

Critical Analysis of Secondary Structure Prediction Algorithms

Elizabeth D. Chao

June 6, 2002

Biochemistry 218: Computational Molecular Biology
Professor Doug Brutlag
Stanford University

INTRODUCTION

The sequence-structure gap:

The explosive accumulation of protein sequences in the wake of large-scale genome sequencing projects is in stark contrast to the much slower experimental determination of protein structures. Despite significant improvements of structure determination methods, the gap between knowledge of protein structure and protein sequence is rapidly increasing. Thus, to acquire a full understanding of the biological role of these proteins, knowledge of their structure and function is essential. Computational structure prediction methods have sought to meet the challenge of bridging the sequence-structure gap in providing valuable information for the large fraction of sequences whose structures have not yet, or may not ever, be determined experimentally.

The basis for secondary structure prediction:

A long-term goal of the protein-folding problem is to be able to predict the folded three-dimensional structure of a protein from its amino acid sequence alone. Secondary structure prediction is often regarded as the initial starting point in predicting the three-dimensional structure of a protein. Fundamentally, it attempts to classify amino acids in protein sequence according to their predicted local structure, which can be subdivided into three states: α -helix, β -sheets, or loops. While the number of states may vary depending on the algorithm employed, we will simplify our analysis to a three-state problem, Q_3 , where turns, coils, or other helices will collectively be called “loops”.

Principle Assumptions:

The fundamental assumption on which all secondary structure prediction methods are based is that there should be a correlation between amino acid sequence and secondary structure. Because the entire information for forming secondary structure is contained in the primary sequence, any short stretch of amino acid sequence will preferentially adopt one kind of secondary structure over another. Thus, many algorithms examine a sequence window of 13-17 residues, assuming that the central amino acid in the window will adopt a conformation that is determined by the side groups of all the amino acids in the window. For α -helices, this window is typically 5-40 residues long, and for β -sheets, this window ranges from 5-10 residues in length. While earlier algorithms assumed that each amino acid within the sequence window was unaffected by other neighboring amino acids, later methods recognized the oversimplification, and accounted for the possibility that more distant interactions within the primary amino acid chain may influence local secondary structure.

Secondary structure prediction algorithms:

The three most widely used methods of protein secondary structure prediction include:

- Chou-Fasman and GOR methods
- neural network models
- nearest-neighbor methods

Chou-Fasman and GOR methods:

In 1974, Chou and Fasman developed a statistical method based on the propensities of amino acids to adopt secondary structures based on the observation of their location in 15 protein structures determined by X-ray diffraction. These statistics derive from the particular stereochemical and physicochemical properties of the amino acids and are shown in **Table 1**. Over the years, these statistics have been refined using a larger set of proteins.

Unlike Chou-Fasman which assumes that each amino acid individually influences secondary structure within a window of sequence, GOR (Garnier, Osguthorpe, and Robson) takes into account the influence on secondary structure of the amino acids flanking the central amino acid residue. In the most recent version of GOR (GOR IV), certain pairwise combinations of amino acids in the flanking region or of a flanking amino acid and the central residue can influence the conformation of the central amino acid. For instance, if a particular amino acid is surrounded with residues that prefer to be in a helix, it is likely to be in a helix, even if its individual helical preference is low. Thus, instead of considering propensities for a single residue, position-dependent propensities for helix, sheet and turn has been calculated for all residue types. It is based on the same P_{ij} values as in the second Chou Fasman method.

Programs: DPM, DSC, GOR IV

Table 1: Chou-Fasman statistics for the secondary structure preferences of amino acids

Name	P(H)	P(E)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	142	83	66	0.06	0.076	0.035	0.058
Arginine	98	93	95	0.07	0.106	0.099	0.085
Aspartic Acid	101	54	146	0.147	0.11	0.179	0.081
Asparagine	67	89	156	0.161	0.083	0.191	0.091
Cysteine	70	119	119	0.149	0.05	0.117	0.128
Glutamic Acid	151	37	74	0.056	0.06	0.077	0.064
Glutamine	111	110	98	0.074	0.098	0.037	0.098
Glycine	57	75	156	0.102	0.085	0.19	0.152
Histidine	100	87	95	0.14	0.047	0.093	0.054
Isoleucine	108	160	47	0.043	0.034	0.013	0.056
Leucine	121	130	59	0.061	0.025	0.036	0.07
Lysine	114	74	101	0.055	0.115	0.072	0.095
Methionine	145	105	60	0.068	0.082	0.014	0.055
Phenylalanine	113	138	60	0.059	0.041	0.065	0.065
Proline	57	55	152	0.102	0.301	0.034	0.068
Serine	77	75	143	0.12	0.139	0.125	0.106
Threonine	83	119	96	0.086	0.108	0.065	0.079
Tryptophan	108	137	96	0.077	0.013	0.064	0.167
Tyrosine	69	147	114	0.082	0.065	0.114	0.125
Valine	106	170	50	0.062	0.048	0.028	0.053

$$P_k(i) = \frac{P(k/i)}{P(k)}$$

In which:

- k designates the given state (helix, sheet, loop).
- $P(k/i)$ is the frequency of observation of the state k for amino acid i .
- $P(k)$ is the frequencies of observing state k .
- $P(k,i)$ is the parameter of amino acid i for state k .

Neural network models

A neural network is comprised of a machine learning approach, providing computational processes the ability to “learn” in an attempt to simulate the complex patterns of synaptic connections formed among neurons in the brain during learning. Computers are trained to recognize patterns in known secondary structures using training sets of non-homologous structures, and tested with proteins of known structure. An example of one commonly used neural network, PHD, is illustrated below in [Figure 1](#).

Neural networks have been able to achieve a level of 73% overall three-state per-residue accuracy. The reasons for improved prediction accuracy is attributed to its ability to align the query sequence with other related proteins of the same family and find protein members with known structures to aid its assignment of secondary structures. While neural networks can detect interactions between amino acids within a window of amino acids, neural networks have great difficulties in dealing with variable length motifs because the input layer is typically a rigid structure with a fixed number of cells, accepting sequences of only one length class. Furthermore, neural nets are designed as black-box methods. While the weights of a weight matrix are usually known to the user and lend themselves to a physical interpretation, the parameters of a neural net are hidden and not meant to be biologically interpretable or of interest to the user. Thus, while neural nets may be very powerful function prediction tools, they usually do not tell us anything about the underlying molecular recognition process.

Programs: PHD, PSIPRED, NNpredict

Nearest-neighbor methods

The basic idea of the nearest-neighbor approach is the prediction of the secondary structure state of the central residue of a test segment, based on secondary structure of homologous segments from proteins with known three-dimensional structure. It is performed by finding some number of the closest sequences (from a database of proteins with known structure) to a subsequence defined by a window around the amino acid of interest. Using the known secondary structures of the aligned sequences (generally weighted by their similarity to the target sequence) a secondary structure prediction is made. For instance, a large list of short sequence fragments is made by sliding a window of defined sequence length along a set of ~400 training sequences of known structure that are non-homologous to each other, and recording the secondary structure of the central amino acid of each window. Subsequently, a window of the same size is then selected from the query sequence and compared to the list of short sequence fragments to identify the 50 best matches. The frequency that the central amino acid in each of the 50 matching fragments will form a particular secondary structure is then used to predict the secondary structure of the central amino acid in the query sequence. The variability in nearest neighbor methods arises from the selection of subsequences closest to a window around the amino acid whose structure is being predicted. Each program uses a different set of parameters, like how similarity is defined, or what sequence window size should be examined.

