Applying Near-Optimal Alignments to Protein Structure Prediction

by Wendy Ching
wching@stanford.edu
March 10, 2003

Project Aim: *Improvement of protein structure prediction by the application of near-optimal alignments to protein family and structural profile database searches and 3D homology modeling.*

Introduction to Protein Structure Prediction

The prediction of three-dimensional protein structures from one-dimensional amino acid sequence information is an important and interesting problem, as much can be learned about a protein's function from the way that it is folded. There are many different ways to approach this problem, but they all tend to fall into three basic categories: homology modeling, fold recognition, and *ab initio* prediction.

In homology modeling, the structure of the target sequence is inferred by comparison to proteins of known structure (templates). Protein structure can be experimentally determined using methods such as X-ray crystallography and nuclear magnetic resonance (NMR), but this can be difficult as well as costly. Using homology modeling, an optimal alignment between the target and template can be used as the basis for structural superposition of the two proteins.

If there is no sequence homolog of known structure available for the target sequence, fold recognition methods, or threading, can be used to try to detect structural similarities that are not accompanied by significant sequence similarity. The general method is to try to fit the target sequence to a compatible three-dimensional structural profile built from a protein with known structure. Once a template protein has been found in this way, the target can be modeled by comparison.

Finally, *ab initio* prediction refers to the determination of three-dimensional structure by applying energetics principles to one-dimensional sequence information alone.

A Closer Look at Homology Modeling and Fold Recognition

Predicting protein structure by homology depends on the ability to identify a template protein and align the target to the template. One way to better find template sequences is to align the target to a protein family or domain profile built from a multiple sequence alignment rather that doing a simple pairwise alignment to proteins of known structure. In this way, residues which are more or less evolutionarily conserved can be weighted accordingly so that more distant relationships can be found at a significant level. There are a number of publicly available protein domain family databases that can be searched.

Pfam (http://pfam.wustl.edu/index.html) is a database of protein families constructed using manually checked constant seed alignments (usually derived with ClustalW) and hidden Markov models (HMMs) to find and align family members (Sonnhammer, Eddy, Durbin 1997). Then for each full protein family, a profile HMM is built. A target sequence can be searched against this HMM database in order to try to align the sequence to a particular protein family profile using dynamic programming methods.

Another method of building protein families is NCBI's PSI-BLAST, which is an iterated BLAST search. IMPALA is a software package designed by NCBI that can search a target sequence against a database of position-specific scoring matrices (PSSMs) built from PSI-BLAST families (Schaffer et al. 1999). PSSMs are calculated by looking at the logarithm of the ratio of predicted to background residue frequency for each position. IMPALA uses the Smith-Waterman algorithm to find the optimal local alignment for each query and assigns the target sequence to this family.

ProDom is a database of protein domains built by automatic clustering of sequences from SWISS-PROT/TrEMBL (Corpet et al. 2000). It can be searched via BLASTp by either consensus sequence or multiple sequence alignment and gives a list of matching domains. DOMO is similar to ProDom but it was built by clustering of domains using multiple criteria and therefore may be more accurate (Gracy and Argos 1998).

Once a template protein family or domain profile has been identified, a good alignment between target and template must be created so that structural superposition will be accurate. Currently, searches such as Pfam/HMMER and IMPALA give a user back the

optimal alignment path between the query sequence and the highest scoring template family found using dynamic programming methods. ProDom and DOMO also return the best alignment for each of the domain profile matches found. In cases where sequence similarity is greater than 40%, these alignments are nearly always correct for structural comparisons. However, as similarity decreases, an increasing number of gaps and errors appear in the optimal path. To try to make up for this, structural information must be taken into account, often manually, to try to refine the alignment. For example, in regions where structure is predicted to be highly constrained (such as the residues in an alpha helix) gaps in the alignment are to be avoided. Indeed, two major causes of error in homology modeling are misalignments and local distortions and shifts in correctly aligned regions. (Marti-Renom et al. 2000)

Similarly, target sequences can be aligned to a database of 3D structure profiles. The Protein Data Bank (PDB) is an archive of experimentally determined 3D structures. A structural profile converts 3D structural information about a template protein into a 1D string of scores for each possible residue at each position. For example, Bowie, Luthy, and Eisenberg (1991) determine the environmental class of each residue in a folded protein structure that is dependent on the total area of the side chain that is buried, fraction of side chain area that is covered by polar atoms or water, and local secondary structure (eg. alpha helices, beta sheets). In this way, structural information is converted into a string of position-specific environmental classes. At each position, a score can be assigned for the probability of finding each of the twenty amino acids. These are referred to as 3D-1D scores. The target sequence can then be aligned to this string of scores using dynamic programming. The resulting alignment is then used for homology modeling. Just like with target-template alignments to protein family and domain profiles, biologically correct alignments to structure profiles are critical for accurate structure superposition. Yet in this case too, it has been observed that the resulting optimal alignments often require further refinement for increased structural accuracy.

Proposed Application of Near-Optimal Alignments

In scenarios like these, the highest scoring alignment may not necessarily be the most biologically or structurally relevant. I propose that by looking at a set of near-optimal alternative alignments filtered with an additional scoring function based on structural criteria, a more relevant path might be found automatically.

