion of other methods with empirical potentials can be found in Hardin's review. (Hardin, 2002)

The ROSETTA method, like many others, uses a reduced representation of the protein as short segments. This representation can be attributed to the observation by Go (Go, 1983) that local segments of the protein sequence have statistically important preferences for specific local structures and that the tertiary structure has to be consistent with this preference. In ROSETTA the protein is represented by short sequence segments and the local structures they can adopt are assumed to be those found in all the known protein structures. (Simons, 1997) The energy function is defined as the Bayesian probability of structure/sequence matches and this forms the basis of the Monte Carlo sampling of the reduced protein conformational space (Simons, 1997). The non-local potential, which drives the protein toward compact folded structure, includes terms that favor paired strands and buried hydrophobic residuals. The solvation effect can also be incorporated in the energy function.

A problem intrinsic to the reduced representation of the protein and the simplified empirical potential is that the energy function is not sensitive enough to differentiate the correct native structures from conformations that are structurally close to the native state. The energy landscape calculated from such energy functions will not be properly funneled but flattened and caldera-like around the native structure. In fact, as the native state is approached, the correlation between the calculated energy and the measure of similarity between predicted and native structures are no longer valid. The usual practice is then to produce a large number of decoy structures and then use various filtering and clustering techniques to pick up the more native like structures. Filters can be used to eliminate structures with poorly formed secondary structures and low contact orders compared with that for sequences with compatible length (Bonneau 2001). The other

important technique is to use multiple sequences similar to the target sequence to generate decoy structures. Structures thus generated usually form dense clusters that are more compatible to the native structures of protein families of similar sequences than those obtained from a single sequence only.

Many Ab Initio methods now can predict long segments of the protein sequence with backbone atom RMSD less than 6 Å. The predicted local structures are usually right, with the correct contacts among residuals. One of the largest sources of errors was identified to be in the contacts between distant residuals in the sequence as measured by the contact order (CO). (Bonneau, 2002).

Discussion

We have discussed the common methods used for protein structure prediction. The most accurate and successful method so far has been comparative modeling based on sequence similarity comparison, especially when there exists a structure template with high sequence identity to the target. One major progress in comparative modeling is the very sensitive profile based sequence comparison method such as PSI-BLAST and profile-profile sequence comparison. Profile-profile based sequence comparison methods are usually superior in that such methods can pick up possible homologous structure templates even when the sequence identity is very low and that profile-profile comparison can align the sequence to the structure template more accurately, producing more accurate structure models. As more and more novel sequences are produced from the genome projects, the profile-based methods can be expected to become even more sensitive. Fold assignments that are traditionally accomplished from threading methods can be done with comparative modeling instead. On the other hand Ab Initio based methods can still be expected to play an important role in identifying new folds as the accuracy of these methods increase.

There is a wide range of possible applications for protein structure prediction, requiring different accuracy of the predicted structures (Baker, 2001). For applications like studying catalytic mechanisms and ligand docking in drug design, high accuracy structures with RMSD within 1 Å of the native structure is required. Low accuracy structures with RMSD in the range of 1.5~3.5 Å for more than 80% of the sequence can be used for tasks like fitting X-ray structures. Reliable functional annotation and active site prediction can usually be achieved with accuracy of 4-8 Å for over 80 amino acids, which is well within the current capability of Ab Initio methods like Rosetta. When structure templates with sequence identity over 50% can be found, the main chain atoms can be modeled to 1 Å RMSD, with the main error from the loop regions and side chains. With sequence identities between 30%-50%, main chain accuracy of 1.5 Å can be expected. When the sequence identity is below 30%, the error in the aligned main chain atoms can be estimated from the sequence difference. Simple linear relation has been found between the structural difference and sequence difference if the sequence difference is taken to be the average of that between the sets of sequences compatible to the structures (Koehl 2002, Koehl 2002). A more serious problem for comparative modeling in cases with low sequence identity is the false positives. With highly sensitive

profile-profile based methods, even if several structures may be identified to have sufficient sequence similarity, it could happen that none of these structures is the correct template for the target sequence. In fact, it has been shown that 95% of the sequence pairs with 20~30% identity (the twilight zone) are not structurally similar (Rost, 1999). In such cases the possible structure templates can be subject to further threading test for validation. Threading method has been shown to be able to avoid false positives (Panchenko, 1999). Threading with the limited number of possible structure templates avoid one of the computational bottlenecks in threading methods.

Further improvements in the predicted protein structures can be expected from several fronts and I briefly discuss these possibilities. (1) First and foremost, the largest improvements will come from more experimentally determined structures. As more and more protein structures are determined experimentally, it is conceivable that more and more target sequences will have compatible structures already deposited in the known structure database. This will increase not only the chance that the comparative modeling can assign the fold correctly but also the likelihood that the fold identified is more structurally similar to the target, thus increasing the accuracy of the structural model. (2) Further improvement in the sequence-structure alignment can also improve the accuracy of the structure model. The current sequence comparison methods can only align a fraction of the residuals that can be aligned in structure alignments (Sauder, 2000). Better-aligned residuals can undoubtedly improve the accuracy of the structure model. Probably there is a limit to which sequence comparison methods can align sequence to structure when the sequence identity is low. One possible way of improving sequence-structure alignment might be using threading based techniques to align the sequence to structures identified in comparative modeling. With better energy functions for evaluating the fit between sequence and target, this could be very effective. (3) Refinements to the structure models generated from homology modeling, threading, or even Ab Initio methods can be accomplished by molecular dynamics (MD) with accurate all-atom physical potentials. The most severe obstacle of the application in MD in protein structure prediction has been the long time it takes for the protein to fold from the completely unfolded states. This is probably due to the energy barriers encountered in the course of folding. If the simulation starts from the near native structures generated from the protein structure prediction methods, the MD simulation perhaps can reach the native structures much easily.