Programs: SOPM, SOPMA, NNSSP, PREDATOR

Figure 1: Schematic of the PHD algorithm

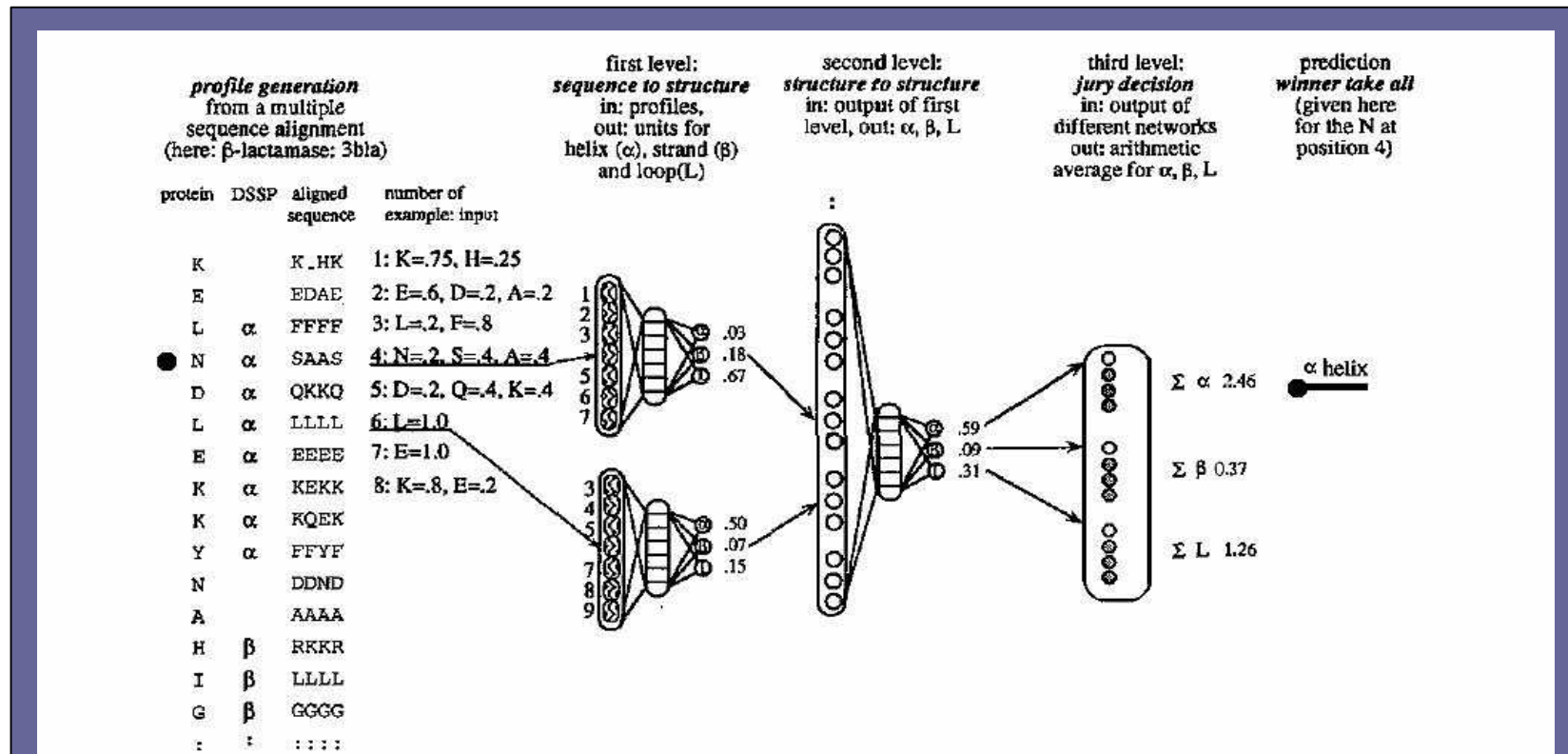


Figure 1: Schematic of PHD algorithm. PHD starts from a multiple sequence alignment and then uses three layers of networks to predict the secondary structure of the central residue. **Step 1:** Homology information is obtained using a Blast search of all related proteins. **Step 2:** A multiple sequence alignment is generated from the homologous proteins using MaxHom. The frequencies of different amino acids (and gaps) in each position and some other information is used for the prediction. **Step 3:** The occurrence of various residues in a sequence window of 13 amino acids is correlated with the secondary structure of the central residue. **Step 4:** In the structure-to-structure layer, the output from the first layer in a window of 17 residues is used to predict the secondary structure of the central residue. In this case, the network will be trained not to predict unreasonably short segments of secondary structure. **Step 5:** Several networks (3-12 depending on PHD version) are combined into a jury prediction network. This improves the prediction accuracy with about 2%. **Step 6:** A simple filtering method that, for instance, changes HHEHH to HHHHH is applied. No significant change in performance is obtained.

SPECIFIC AIMS

In this paper, I have used a fully crystallized protein from each class of proteins: all α -helix, all β -sheet, alpha/beta, and engineered alpha+beta protein to evaluate the strengths and weaknesses of nine different secondary structure prediction algorithms. In doing so, I also hoped to address the following questions:

- Are there differences in the ability of secondary structure prediction algorithms to detect alpha-helices vs. beta-strands? If so, what differences? And why?
- Are the algorithms capable of predicting the secondary structure of an engineered protein?
- Are tertiary interactions critical for accurate secondary structure prediction? If so, how?
- What affect does helical capping have on secondary structure prediction?
- Are buried helices predicted at lower accuracy than exposed helices?
- Is the use of multiple sequence alignments in secondary structure prediction a great advantage?
- Does prediction accuracy increase by combining the results from multiple programs into a consensus sequence?

METHODS

To aid us in the evaluation of the strengths and weaknesses of each secondary structure algorithm, I have provided a detailed description of the basic principles behind each of the programs used in this paper.

ALGORITHM	DESCRIPTION
DSSP	Dictionary of protein secondary structure. Pattern recognition of hydrogen-bonded and geometrical features. Defines secondary structure, geometrical features and solvent exposure based on the atomic co-ordinates from PDB files, but does not predict structures. It bases assignments on hydrogen bonding patterns and backbone dihedral angles.
	NEURAL NETWORK ALGORITHMS
PHD	Jury decision neural networks. Neural networks of multiple sequence alignment. PHD was the first program to use evolutionary information derived from aligned homologous sequences. It is based on a two-layered feed-forward neural network. In the neural network, aligned homologous sequences of known structures are used to "train" the network, which then can be used to predict the secondary structure of the aligned sequences of the unknown protein. See Figure 1 for details. The method also applies balanced training, percentage amino acid composition and conservation, sequence length, and insertions and deletions to enhance prediction accuracy
PSIPRED (Secondary structure prediction based on Position Specific Iterated-Blast)	Divergent profile (PSI-Blast) based neural network prediction. PSIPRED is a simple and reliable secondary structure prediction method, incorporating two feed-forward neural networks that perform an analysis on output obtained from PSI-BLAST (Position Specific Iterated-BLAST). PSI-BLAST refers to a feature of BLAST in which a profile (or position specific scoring matrix, PSSM) is constructed from a multiple alignment of the highest scoring hits in an initial BLAST search. Highly conserved positions receive high scores and weakly conserved positions receive scores near zero. The profile is used to perform a second (etc.) BLAST search and the results of each "iteration" used to refine the profile. This iterative searching strategy results in increased sensitivity. PSIPRED is capable of achieving an average Q ₃ score of nearly 78%. The results are returned as a graphical jpeg representation of the secondary structure prediction.
nnPREDICT	Neural network method. The nnPREDICT algorithm uses a two-layer, feed-forward neural network to assign the predicted type for each residue. In making the predictions, the server uses a FASTA format file with the sequence in either one-letter or three-letter code, as well as the folding class of the protein. Residues are classified as being within an α -helix, β -strand, or neither. For the best-case prediction, the accuracy rate of nnpredict has been reported as being over 65%.

	Chou-Fasman/GOR algorithms
GOR IV	Secondary structure prediction using information theory. Consideration of residue pairs. The GOR method uses information theory to formulate the influence of local sequence upon the conformation of a given residue. However, the existing database does not allow the evaluation of parameters required for an exact treatment of the problem. GOR IV considers all possible pair frequencies within a window of 17 amino acid residues, improving GOR I to a mean accuracy of 64.4% for a three state prediction (Q3). The predicted secondary structure is the one of highest probability compatible with a predicted helix segment of at least four residues and a predicted extended segment of at least two residues.
DPM (Double Prediction Method)	Chou-Fasman and class prediction. DPM consists of a first prediction of the secondary structure from a new algorithm that uses parameters of the type described by Chou-Fasman (Table 1), and the prediction of the class of the proteins (α , β , α/β , $\alpha+\beta$) from their amino acid composition. These two independent predictions allow one to optimize the parameters calculated over the secondary structure database to provide the final prediction of secondary structure. It can be summarized as 4 successive steps: (1) Prediction of the structural class of a protein from amino acid composition according to Nakashima <i>et al.</i> , 1986. The parameters used for the prediction of the class have been determined onto a set of 135 proteins with known structures. They are just amino acid percentages calculated on isolated classes of α , β , α/β , and $\alpha+\beta$ proteins. (2) Preliminary secondary structure estimation from a simple algorithm, (3) Comparison between the 2 independent predictions, (4) Optimization of parameters and re-prediction of secondary structure.
DSC (Discrimination of Protein Secondary Structure Class)	GOR and linear discrimination of multiple sequence alignment. DSC applies GOR residue attributes, with the addition of hydrophobicity and amino acid position, which are combined with information from the multiple sequence alignment. The important concepts in secondary structure prediction are identified as: residue conformational propensities, sequence edge effects, moments of hydrophobicity, position of insertions and deletions in aligned homologous sequence, moments of conservation, auto-correlation, residue ratios, secondary structure feedback effects, and filtering. Optimal weights are deduced by linear discrimination, with filtering applied to remove erroneous predictions. This method has an advantage in that the prediction method is both implicit and effective.