Before going into that, however, a brief review of near-optimal alignments is in order. Near-optimal alignments are alignments whose scores lie within a certain user-specified range from the optimal score. They can be calculated algorithmically as an extension of standard dynamic programming methods. Regions that are shared by all near-optimal alignments are 'uniquely defined' and are the most reliably aligned. Therefore this method provides a means of obtaining local quality scores for an alignment as well as providing alternative high-scoring alignments. (Vingron and Argos 1990)

The application of near-optimal alignments to protein structure prediction is not completely unprecedented, as this method has previously been used to improve the quality of pairwise alignments for homology modeling. Saqi, Bates, and Sternberg (1992) showed that by filtering non-trivial near-optimal alignments between a target and template sequence using measurements of structural criteria (packing potentials and core volumes), alignments were obtained that corresponded more closely to the structurally correct alignment than by looking only at the optimal path. Therefore an extension of the application of near-optimal alignments to homology modeling using protein domain family or structural profiles might prove to be a more efficient way to find structurally correct alignments. This is accomplished by reducing the need for refinement of the target-template alignment after the most relevant alignment has been found. Furthermore, this method defines which regions are more or less reliably aligned. This information can be used to build a more accurate model of target structure.

Method

As mentioned earlier, protein family database searches used by both the Pfam database of profile HMMs and IMPALA implement dynamic programming methods to find and return the highest scoring alignment path. Structure profiles such as 3D-1D environmental strings are also aligned in this way. Using the method of Naor and Brutlag (1994), non-trivial alternative alignments within a certain distance from the optimal path could also be found and enumerated. The Prodom and DOMO domain databases are currently searchable by either consensus or multiple sequence alignment using BLASTp. Perhaps by creating domain profiles for these databases, dynamic programming may also be applied here to search for near-optimal alignments.

Finding a Relevant Alignment: Structural Scoring Functions

Once a set of near-optimal alignments has been generated, the best structural match must be distinguished from the rest. This can be done using a scoring function based on structural criteria. The scoring function is a means to represent how well each alignment matches the template in terms of various energetic considerations.

For example, one measure that can be used as an indicator of structural similarity between target and template is core volume. A protein's core volume is made up of the amino acid side chains that point inwards toward the buried center of the protein. Because packing is complementary, conserved structures generally have a conserved overall volume in the core even if individual residues that point inwards are not identical (Bordo and Argos 1990). Core volume can be determined based on alignment information by calculating the volume of core residues in the template profile and that of corresponding residues in the target sequence. Though these scores will be rough estimates of similarity, the fact that they can be calculated from alignment information alone makes this a good criteria for filtering. Depending on the number of alignments under consideration, things could be very slow and cumbersome if a detailed 3D model had to be built for each alignment. This is a potential weakness of the method, but we can try to avoid it by choosing scoring criteria like these whose calculations do not require detailed models.

Pair potentials, or packing densities, can also be used to score similarity of the target to the template based on structure. Each amino acid is represented by either one, two (histidine, tyrosine, and phenylalanine), or three (tryptophan) spheres in this simplified model of residue packing density. Thus it does not require a detailed 3D model. Proteins with more similar fold structures should have similar packing densities. (Gregoret and Cohen 1990)

Contact potentials represent the energy of interaction between two residues as a function of their 3D distance from each other. Contact potentials can be calculated for all residue pairs when the template structure's 3D coordinates are known (such as in PDB entries). This information can be compared to contact potentials calculated from the target sequence once it has been superpositioned onto the carbon backbone structure of the template. (Hendlich et al. 1990) Building these types of 3D models for each alignment can be

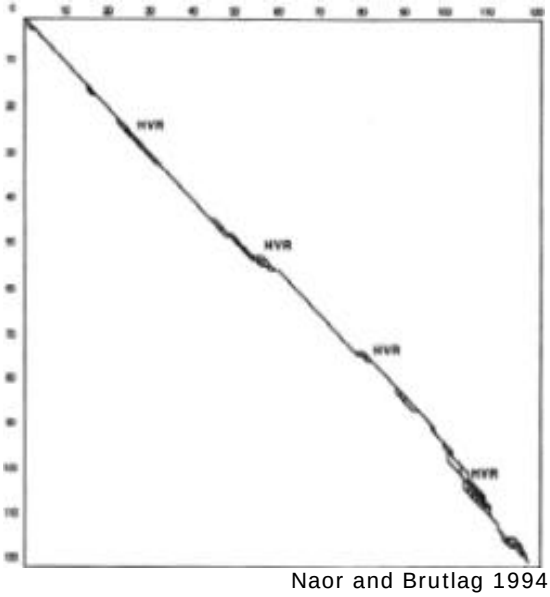
more intensive and this scoring method only works when the template structure's 3D coordinates are available.

Scoring functions that take into account one or more structural considerations such as the ones mentioned here might therefore be used to find the most relevant alignment of the target sequence to a protein domain family or structural profile from a set of near-optimal alignments. The selected alignment can then be used to model the target sequence comparatively.

