Roopal Sampat
Biochemistry 118Q
Professor Doug Brutlag
3/8/99

Probabilistic Approaches to Predicting the Secondary Structure of Proteins

Today's increasingly unaffordable medical treatment forces genomic research to have far-reaching consequences. Most members of the public do not realize that the genetic sequence does not only encode information about hereditary make-up, but that it also contains the necessary blueprints for the structural formation of essential proteins. As the central dogma of molecular biology declares, DNA is transcribed into RNA, which is then translated into amino acids that make up proteins. Malfunctions in these proteins result in phenotypes that may be classified as diseases. As health care functions today, doctors assess symptoms, resulting in a diagnosis for a disease. Physicians must make educated guesses based upon the symptoms and run a series of tests, the process of which may sometimes prove impractical or extremely expensive. Bioinformatics has emerged as providing a new perspective for the treatment of genetically inherited diseases. The central paradigm of bioinformatics states that genetic information can be used to predict molecular structure of proteins, and the function of these proteins can then be determined, providing a cause for symptoms of a disease. If the structure and function of every protein encoded by DNA were known, the underlying causes of symptoms could be easily pinpointed. Elucidating these structures, however, is a process that could occupy scientists for hundreds of years. As a result, much research has been and

continues to be done regarding the prediction of secondary structures of proteins based upon determined amino acid sequences.

X-ray crystallography has been the traditional method for determining the structure of a protein. Protein samples are crystallized, and a fine beam of x-rays is targeted at them. The x-ray diffraction detected is then used to generate a model of the electron density of the protein. Several disadvantages, however, exist to using x-ray crystallography. First of all, the crystallization of proteins is usually a difficult and time consuming process that requires a great deal of skill. Secondly, x-ray diffraction provides a static model of protein structure, with atoms and molecules mapped in fixed-space. Although this representation is useful, proteins do not usually acquire a fixed structure and instead are continuously bending and shifting, characteristics that may be crucial to the function of the protein. Thirdly, the time needed to crystallize and x-ray, much less identify, every single protein that is encoded by the genetic sequence could span centuries of work. As a result, scientists would prefer to be able to accurately predict structure rather than actually determining it.

The prediction of protein structure from the amino acid sequence is a work-in-progress. Scientists are cataloguing and using the known structures of thousands of proteins to help them through this process. The Protein Data Bank, or PDB, is maintained through Brookhaven National Laboratory. As of March 3, 1999, the PDB holds 9419 coordinate entries, of which 8751 are proteins, 656 are nucleic acids, and 12 are carbohydrates (Protein Data Bank). These structures are classified into groups, the most general of which being the Class (α, β, α/β, and $\alpha+\beta$); major structural similarities place proteins in the same Fold category; some

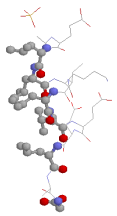
degree of sequence similarity implies a probable common ancestry, and puts proteins in the same Superfamily; and greater than 25 percent sequence similarity demonstrates a clear evolutionary ancestry, which places proteins in the same Family (Brutlag lecture, 2/1). The classification of proteins into such groups aids in understanding and attempting to predict protein structures by allowing easy observation of and comparisons between patterns in amino acid sequences.

The Asilomar conferences of 1994 and 1996 discussed four approaches to secondary structure prediction. The first is homology modeling. Two proteins are generally agreed to have the same structure if their sequences are 25-30 percent homologous (Brutlag lecture, 2/1). This approach utilizes knowledge of a closely related protein to predict the structure of a protein in question. If the sequences and/or structures of no closely related proteins are known, however, *ab initio* prediction appeals as a second approach. *Ab initio* methods attempt to predict secondary structure through knowledge of only the amino acid sequence of the protein in question (Altman lecture). One *ab initio* method that has been worked on is determining the lowest energy configuration possible, determined through a hidden Markov models and computer modeling, using the given sequence of amino acids. Such an approach, however, has not proven successful beyond predicting the secondary structure of small proteins because naturally occurring proteins often do not exist in their minimum energy configuration for reasons that may or may not be known (Brutlag lecture, 2/4). For proteins that have some (< 25 percent) sequence homology with known structures, a third approach to structure prediction is taken. Fold recognition utilizes knowledge of existing structures to hypothesize whether or not new sequences could acquire

such structures (Altman lecture). Predicting how two proteins fit together is done by the fourth approach, protein docking. The geometry of the physical association between two proteins is predicted by studying the surface-to-surface interactions to determine the best way in which they would fit together. Homology and *ab initio* are the two current methods that will be concentrated on as efforts towards protein structure prediction.