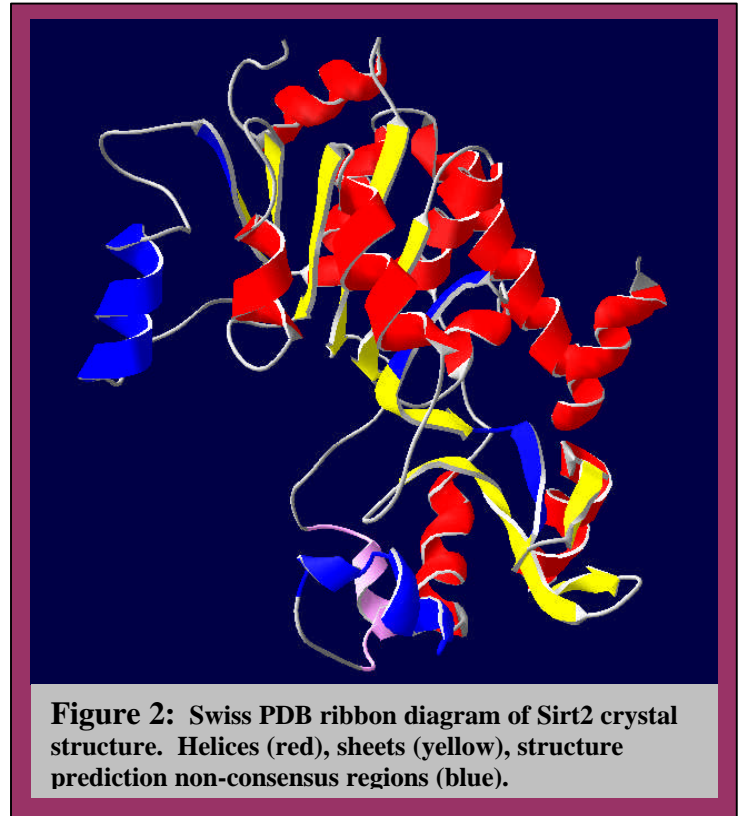
	Nearest-neighbor Algorithms
SOPM (Self-Optimized Method for protein secondary structure prediction) & SOPMA	<p>Nearest-neighbor method. A new method called the self-optimized prediction method (SOPM) seeks to improve the success rate in the prediction of the secondary structure of proteins in the following way. The first step of the SOPM is to build subdatabases of protein sequences and their known secondary structures drawn from a 'DATABASE.DSSP' of 239 proteins by (i) making binary comparisons of all protein sequences and (ii) taking into account the prediction of structural classes of proteins. The second step is to submit each protein of the subdatabase to a secondary structure prediction using a predictive algorithm based on sequence similarity. The third step is to iteratively determine the predictive parameters that optimize the prediction quality on the whole sub-database. The last step is to apply the final parameters to the query sequence. This new method correctly predicts 69% of amino acids for a three-state description of the secondary structure in the whole database (46,011 amino acids). Improvements on SOPM were brought about by predicting all the sequences of a set of aligned proteins belonging to the same family. This improved SOPM method (SOPMA) correctly predicts 69.5% of amino acids for a three-state description of the secondary structure in a whole database containing 126 chains of non-homologous (less than 25% identity) proteins.</p>
PREDATOR	<p>Hydrogen-bonding propensities and nearest neighbor classifier. PREDATOR incorporates non-local interactions in protein secondary structure prediction from the amino acid sequence. It is based on recognition of potentially hydrogen-bonded residues in the amino acid sequence. PREDATOR uses database-derived statistics on residue-type occurrences in different classes of local hydrogen-bonded structures in such a way that it can differentiate the hydrogen-bonding interactions between adjacent $\hat{\alpha}$-strands, parallel vs. anti-parallel $\hat{\alpha}$-strands, and amino acids (i, i+4) on $\hat{\alpha}$-helices. Seven different secondary structure propensities are generated for the query sequence, with a nearest neighbor implementation applied to calculate propensities for $\hat{\alpha}$-helix, $\hat{\alpha}$-strand and loop. The novel feature of PREDATOR is its reliance on local pair-wise alignment of the sequence to be predicted between each related sequence. The algorithm has a prediction accuracy of 68% in three structural stages, relies only on a single protein sequence as input and has the potential to be improved by 5-7% if homologous aligned sequences are also considered.</p>
NNSSP (Nearest-neighbor secondary structure prediction)	<p>Scored nearest neighbor method. The NNSSP method combines nearest-neighbor algorithms and multiple sequence alignments. It improves upon the local structural environment scoring scheme developed by Bowie <i>et al.</i>, which assigns every residue of a protein with known three-dimensional structure to an "environment class" based on the local structural features of the residue position, such as solvent accessibility, polarity, and secondary structure. In addition to incorporating the Bowie scheme, NNSSP takes into consideration N- and C-terminal positions of α-helices and β-strands and also β-turns as distinctive types of secondary structure. Another improvement, which also significantly decreases the time of computation, is performed by restricting a database with a smaller subset of proteins that are similar with a query sequence. Using multiple sequence alignments rather than single sequences and a simple jury decision method we achieved an over all three-state accuracy of 72.2%, which is better than that observed for the most accurate multilayered neural network approach, tested on the same data set of 126 non-homologous protein chains. The size of the database used for scanning is also altered to reflect similarity to the query sequence, reducing computation time, and improving the final accuracy.</p>

RESULTS AND DISCUSSION

Case Study 1: alpha/beta proteins

Sirt2 histone deacetylase (*Homo sapiens*)

The deacetylation of histones is an important phenomenon in eukaryotic gene regulation that results in tighter chromatin structure and transcriptional repression. *Saccharomyces cerevisiae* Sir2, the defining member of a novel family of NAD⁺-dependent histone deacetylases, called sirtuins (Sir2-like proteins), functions in the establishment of silenced chromatin at the mating-type, rDNA, and telomeric loci, and participates in cell cycle regulation, double-stranded break repair, meiotic checkpoint control, and aging. Uniquely, Sir2 has been found to play a role in lengthening the life span of yeast due to caloric restriction. The sirtuin family is evolutionarily conserved across numerous prokaryotic and eukaryotic organisms, such as *Salmonella typhimurium*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Homo sapiens*, suggesting that their mechanisms of action may be similar. To date, the most extensively studied sirtuin is the yeast Sir2 protein itself, while the physiological functions and molecular interactions associated with other classes of sirtuins are still largely unknown.



Recently, the 1.7 Å crystal structure of the 323 amino acid catalytic core of human Sirt2, a homolog of yeast Sir2, was obtained (Finnin *et al.*, 2001). Sirt2 has a 304-amino acid catalytic core and a 19-residue N-terminal helical extension. It is comprised of two domains: a larger domain (residues 53-89, 146-186, 241-356) that is an inverted variant of the classical Rossmann fold, and a smaller domain (residues 90-145 and 187-240) that contains the zinc binding domain. Sirt2 falls under the SCOP classification of an alpha/beta protein with mainly parallel β -sheets and β - α - β units. More specifically, Sirt2 is comprised of a deoxyhypusine synthase (DHS)-like NAD/FAD-binding domain with parallel β -sheets of six strands in the order 321456 and six α -helices packed against the β -sheet. Furthermore, as part of the Sir2 family of transcriptional regulators, it contains an insertion of a rubredoxin-like zinc finger domain.

Table 2: Pairwise Comparison of Secondary Structure Prediction Algorithms for Sirt2

	DPM	DSC	GOR4	PHD	Predator	SOPM	SOPMA	Consensus	DSSP
DPM	100%	75%	71%	62%	77%	74%	74%	81%	66%
DSC	75%	100%	69%	71%	74%	78%	75%	83%	66%
GOR4	71%	69%	100%	57%	72%	78%	72%	79%	65%
PHD	62%	71%	57%	100%	62%	69%	68%	72%	67%
Predator	77%	74%	72%	62%	100%	73%	71%	82%	67%
SOPM	74%	78%	78%	69%	73%	100%	84%	89%	76%
SOPMA	74%	75%	72%	68%	71%	84%	100%	84%	71%
Consensus	81%	83%	79%	72%	82%	89%	84%	100%	73%
DSSP	66%	66%	65%	67%	67%	76%	71%	73%	100%

The most successful algorithms for predicting the secondary structures of the alpha/beta protein, Sirt2, were PSIPRED (79%) and SOPM (76%), while the worst algorithms were Chou-Fasman (39%) and GOR I (48%). PSIPRED's incorporation of evolutionary information through conserved regions in multiple sequence alignments performed by PSI-BLAST probably helped achieve the extraordinary level of structure prediction over the

other algorithms. In this specific case, Sirt2 belongs to a highly conserved, and well-defined family of Sir2 histone deacetylases, enabling the PSI-BLAST sequence alignment to be particularly effective. In contrast, Chou-Fasman's assumption that each amino acid operates independently in affecting secondary structure is clearly insufficient for accurate secondary structure prediction. Even GOR's attempt to take into account the influence on secondary structure of the amino acids flanking the central amino acid residue may not suffice.

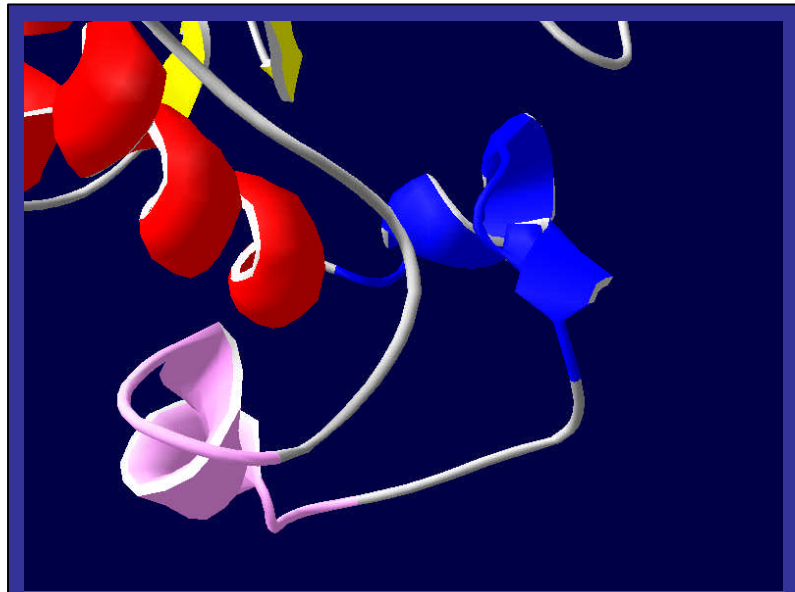


Figure 3: Sirt2 crystal structure, residues 100-115 (pink). DPM, PHD, SOPM, and nnPREDICT are the only methods sensitive enough to correctly predict that the region suspended outside on the surface of the protein between residues 100-115 are alpha-helical. On the other hand, Chou-Fasman predicts the residues form a beta-strand, and the other programs, GOR4, Predator, PSIPRED, NNSSP predict that it is just a loop. Note that it is possible that the “loop” region may also include another kind of alpha-helical structure that is not defined using the three-state per-residue prediction assumptions we have made. Thus, for our purposes we will ignore the fact that it is not entirely correct to say that GOR4, Predator, PSIPRED, and NNSSP were inaccurate in predicting a loop—in fact, they may have realized a more subtle distinction in secondary structure that doesn't exactly fit the oversimplified categories we have defined.