Using Local Alignment Quality for 3D Homology Modeling

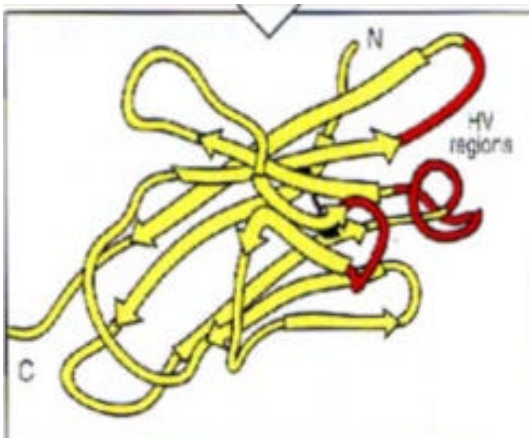
Homology modeling begins with a superposition of a target sequence onto the best template profile, and can then be manually refined. Here, too, information gained by near-optimal alignment methods can be used to improve structure prediction.

As mentioned earlier, regions that are common to all or many near-optimal alignments are the most reliably aligned. Therefore local quality scores can be assigned over an entire alignment based on a comparison of just one target sequence to one template profile. Highly conserved regions among near-optimal alignments usually correspond to highly structured regions of the folded protein. Often these are elements of secondary structure such as alpha helices and beta sheets. Meanwhile, loops and turns connecting regions of secondary structure are least highly conserved. This has been illustrated with a graphical representation of a set of near-optimal alignments between an immunoglobulin heavy and light chain.



Naor and Brutlag 1994

Near-optimal paths diverge at the hypervariable regions (HVR) of human immunoglobulin (Ig). These regions are highly divergent and represent the loops that connect the beta strands that make up the heavy and light chains. In the figure below of an Ig light chain the hypervariable regions are shown in red. These regions make up the antigen-binding sites of the Ig, leading to a diverse range of specific binding-affinities in different members of the Ig family.



(Janeway and Travers 1997)

With this in mind, information on local quality of an alignment can be used when thinking about building a 3D model of the target protein. Reliably aligned regions are highly conserved between the template and the target, and probably represent highly structured regions such as alpha helices or beta sheets. Regions where near-optimal alignments diverge usually represent areas with looser structural constraints. Therefore by looking at changes in local alignment quality, delineation of elements of secondary structure might be improved.

Conclusions

The application of near-optimal alignments to protein structure prediction may prove helpful in the search for more biologically relevant target-template alignments by providing alternative alignment paths that can be filtered according to structural similarity criteria. In addition, by providing information about local alignment quality, near-optimal alignments may aid in the subsequent challenge of modeling the 3D structure of the target sequence using homology. These are just some of the many ways in which the application of near-optimal alignments to protein structure prediction may prove to be beneficial.

REFERENCES:

Bowie JU, Luthy R, Eisenberg D. “A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure.” *Science* 258(5016):164-170, 1991.

Bordo D, Argos P. “Evolution of protein cores: constraints in point mutations as observed in globin tertiary structures.” *J. Mol. Biol.* 211(4):975-988, 1990.

Corpet F, Servant F, Gouzy F, Kahn D. “ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisions.” *Nucleic Acids Research* 28(1):267-269, 2000.

E.L.L. Sonnhammer, S.R. Eddy, and R. Durbin. “Pfam: A Comprehensive Database of Protein Domain Families Based on Seed Alignments.” *Proteins* 28:405-420, 1997.

Gracy J, Argos P. "Automated protein database classification: I. Integration of compositional similarity search, local similarity search and multiple sequence alignment. II. Delineation of domain boundaries from sequence similaritites." *Bioinformatics* 14(2): 164-187, 1998

Gregoret LM, Cohen FE. “Novel method for the rapid evaluation of packing in protein structures.” *J. Mol. Bio.* 211(4):959-974.

Hendlich M, Lackner P, Weitckus S, Floeckner H, Froshauer R, Gottsbacher K,Casari G, Sippl MJ. “Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force.” *J. Mol. Biol.* 216(1):167-80, 1990.

Janeway C, Travers P. Immunobiology 3rd ed. *Current Biology Ltd*. Figure 3-8, 1997.

Naor D, Brutlag DL. “On near-optimal alignments of biological sequences.” *J. Comput. Biol.* 1(4):349-366, 1994.

Saqi MAS, Bates PA, Sternberg MJE. “Towards an automatic method of predicting protein structure by homology: an evaluation of suboptimal sequence alignments.” *Protein Engineering* 5(4):305-311, 1992.

Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF. “IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific scoring matrices.” *Bioinformatics* 15(12):1000-1011, 1999.

Vingron M. “Near-optimal sequence alignment.” *Current Opinion in Structural Biology* 6(3):346-352, 1996.

Vingron M, Argos P. “Determination of reliable regions in protein sequence alignments.” *Protein Engineering* 3(7):565-569, 1990.

http://pfam.wustl.edu/index.html (Pfam database)

http://www.smi.stanford.edu/projects/helix/bmi214 (BMI 214 class notes)