References:

- Altschul, S. F., Gish W., Miller W., Myers E. W., Lipman D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403.
- Altschul, S. F., Madden T. L., Schaffer A. A., Zhang J., Zhang Z., Miller W., Lipman D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389.
- Baker D., Sali A. (2001). Protein structure prediction and structural genomics. *Science*. **294**, 93.
- Bonneau R. Baker, D. (2001) Ab Initio protein structure prediction: progress and prospects. *Annul. Rev. Biophys. Biomol. Struct.* **30**, 173.
- Bonneau R., Tsai Jerry, Ruczinski I., Chivian D., Rohl C., Strauss C. E. M. and Baker D. (2001) ROSETTA in CASP4: Progress in Ab Initio protein structure prediction. *Proteins: Structure, Function, and Genetics Suppl 5*, 119
- Bonneau R., Ruczinski I., Tsai J., and Baker D. (2002). Contact order and Ab Initio protein structure prediction. *Protein Science*. **11**, 1937.
- Bowie J. U., Luthy R. & Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164
- Brenner S. E., Chothia C. & Hubbard, T. J. P. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA*, **95**, 6073.
- Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature*, **357**, 543.
- Chung, S., Subbiah, S. (1996). A structural explanation for the twilight zone of protein sequence homology. *Structure*, **4**, 1123
- Doolittle, R. F. (1986). Of URFs and ORFs: A primer on how to Analyze Derived Amino Acid Sequences, *Series University Science Books*, Mill Valley, CA
- Gerstein, M., Levitt, M., (1998). Comprehensive assessment of automatic alignment against a manual standard; the SCOP classification of proteins. *Protein Sci.* **7**, 445
- Go N. (1983). Theoretical studies of protein folding. *Ann. Rev. Biophys. Bioeng.* **12**, 183
- Gribskov, M., McLachlan, A. D., Eisenberg, D. Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci USA*. **84**, 4355
- Hardin C. H., Pogorelov T. V. and Luthey-Schulten Z. (2002) Ab Initio protein structure prediction. *Current opinion in structural biology*. **12**, 176
- Jones D., Taylor W. R., and Thornton, J. M. (1992) A new approach to protein fold recognition. *Nature*. **358**, 86.
- Jones D. T. (1998) THREADER: protein sequence threading by double dynamic programming. In *Computational methods in biology* (ed. S. Salzberg, D. Searl, and S. Kasif). Elsevier, Amsterdam.
- Jones D. T. (1999) GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, **287**, 797.
- Jones D. T., Hadley C. (2001) Threading methods for protein structure prediction. In *Bioinformatics: Sequence, Structure and Databanks: A Practical Approach* (ed. Higgins D., and Taylor W.). Oxford University Press, Oxford.
- Karplus K., Barrett C., Hughey R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics*. **14**, 846.

- Koehl P., Levitt M. (2002). Improved recognition of native-like protein structures using a family of designed sequences. *Proc. Natl. Acad. Sci.* **99**, 691.
- Koehl P., Levitt M. (2002). Protein topology and stability define the space of allowed sequences. *Proc. Natl. Acad. Sci.* **99**, 1280.
- Krogh A., Brown M., Mian I. S., Sjolander K., Haussler, D. (1996). Hidden Markov models in computational biology: application to protein modeling. *J. Mol. Biol.* **235**, 1501.
- Murzin, A. G., Brenner, S. E., Hubbard, T., Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536
- Pachenko A., Marchler-Bauer A., Bryant S. H. (1999). Threading with explicit models for evolutionary conservation of structure and sequence. *Proteins: structure, function, and genetics*. **Suppl 3**, 133.
- Rost B. (1997). Protein structures sustain evolutionary drift. *Fold Des.* **2**, 519
- Rost B. (1999) Twilight zone of protein sequence alignment. *Protein Eng.* **12**, 85
- Sauder J. M., Arthur W., and Dunbrack R. L. (2000) Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins: Struct., Func., and Genetics.* **40**, 6.
- Shindyalov I. N., Dourne P. E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Prot. Eng.* **11**, 739
- Simons K. T., Kooperberg C., Huan E., Baker D. (1997) (Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209
- Simons K. T., Strauss C., David B. (2001) Prospects for Ab Initio protein structural genomics. *J. Mol. Biol.* **306**, 1191
- Sippl M. J. (1990). Calculation of conformational ensembles from potentials of mean force: An approach to knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, **213**, 859.
- Orengo, C. A., Jones, D. T., and Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature*, **372**, 631.
- Taylor, W. R. and Orengo, C. A. (1989) Protein structure alignment. *J. Mol. Biol.* **208**, 1
- Xu Y., Xu D. Protein threading using PROSPECT: design and evaluation. *Proteins: struct., Funct., and Genetics.* **40**, 343.
- Xu Y., Xu D., Crawford O. H., and Einstein J. R. (2000). A computational method for NMR-constrained protein threading. *J. Comp. Biol.* **7**, 449.
- Xu D., Crawford O. H., LoCascio P. F., and Xu Y. (2001) Application of PROSPECT in CASP4: characterizing protein structures with new folds. *Proteins: struct., Funct., and Genetics*. **Suppl. 5**, 140.
- Yona G., Levitt M. (2002). Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J. Mol. Biol.* **315**, 1257.