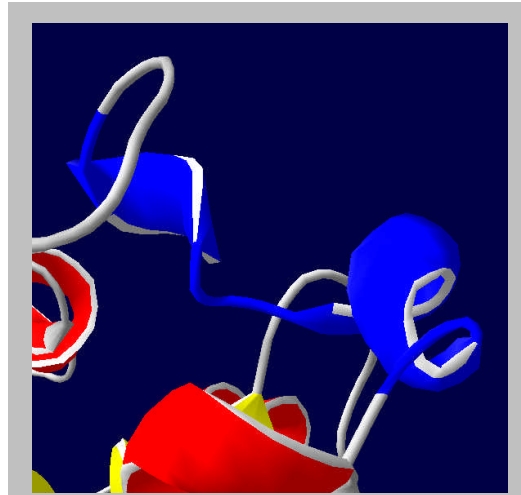


Figure 4: Sirt2, residues 115-127 (blue). All the algorithms contradict each other regarding the alpha/loop/alpha structure of the residues between 115-127. The confusion is understandable, as the alpha helices are short and the loop regions are twisted into a structure that is ambiguous, and can be easily mistaken to be a beta-strand. PHD and Predator predict a beta-strand that spans the first alpha helical region, while SOPM predicts a beta-strand in the middle of a helical region. While DSC over predicts the length of the alpha helical region, GOR4 under predicts it. Only PSIPRED and NNSSP are the closest to being correct. PSIPRED correctly returns a short alpha-helix between 115-118, followed by a loop, and then the alpha helical region 122-126. Similarly, NNSSP predicts an alpha helical region that spans the mini-loop. Generally, NNSSP is better at predicting alpha-helical regions because it takes into consideration N- and C-terminal positions of alpha-helices as distinct types of secondary structure. Moreover, it filters out short helical regions rigorously to increase its accuracy by converting any mixed alpha/beta regions shorter than 3-4 amino acids completely to alpha helices. By increasing its sensitivity towards detecting alpha helices, NNSSP also sacrifices the ability to predict shorter alpha helical fragments, as in this case. It is surprising that Predator failed to recognize the alpha-helicity of this region. The most reasonable explanation is that the database from which Predator derives its statistical information on residue-type occurrences in different classes of hydrogen-bonded structures was inadequate to encompass the type of hydrogen-bonding interaction described by this amino acid region.

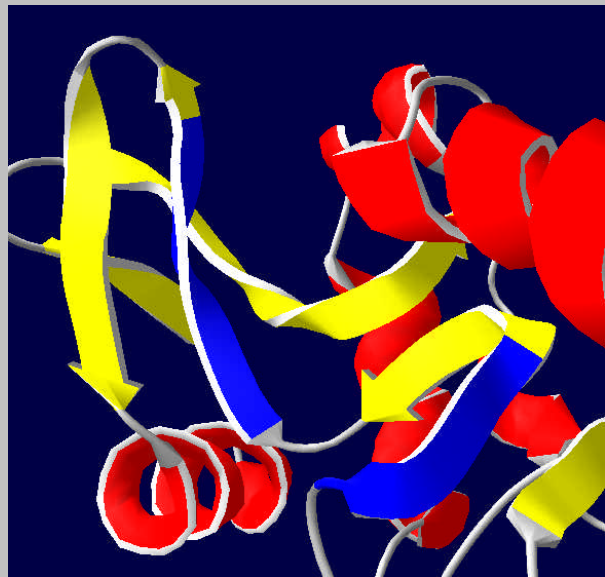


Figure 5: Sirt2, Residues 181-185 and 191-195 (blue). Another region that seemed to have caused some confusion were the two beta-strands between residues 182-185 and 191-195. The two strands seem to be twisted in such a way that may make it easy to mistaken them for helical or loop regions. While PHD and SOPM both predict helical regions, Predator predicts a loop region. Indeed, a region may have a higher preference for forming a helix than a strand, but interactions non-local in the sequence may result in that the formation of a β -sheet is energetically more favorable. Indeed, the confusion between helices and strands can often be attributed to hydrogen bonds stabilized by non-local inter-residue contacts.

Figure 6: Sirt2, residues 295-304 (blue).

Surprisingly, none of the algorithms are capable of predicting that the region enclosed by residues 295-304 is alpha-helical. PSIPRED, SOPMA, NNSSP, and nnPREDICT predict that it is a beta-strand, while PHD and Predator predict that it forms a loop. Apparent from the crystal structure, the helical region is suspended in mid-air between two loops. While it is possible to imagine that the algorithms were simply confused by the loop regions flanking the alpha-helical segment, it would also be interesting to see if this region is initially unstructured, but once it is solvent exposed, it transforms into an alpha helix due to hydrophobic interactions. For instance, a peptide region in the SH3 domain forms into an alpha helix upon exposure to solvent. Thus, a protein's secondary structure may very well be governed by interactions beyond just tertiary interactions within itself.

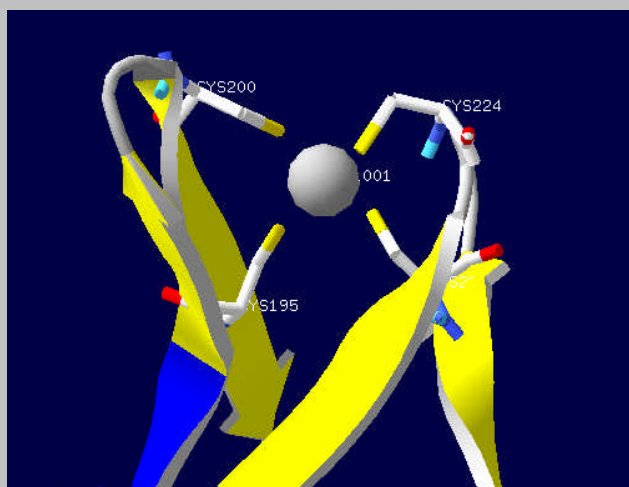
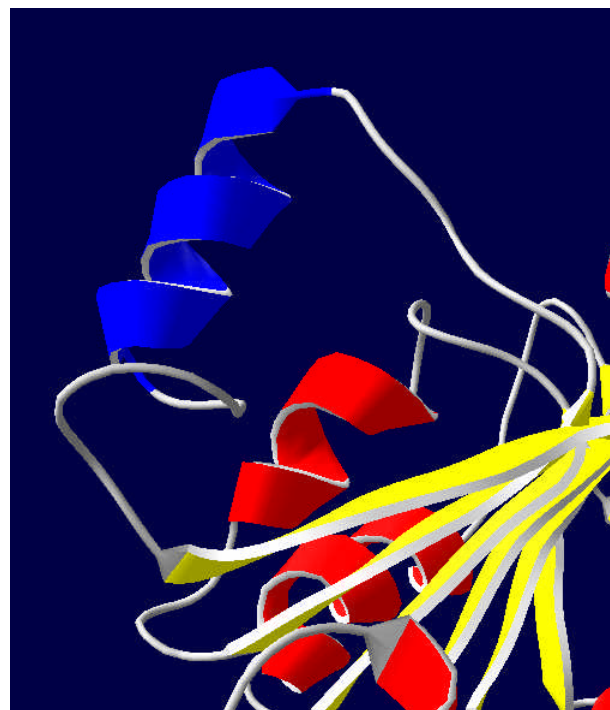


Figure 7: Sirt2 zinc binding domain with 4 cysteine residues.

The pair of beta-strand/loop motifs forming the highly conserved zinc-finger binding region does not cause any problems in the secondary structure prediction programs, most likely because it is such a well-defined, and highly conserved region in the protein.

Case Study 2: all alpha proteins

TAFII18 (*Homo sapiens*)

In eukaryotic transcription initiation, RNA polymerase II assembles into a macromolecular complex comprised of the general transcription factors TFII(A-H) at the basal promoter, which binds to the TATA element with the help of the TATA-binding protein (TBP). Associated with TBP are TAFs, or TBP-associated factors, which play several important roles in transcriptional regulation (reviewed in Goodrich *et al.*, 1996), including promoter recognition, transcriptional activation, and acting as specific coactivators. In human TFIID, cDNAs encoding eleven TAFs have been characterized, and exhibit remarkable evolutionary conservation. Human TAFII18 is a novel TAF shown to exhibit homology to the N-terminal region of the well-characterized yeast TAFII, SPT3, which interacts with TBP and is required for transcription from a subset of yeast promoters and is part of the large SAGA complex containing ADA coactivators and the Gcn5 histone acetyltransferase.

Recently, the 2.6 Å resolution crystal structure of the human TBP-associated factor (hTAFII)28/hTAFII18 heterodimer was solved, showing that these TAFs form a novel histone-like pair in the TFIID complex. The histone folds in hTAFII18 were not predicted from its primary sequence, indicating that it defines a novel family of atypical histone fold sequences, unlike core histones and other known histone fold-containing proteins. The TAFII18 histone fold motifs are also present in the N- and C-terminal regions of the SPT3 proteins, suggesting that the histone fold is a more commonly used motif for mediating TAF–TAF interactions than previously believed.

Under the SCOP classification, human TAFII18 is composed entirely of α -helices that form a non-canonical histone fold motif seen in Spt3-like transcription factors. It consists of three helices, and one long middle helix flanked at each end with shorter helices. The hTAFII18 fragment (residues 31–75) resembles a histone fold motif lacking the C-terminal 3 helix. A ten residue N-terminal 1 helix is linked to a 25-residue 2 helix by an 8-residue L1. A unique charged residue, K34, is exposed at the surface of the 1 helix, while the hydrophilic face of the 2 helix is mainly acidic (E52, D55, E58, D59, E63, and E67).