Most efforts at predicting secondary structure concentrate on predicting the state of an amino acid in the center of a local window of residues (Schmidler). Because the twenty amino acids do not occur in equal distribution in proteins, the beginnings of structure prediction attempted to utilize the frequency of an amino acid's occurrence in different conformations. For example, proteins usually have low levels of methionine and tryptophan and higher levels of leucine and serine (Stryer). In particular, however, the amino acids do not have the same proportions in particular regions of a protein forms a secondary structure as they do in the protein overall. The side chains on the amino acids can either promote or hinder secondary structure formation. Proline disrupts α-helical structure because it has no hydrogen on its N-terminus, prohibiting it from participating in hydrogen-bonding. As a result, it is often found near the ends of helical regions where turns in the chain are located; glycine and asparagine also have a propensity for forming such turns. Amino acids with large, bulky R groups, such as isoleucine, would tend to destabilize an α-helix, thus preventing helix formation. Depending upon the pH of the surrounding solution, charges on side chains can prevent helix formation. For example, at physiological pH, a polyarginine molecule would not become helical because its

R group would be positively charged. Bulky, charged side chains on amino acids also hinder the formation of β-pleated sheets. The hydrophobicity / hydrophilicity of side chains must be taken into account in all cases, as these forces are the strongest in guiding structural conformation.



Attempting to predict secondary structure on a residue-by-residue basis, as common sense would dictate, is not the best approach. Indeed, the interactions between amino acids between immediate neighbors, and possible interactions between amino acids that are located at some distance away from each other, should be taken into account in order to provide a more honest method of prediction. For example, in an α-helix, the C=O group of residue *n* is hydrogen bonded to the N-H group of residue (*n* + 4). β-sheets have hydrogen bonds between C=O and N-H groups in distant regions of the same chain of amino acids, or even on different strands of residues. Consideration of the environment of the protein can also provide great insight into the possible structure of the protein. For instance, a repetitive nature in the degree of hydrophobicity of amino acids can indicate a resulting secondary structure in which the hydrophobic side-chains face one side of the molecule while the hydrophilic side-chains face the other, forming an 'exterior' and an 'interior' to the region of the protein. An example is shown in the amphiphilic α-helix above, where all the hydrophobic residues (shown in ball-and-stick form) occupy one side of the helix. As a result, taking into account possible non-local interactions between, as well as periodicity among, residues is a must when attempting to predict secondary structure.

The most successful model yet constructed for secondary structure prediction is one proposed by Frishman and Argos (1997). At a level of 75 percent accuracy, the model relies on a local pairwise alignment of the sequence with each related sequence it is being compared to rather than initially conducting a multiple alignment, in addition to taking into account regional and nonlocal interactions between residues. Accuracy is reported on a per-residue-basis. The multiple alignment procedure assumes an evolutionary relationship between homologous sequences of different proteins.

The most accurate *ab initio* prediction of a sequence without known structural homologues was achieved by Salamov and Solovyev (1997) by using a variant on the nearest-neighbor approach. The nearest neighbor method starts with a region of residues and searches the Protein Data Bank for the sequence's 'nearest neighbors,' determining the structure of the sequence as the conformation that most of the nearest neighbors take. Salamov and Solovyev modify this procedure by allowing gaps in the alignment process, and also by combining sequence scores with environmental scores. The environmental score is determined by the area of the residue buried in the protein and inaccessible to the solvent in which the protein resides, the fraction of side-chain area that is covered by polar atoms (O and N), and the local secondary structure (Bowie *et al.*). By combining these techniques, a 72 percent accuracy in structural prediction was achieved.

Another probabilistic *ab initio* approach to structure prediction is utilization of the hidden Markov model. The hidden Markov model (HMM) attempts to statistically represent stationary signals. It has been most greatly exploited in speech recognition projects, but its significance has carried over into

hand-written script recognition and, more relevantly, the modeling of protein chains. The idea of using a HMM to predict secondary structure was first introduced by K. Asai *et. al.* in 1993. A programmed HMM can 'learn' protein secondary structures such as the α-helix, β-sheet, and the turn, and these HMMs can then, in turn, be applied to new sequences whose structures are unknown. The HMM gives an output of probabilities for the secondary structure, which are used to predict the secondary structures of the sequences. The model is used to predict a structure using only the sequence in question, rather than using homologous sequences, as well.