Table 3: Pairwise Comparison of Secondary Structure Prediction Algorithms for TAFII-18

	DPM	DSC	GOR4	PHD	Predator	SOPM	SOPMA	Sec.Cons.	DSSP
DPM	100%	68%	66%	66%	75%	72%	74%	74%	80%
DSC	68%	100%	89%	87%	84%	82%	81%	90%	76%
GOR4	66%	89%	100%	82%	87%	81%	78%	86%	73%
PHD	66%	87%	82%	100%	79%	78%	76%	83%	74%
Predator	75%	84%	87%	79%	100%	85%	85%	93%	81%
SOPM	72%	82%	81%	78%	85%	100%	97%	88%	81%
SOPMA	74%	81%	78%	76%	85%	97%	100%	88%	84%
Sec.Cons.	74%	90%	86%	83%	93%	88%	88%	100%	79%
DSSP	80%	76%	73%	74%	81%	81%	84%	79%	100%

Generally, the algorithms are better at predicting alpha helices and proteins dominated by alpha-helical regions. SOPMA returned an extraordinary success rate of 85% in the secondary structure prediction of TAFII-18. Moreover, the consensus sequence was, on average, better than using only one algorithm alone as the predictive method. Surprisingly, the neural network algorithm, PHD, did very poorly in the predictions of an entirely alpha-helical protein (see **Figure 8**), in fact, predicting more beta-strand regions than any of the other algorithms. Both DPM and SOPM/A may have increased success because of the use of structural class prediction—in this case, TAFII-18 easily falls into the all-alpha structural class of proteins. Falling cleanly into one of these categories allows both algorithms to properly optimize their predictive parameters for the most accurate prediction. Fortunately, this protein did not form any coiled-coils, which could not have been predicted using the algorithms here.

Surprisingly, however, despite being an entirely alpha-helical protein, two regions were confused as being beta-strands.

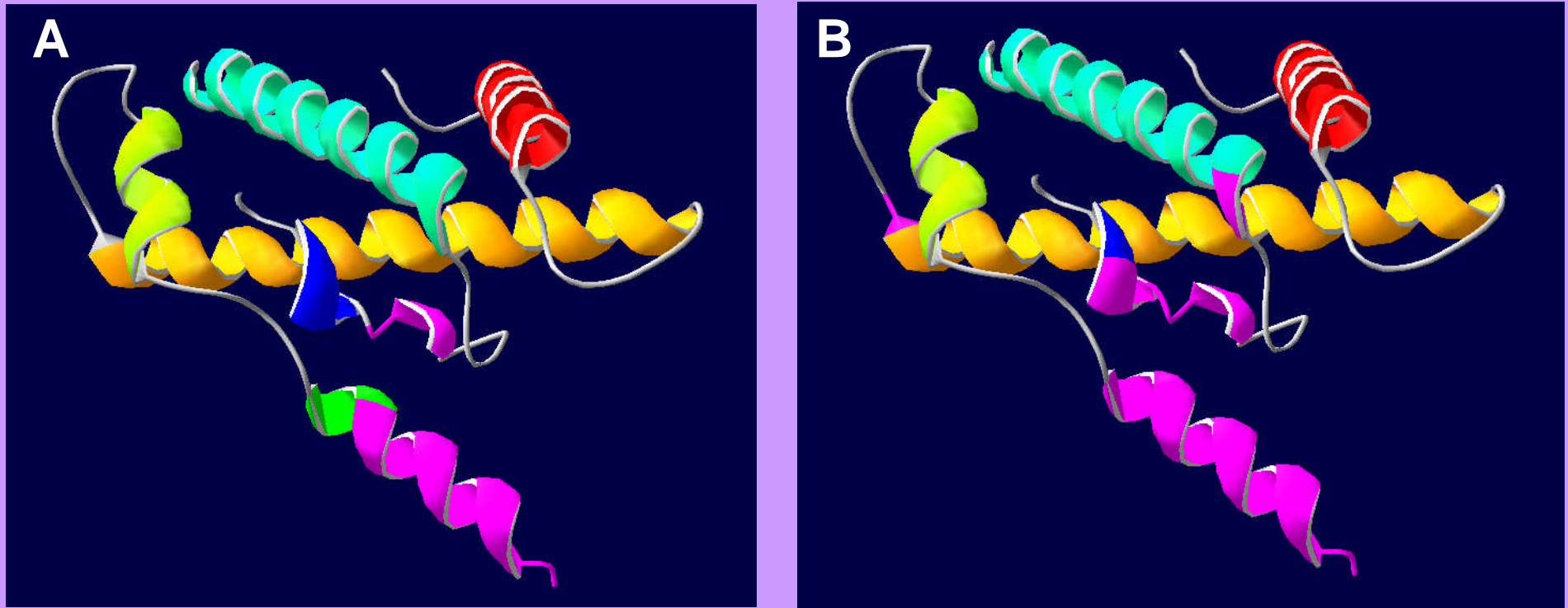


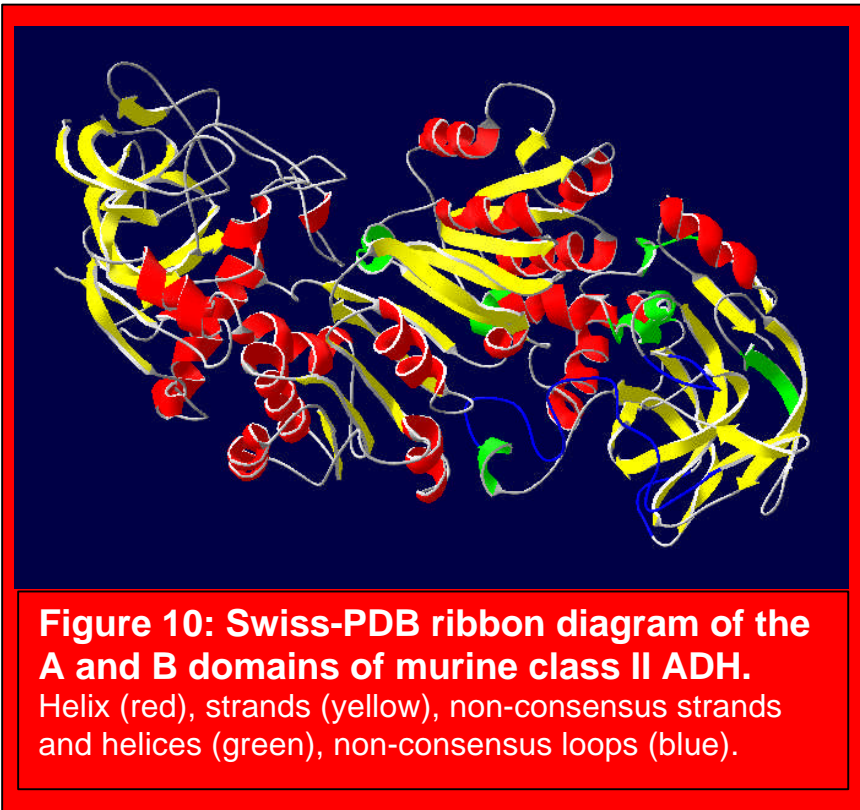
Figure 9: Swiss PDB ribbon diagram of TAFII8-18, secondary structure colored in succession. (A) Regions of confusion among secondary prediction algorithms (magenta) (B) PHD inaccurately predicted regions of secondary structure (magenta).

The first region constitutes residues 38-41, which were predicted by DSC, GOR4, PHD, SOPM/A, and the consensus sequence as being beta sheets. According to the crystal structure, this region is a short alpha helix flanked by two disordered loops at the beginning of the protein.

The second region, a segment of the long alpha helical region between residues 107-145, seemed to have caused some problems among the various algorithms. The segments spanning residues 107-118 and 121-126, were predicted to be beta sheets by virtually all the algorithms, except DPM, which was able to predict the entire helical region correctly. SOPMA was the next algorithm closest to being correct, having missed a short stretch of 112-118 residues of the helix but correctly predicting the remainder of the helical region. It is interesting to note that both regions that caused problems were confined to the C-terminal beginning and N-terminal end of TAFII-18. These regions may be floppier since they are pointed away from the more ordered and structured internal region of the protein and exposed entirely to solvent.

Case Study 3: all beta proteins

Alcohol dehydrogenase, Class II (Mus musculus)



Alcohol dehydrogenase (ADH) is a dimeric zinc-metalloenzyme that catalyzes the reversible oxidation of alcohols to a cetaldehyde/ketones with the concomitant reduction of NAD. The functional roles of the ADH classes are not fully established, but catalytic activities suggest roles in the metabolism of steroids, retinoids, biogenic amines, lipid peroxidation products, hydroxy fatty acids as well as xenobiotic alcohols and aldehydes. Moreover, zinc-containing ADH's are found in bacteria, mammals, plants, and fungi. The murine class II ADH defines a functionally distinct group of class II alcohol dehydrogenases that exhibits interesting catalytic properties, such as low catalytic efficiency as a consequence of slow hydride transfer and the fact that it is not possible to saturate it with ethanol, while beta-hydroxy fatty acids function as tight inhibitors rather than as substrates.

Moreover, it has been shown to catalyze the reduction of some benzoquinones and benzoquinoneimines. Only three classes of murine ADH's (I, III, IV) have previously been identified, and the extreme evolutionary divergence of class II ADH's have made this group a challenge to characterize. The human form was the first class II ADH to be identified and shown to prefer unsaturated hydrophobic aldehydes in noradrenaline metabolism (Ditlow *et al.*, 1984).

According to the SCOP classification system, the murine class II ADH contains a GroES-like fold with a partly opened beta-barrel, a classical alpha/beta Rossmann-fold C-terminal domain, and a zinc-finger subdomain comprising residues 94-117. Recently, the 2.2 Å resolution crystal structure of murine class II alcohol dehydrogenase was determined in a binary complex with the coenzyme NADH. The ADH2 dimer is asymmetric in the crystal with different orientations of the catalytic domains relative to the coenzyme-binding domains in the two subunits, resulting in a slightly different closure of the active-site cleft. The semi-open conformation and structural differences around the active-site cleft contribute to a substantially different substrate-binding pocket architecture as compared to other classes of alcohol dehydrogenase, and provide the structural basis for recognition and selectivity of alcohols and quinones. The loop with residues 296-301 from the coenzyme-binding domain is short, thus opening up the pocket towards the coenzyme. On the opposite side, the loop with residues 114-121 stretches out over the inter-domain cleft. A cavity is formed below this loop and adds an appendix to the substrate-binding pocket. Asp301 is positioned at the entrance of the pocket and may control the binding of omega-hydroxy fatty acids, which act as inhibitors rather than substrates.

Table 4: Pairwise Comparison of Secondary Structure Prediction Algorithms for ADH

	DPM	DSC	GOR4	PHD	Predator	SOPM	SOPMA	Sec.Cons.	DSSP
DPM	100%	62%	70%	57%	67%	67%	59%	72%	52%
DSC	62%	100%	71%	77%	63%	67%	80%	77%	62%
GOR4	70%	71%	100%	65%	74%	73%	68%	82%	58%
PHD	57%	77%	65%	100%	62%	67%	81%	77%	73%
Predator	67%	63%	74%	62%	100%	67%	63%	75%	61%
SOPM	67%	67%	73%	67%	67%	100%	76%	81%	61%
SOPMA	59%	80%	68%	81%	63%	76%	100%	79%	68%
Sec.Cons.	72%	77%	82%	77%	75%	81%	79%	100%	68%
DSSP	52%	62%	58%	73%	61%	61%	68%	68%	100%

The accuracy of the algorithms declined in their attempt to predict the structure of the predominantly beta-sheet protein, murine class II ADH. In this case, the consensus sequence (68%) rivals the success rate of the PHD algorithm (73%) in being the most accurate in predicting the secondary structure of ADH, which has typically not been the case in other situations. In contrast to the success of predicting the secondary structure of TAFII-18, the algorithms are generally not as keen in detecting beta-strands.

The first beta sheet appears between residues 9-15, but DPM, SOPM, and SOPMA are mildly confused in thinking that there is an alpha helix wedged between two beta strands. DSC and GOR IV are accurate, but their predictions are a bit longer than the observed beta strand. PHD and Predator predict that the beta strand is followed immediately by an alpha helix. Interestingly, the consensus prediction decides to interpret the muddle as a loop.

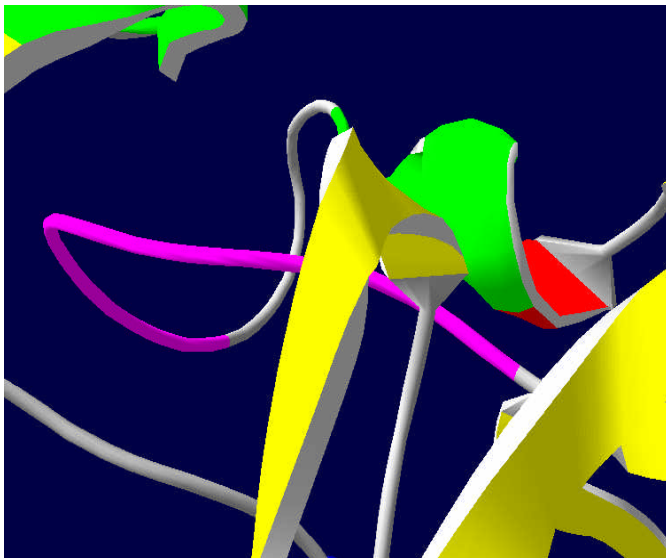


Figure 11: ADH, residues 60-66.

The loop region between residues 60-66 was predicted by all the algorithms to form a beta-strand structure. The same mistake happens again around residue 80-84, but this time, only a minority of the algorithms (DSC, PHD, and SOPM/A) interpret the region to be a beta-strand.

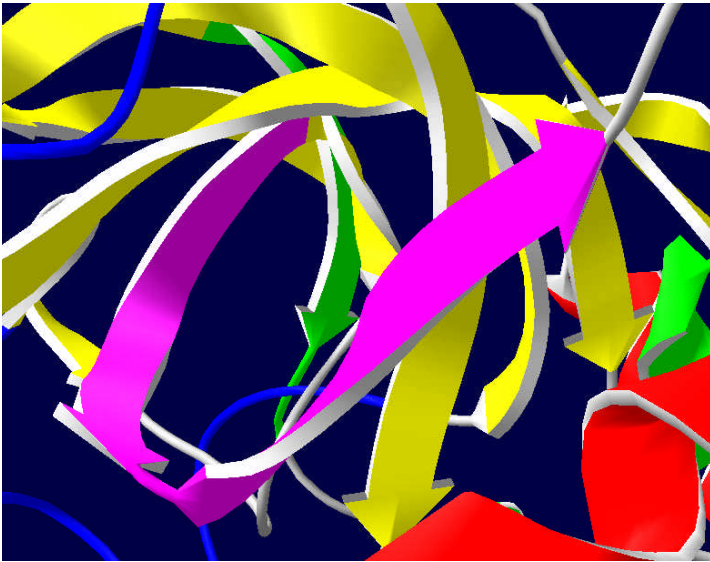


Figure 12: ADH, residues 153-164.

The two short beta-strands, separated by a loop, comprising residues 153-164, caused the algorithms to predict the presence of helices, especially GOR 4. Only DSC, PHD, and SOPMA are exempt in that they are able to predict the second short beta strand. The other programs believe the beta strand is followed immediately by an alpha helix.

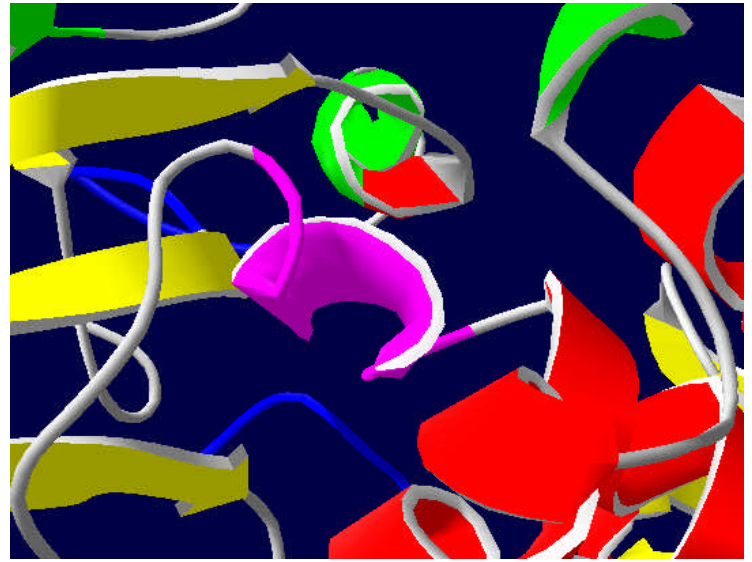


Figure 13: ADH, residues 173-178.

Reminiscent of the residues between 49-53 and 101-104, the short helical fragment between residues 173-178 is predicted to be a beta-sheet by almost every algorithm, except Predator. Predator correctly predicts a short alpha helical region followed by a loop. The crystal structure identifies the helical fragment as being buried deep inside the protein, following a beta strand/loop, and preceding another cluster of alpha helices. The hydrogen bonding interactions must predominate to generate the alpha-helicity of the residues in this area.

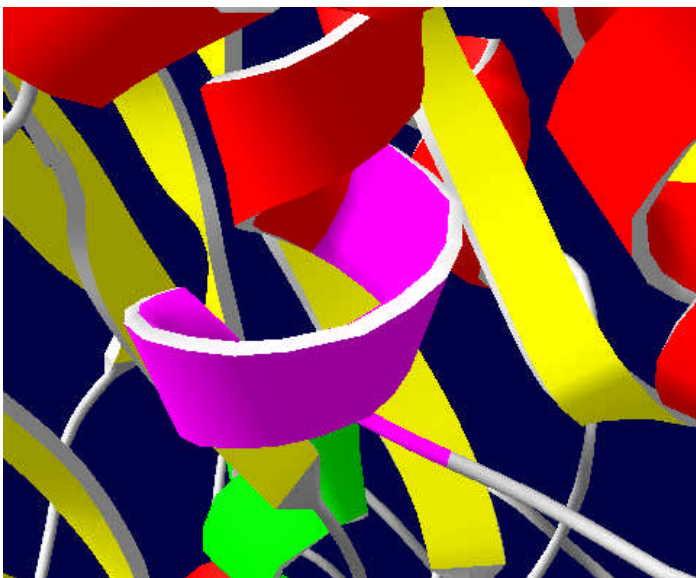


Figure 14: ADH, residues 187-192.

The termination of an alpha helix around residues 187-192 causes a muddled list of predictions, ranging from BLLB, to all BBBB, to ABBA, and no consensus among any of the algorithms. Only PHD predicts the correct secondary structure. The confusion may be a consequence of N-capping, as the last two residues, the asparagine and threonine, are often predicted to form N-caps in alpha helical regions.