New approaches to protein modeling are continually breaching the expected limits to the accuracy of secondary structure prediction. One possible reason for this is that any predictive method is immediately outdated due to the rate at which new structures are being entered into the Protein Data Bank. The PDB is updated weekly, and the last update resulted in 69 new entries. If as few as 50 structures per week were entered, 2600 structures would have been added in the past year. Schmidler *et al*. developed a technique using HMMs that, though it does not yield the highest level of accuracy, has advantages over the existing models. The model easily incorporates the ever-expanding information on protein structure and is more flexible than other HMMs in allowing use of information that is already available. For instance, the N-terminus of a helical cap shows strong signals for the first and second positions; proline and alanine most often occupy position N1, while glutamic acid is most frequently found in position N2. Such positions which are most important for predictive purposes can be highlighted in the model, which allows the most significant differences from sequences in the database to be focused on. The incorporation of this

knowledge, however, would be difficult to execute in the standard window-based approach. Intra-segment residue correlations can also be deduced in the Schmidler *et al.* models. In addition, these models calculate the exact degree of uncertainty at each position within the segment, rather than the uncertainty of the secondary structure of the region as a whole. As a result, the models can be easily modified to be conditional upon specific positions or segments taking known conformations. For these reasons, though the approach stands at 68.8 percent accuracy while the best method using single sequences stands at 71 percent (Salamov and Solovyev), the possibility of the model achieving higher rates of accuracy in prediction as more information is gathered and applied is not ruled out. In addition, application of the model to multiple sequence alignments, as well as consideration of non-local interactions, could produce even more accurate performance.

While the accuracy of protein structure prediction is worked on, the actual structures of proteins continue to be derived. At a 1998 meeting sponsored by the National Institute of General Medical Sciences (NIGMS) at the National Institutes of Health, scientists agreed that a database of 3,000-5,000 new protein structures must be determined experimentally in order to discover all protein folding motifs (National Institutes of Health). Since the database was started in 1973 approximately 7000 structures had been submitted and released at the end of 1997 (Protein Data Bank), compared to the approximately 9000 protein structures that exist in the database today, a little over a year later. The pace at which such information is being acquired continues to provide insight into the classification of protein motifs. The determination of as many types of protein

folds as possibly will invariably expedite the currently slow but steady increase in the accuracy of secondary structure prediction models.

Much progress has been made since attempts to predict secondary structure began. Predictive methods have increased almost 20 percent from the original 56 percent accuracy rate reported by pioneers in the field (Garnier *et al.*) . Though limits to accuracy without resorting to looking at thermodynamic perspectives may exist, many present models may be significantly improved upon. As a result, the limits on the accuracy of probabilistic approaches to protein structure prediction may not yet have been reached. The eventual production and application of a highly accurate model would provide an extremely beneficial perspective on the treatment of inherited diseases. The development of such a model would begin to solve the puzzle of how an amino acid sequence inherently holds the necessary information to bend, twist, and fold into the proper conformation in order to function correctly- a problem that has been plaguing scientists since the genetic code was cracked more than 30 years ago. The capability to accurately predict secondary structure would then allow scientists to focus on tertiary and quaternary structure, which would culminate as the ability to map a direct link from amino acid sequence to full structure, from a linear map to a phenotype, of a protein.

References

Altman, Russell. Slides from Computational Molecular Biology- Protein Structure Prediction and Threading lecture in Biochemistry 218: Genomics and Bioinformatics, November 16, 1998. http://cmgm.stanford.edu/biochem218/16Threading.html.

Asai, K., Hayamizu, S. and Handa, K. "Prediction of protein secondary structure by the hidden Markov model." *Computer Applications in the Biosciences*, 1993 Apr, 9(2):141-6.

Bowie, JU, Lüthy, R and Eisenberg, D. "A method to identify protein sequences that fold into a known three-dimensional structure." *Science*, 1991 Jul 12, 253(5016):164-70.

Brutlag, D. Lectures in Biochemistry 118Q: Genomics and Bioinformatics. Stanford University, Winter quarter, 1999.

Frishman, D. and Argos, P. "Seventy-five percent accuracy in protein secondary structure prediction." *Proteins*, 1997 Mar, 27(3):329-35.

Garnier, J., Osguthorpe, D.J. and Robson, B. "Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins." *Journal of Molecular Biology*, 1978 Mar 25, 120(1):97-120.

National Institutes of Health homepage: NIGMS -- Protein Structure Initiative Meeting Summary 4-24-98. http://www.nih.gov/nigms/news/reports/protein_structure.html.

Protein Data Bank homepage. http://pdb.pdb.bnl.gov.

Salamov, A.A. and Solovyev, V.V. "Protein secondary structure prediction using local alignments." *Journal of Molecular Biology*, 1997 Apr 25, 268(1):31-6.

Schmidler, S. C., Liu, J.S. and Brutlag, D. L. "Bayesian Segmentation of Protein Secondary Structure."

Stryer, Lubert. *Biochemistry*. Fourth edition. New York: W.H. Freeman and Company, 1995.