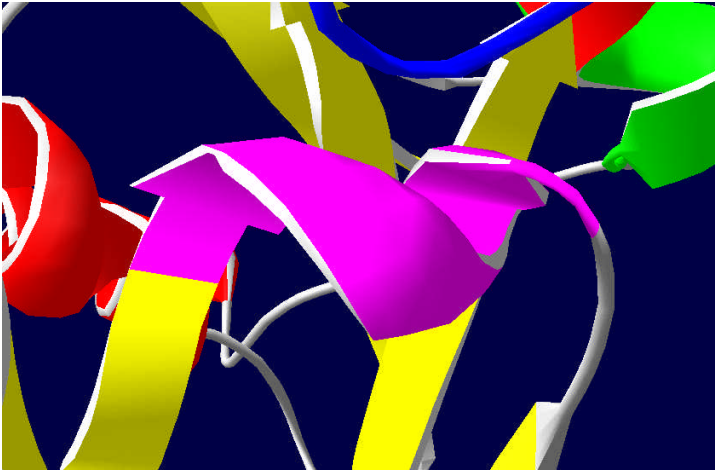


Figure 15: ADH, residues 308-312.

Another completely helical region that immediately follows a beta strand is interpreted to be a beta strand by all the secondary structure prediction algorithms. In this case, the crystal structure reveals why this is the case. Usually a loop separates strands and helices, but in this case, the two structures almost seem to melt into each other—the border between being a helix or a strand is so nebulous that it is very easy to see how the programs can be confused between the two structures.

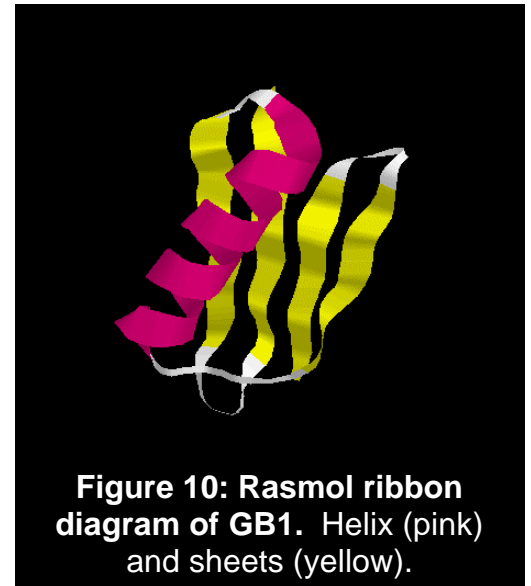
Case Study 4: an engineered alpha+beta protein

Chameleon (IgG-binding protein1, Streptococcus)

The “chameleon” sequence is an artificially engineered 11-amino acid sequence that folds as an α -helix (chameleon- α) when in one position but as a β -sheet (chameleon- β) when in another position of the primary sequence of the Immunoglobulin G-binding domain of protein G (GB1). GB1 is comprised primarily of antiparallel β -sheets with segregated α -helix and β -sheet regions. The 11-amino acid sequence corresponding to AWTVEKAFKTF was designed to replace the α -helix residues 23-33 and β -sheet residues 42-52 of GB1 in such a way that would still preserve the hydrophobic nature of the residues that constitute the interface between each of the secondary structure elements and the core of the protein GB1. Peter Kim identified three types of environments in comparing residues 23-33 and 42-52 of GB1:

- Class I, sites where a residue was buried in one secondary structure but exposed to the other
- Class II, sites where a residue occupied a buried position in both positions but was very different in size or polarity in each structure
- Class III, sites that had no conflict in size, polarity, or burial

By itself, the chameleon sequence is unfolded, having no strong preference for either the α -helix or β -sheet conformation, as determined by both circular dichroism and NMR. Yet, when embedded in different local environments, the chameleon sequence adopts one secondary structure preferentially over the other. Thus, the secondary structure formed by the chameleon sequence is specified by tertiary, non-local interactions, and underscore the importance of environment-dependent effects in protein folding.



SEQUENCES		
GB1:	23	33
TTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTEK		
Chameleon α :	23	33
TTYKLILNGKTLKGETTTEAVDAWTVEKAFKTFANDNGVDGEWTYDDATKTFTVTEK		
Chameleon β :	42	52
TTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGAWTVEKAFKTFVTEK		

Table 5: Pairwise Comparison of Secondary Structure Prediction Algorithms for GB1, Chameleon-alpha, and Chameleon-beta

	DPM	DSC	GOR4	PHD	Predator	SOPM	SOPMA	Sec.Cons.
GB1	54%	64%	64%	73%	82%	64%	64%	63%
Chm- α	53%	69%	66%	74%	81%	62%	66%	64%
Chm- β	54%	66%	57%	70%	73%	55%	61%	61%

Overall, Predator was the most successful (82%) algorithm in correctly predicting the secondary structure governing the 57 amino acid long GB1 sequence. PHD came in second with a success rate of 73%. The consensus sequence was hardly any better than any of the individual algorithms alone. Interestingly, the same trend held throughout the predictions for both chameleon sequences, although the accuracy of Predator dropped dramatically to 73% in the attempt to predict the secondary structure of chameleon-beta. DPM generally faired very poorly in predicting the secondary structure of any of the proteins. Possibly class prediction was difficult due to the short length of the query sequence and the protein being unique to the 135 proteins which define the structure classes. Else, the Chou-Fasman is inadequate in describing the non-local interactions that often govern secondary structures.

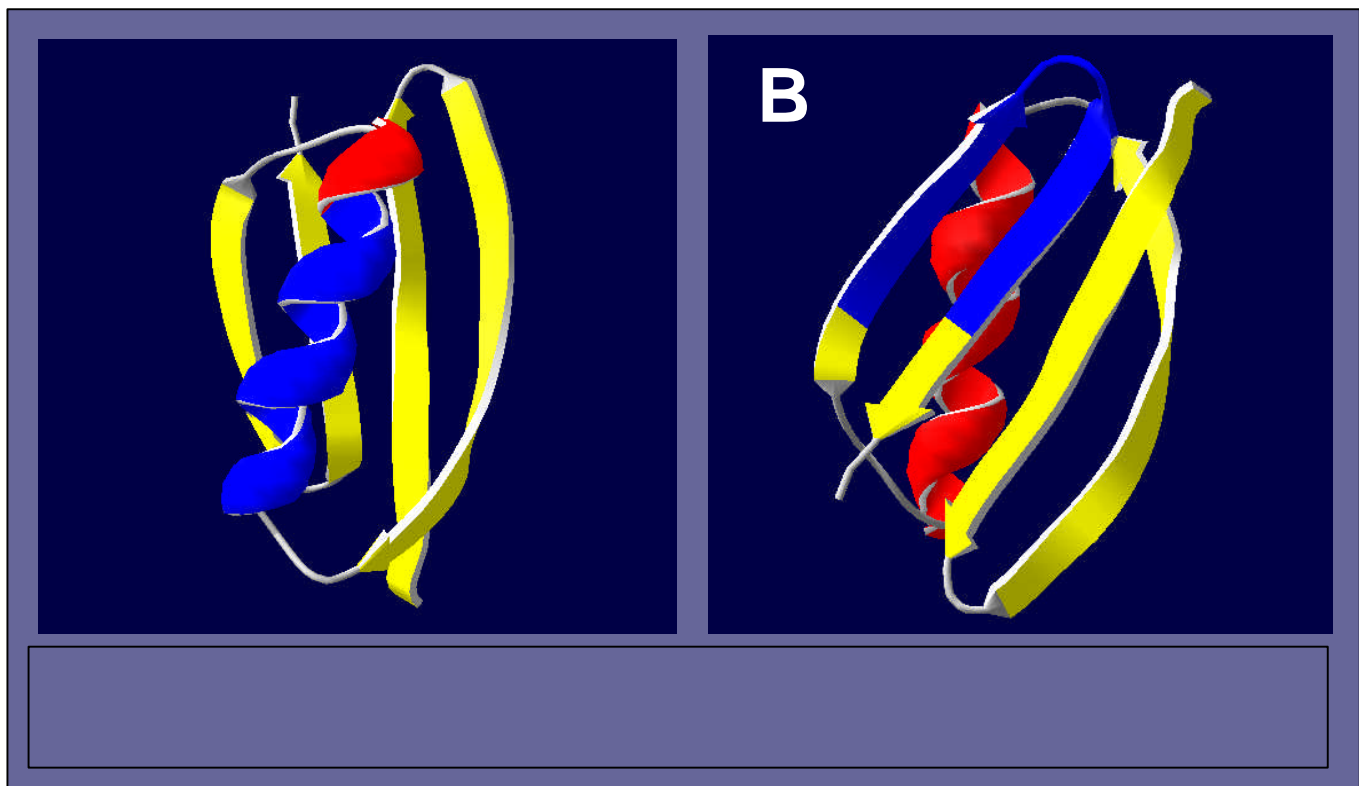


Figure 12: Secondary Structure Prediction alignment of GB1 against DSSP

	10	20	30	40	50
GB1	TTYKLILNGKTLKGETTTEAVD	AATAEKVFKQY	ANDNGVDGEW	TYDDATKTFTVTEK	
DPM	LLBBBBBLLLLLLLLLLLLL	AAAAA	AAAAA	BAAAA	LLLLLLLLLLLLLLLLL
DSC	LLBBBBBLLLLLLLLLLLLL	BBBBB	AAAAA	AAAAA	BBBBB
GOR4	LLBBBBBLLLLLLLLLLLLL	AAAAA	AAAAA	AAAAA	BBBBB
PHD	LBBBBBBLLLBLLBBBBB	AAAAA	AAAAA	AAAAA	BBBBB
Predator	LBBBBBBLLLBLLBBBBB	AAAAA	AAAAA	AAAAA	BBBBB
SOPM	BBBBBBBLLLLLLLLLLLLL	AAAAA	AAAAA	AAAAA	BBBBB
SOPMA	LBBBBBBLLLLLLLLLLLL	AAAAA	AAAAA	AAAAA	BBBBB
Consensus	LLBBBBBLLLLLLLLLLLLL	AAAAA	AAAAA	AAAAA	BBBBB
DSSP-GB1	LLBBBBBBLLLBLLBBBBB	LLLLL	AAAAA	AAAAA	BBBBB

- **Confusion between alpha/beta/loop:**

Predator predicted the correct beta strand between residues 14-20, but the other programs, namely DPM, GOR4, SOPM/A, predicted overlapping loop and helical regions spanning the beta-strand. DSC and PHD predicted some beta-sheet

- **Unable to predict N-terminal cap of the alpha helix:**

With the exception of PREDATOR, all the other algorithms had difficulty predicting the last four amino acids of the alpha helix included in residues 33-36, consisting of two amino acids typically found in N-capping structures: asparagine and aspartic acid. Rather, they terminated prematurely with a prediction for loops because they were unable to detect that the Ncap.

- **DPM has problems detecting secondary structure:**

For the region between 23-33, all the algorithms were consistently able to predict the alpha-helical nature of this region with the natural amino acid sequence in place. Similarly, the beta-strand/turn/beta-strand region was predicted by virtually all the algorithms with the exception of DPM, which missed the beta strand region between residues 40-46, and also mistook the loop/turn region for an alpha-helical structure.

Figure 13: Secondary Structure Prediction alignment of Chm-alpha against GB1-DSSP

	10	20	30	40	50
Chm-alpha	TTYKLILNGKTLKGETTTEAVD	AWTVEKAFKTF	ANDNGVDGEW	TYDDATKTFTVTEK	
DPM	LLBBBBBLLLLLLLLLLLLL	BAAAA	AAAAA	AAAAA	LLLLLLLLLLLLLLLLL
DSC	LLBBBBBLLLLLLLLLLLLL	BBBBB	AAAAA	AAAAA	BBBBB
GOR4	LLBBBBBLLLLLLLLLLLLL	AAAAA	AAAAA	AAAAA	BBBBB
PHD	LBBBBBBLLLBLLBBBBB	AAAAA	AAAAA	AAAAA	BBBBB
Predator	LBBBBBBLLLBLLBBBBB	AAAAA	AAAAA	AAAAA	BBBBB
SOPM	BBBBBBBLLLLLLLLLLLLL	AAAAA	AAAAA	AAAAA	BBBBB
SOPMA	LBBBBBBLLLLLLLLLLLL	AAAAA	AAAAA	AAAAA	BBBBB
Consensus	LLBBBBBLLLLLLLLLLLLL	AAAAA	AAAAA	AAAAA	BBBBB
DSSP-GB1	LLBBBBBBLLLBLLBBBBB	LLLLL	AAAAA	AAAAA	BBBBB

The conversion of the sequence from AATAEKVFKQY → AWTVEKAFKTF (chm sequence) using site-directed mutagenesis should preserve the alpha helical structure of this region. Nuclear overhauser (NOE) spectra and nuclear magnetic resonance (NMR) have confirmed this experimentally. Surprisingly, all the algorithms, with the exception of DPM, were capable of predicting the correct secondary structure for this altered region. In fact, the overall accuracy in alignments slightly improved for each algorithm. In contrast, DPM predicted a muddled string of predictions, confusing alpha helices with beta-strands.

Figure 14: Secondary Structure Prediction alignment of Chm-alpha against GB1-DSSP

	10	20	30	40	50	
Chm-beta	TTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDG	AWTVEKAFKTF	TVTEK			
DPM	LLBBBBLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL	AAAAAAAAAAAA	AAAAAAAAAAAA	LLLLLLLL	ABABAAAAABBBBLL	
DSC	LLBBBBLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL	BBBBAAAAAA	AAAAAAAAAAAA	LLLLLLLL	BBBBLLLLLBBBBLL	
GOR4	LLBBBBLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL	AAAAAAAAAAAA	AAAAAAAAAAAA	LLLLLLLL	AAAAAALLBBBBBL	
PHD	LBBBBBLLLLLLLLBBBBL	AAAAAAAAAAAA	AAAAAAAAAAAA	LLLLLLLL	LBBLLLLLLBBBBBL	
Predator	LBBBBBLLLLLBBBBBBL	LLLLAAAAAA	AAAAAAAAAAAA	LLLLLLLL	LLAAAAALBBBBBL	
SOPM	BBBBBBLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL	AAAAAAAAAAAA	AAAAAAAAAAAA	LLLLLLLL	BBBBAAAAABBBLL	
SOPMA	LBBBBBLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL	AAAAAAAAAAAA	AAAAAAAAAAAA	LLLLLLLL	BBBBBLLLABBBBL	
Consensus	LLBBBBL	LLLLLLLL	LLLLAAAAAA	AAAAAAAAAAAA	LLLLLLLL	LBBBAAALBBBBBL
DSSP-GB1	LLBBBBBBL	LLLLBBBBBBL	LLLLAAAAAA	AAAAAAAAAAAA	LLLLL	LBBBBBLLLLLBBBBBL

Transposing the chameleon sequence to residues 42-52 preserves the beta-strand/turn/beta-strand nature of this region, but causes substantial confusion among the various secondary structure prediction algorithms. DPM, GOR4, and Predator incorrectly predict the chameleon sequence to be predominantly alpha helical, suggesting that they are unable to factor in the amino acid environment surrounding the chameleon sequence into its prediction calculations. It is possible that the non-local interactions important for this structural prediction are not governed by hydrogen-bonding, since Predator, an algorithm that is based on the recognition of hydrogen-bonding propensities, was unable to offer the correct prediction. SOPM is also incorrect in predicting a beta/alpha/beta region. Only DSC and PHD, which both incorporate multiple sequence alignment data, are capable of correctly predicting the beta-strand/loop/beta-strand motif. SOPMA is close to being correct, but confuses one of the loop forming residues to be an alpha-helical residue. In this situation, the consensus sequence obtained from averaging all the programs is complete nonsense: **LBBBAAALBB**.

CONCLUSIONS

Are there differences in the ability of secondary structure prediction algorithms to detect alpha-helices vs. beta-strands? If so, what differences? And why?

As evidenced from the results of the four case studies, it is apparent that secondary structure prediction-based modeling is riddled with errors in attempting to predict helices and strands. Oftentimes, the confusion seems to arise due to the sole use of local information in regions stabilized by long-range, non-local interactions. Upon comparing the ability of the algorithms to predict the secondary structure of an all alpha-helical protein (TAFII-18) with one that was predominantly beta-sheet (ADH), we can see that beta-strands are generally predicted much more poorly than helix or coil residues. More distant interactions may account for the observation that beta strands are predicted more poorly by analysis of local regions. Another source of confusion may result in a situation where a region may have a higher preference for forming a helix than a strand (and vice versa), but interactions non-local in sequence may result in that the formation of a β -sheet (α -helix) is energetically more favorable. Indeed, the confusion between helices and strands can often be attributed to hydrogen bonds stabilized by non-local inter-residue contacts.

More specifically, a major shortcoming of the PHD neural network algorithm is that it often predicts helices that are too long and misses short, fragmented helices in the process. Predator and NNSSP have also exhibited this behavior. PHD has, however, been able to improve its accuracy through evolutionary information from multiple sequence alignments and using a multilevel system for secondary structure prediction. On the other hand, nearest neighbor algorithms like NNSSP has a slightly better overall accuracy than PHD and better predicts α -helices, while PHD more accurately predicts beta-sheets. The larger average segment length of helices predicted by NNSSP is a result of rigorous exclusion of the predicted helices with lengths below 5 through filtering methods. Class prediction programs, such as DPM, fair poorly in distinguishing alpha helices from beta sheets when the protein cannot be easily sorted into one of the four secondary structure classes. Also, it is much easier to predict structure class starting from the detailed information about evolutionary profiles for the entire sequence than by restricting the input to just merely amino acid composition of the protein.

Are the algorithms capable of predicting the secondary structure of an engineered protein?

Since all of the secondary structure prediction algorithms are trained on naturally evolved proteins, it would be interesting to see if these algorithms can be applied to engineered sequences. In the native sequence of the IgG binding domain of protein G (GB1), the deca-peptide AATAEKVFKQY (chameleon alpha: residues 23-33) is embedded in an alpha-helix, whereas the deca-peptide EWTYDDATKTF (chameleon beta: residues 42-52) forms a beta-strand. When replacing both naturally occurring deca-peptides by the engineered deca-peptide AWTVEKAFKTF, the natural structures are maintained in such a way that the deca-peptide AWTVEKAFKTF switches from helical to strand conformation.

Remarkably, DSC and PHD, two algorithms which both incorporate information from multiple sequence alignments, were capable of predicting the correct structure for the engineered peptide. In general, such prediction methods may not be valid if applied to engineered proteins. Interestingly, for the particular case of chameleon, both DSC and PHD was successful. When basing the prediction on a multiple alignment (rather than on single sequences) the peptide AWTVEKAFKTF was correctly predicted in both conformations.

Are tertiary interactions critical for accurate secondary structure prediction? If so, how?

Firstly, there is experimental evidence suggesting that non-local interactions within the primary amino acid chain may influence local secondary structure. For instance, tertiary interactions such as hydrogen bonding play a dominant role in determining whether an amino acid will form α -sheets (Kim *et al.*, 1994). Moreover, it has been shown that the same amino acid sequence up to five (Kabsch *et al.*, 1983) and eight (Sudarsanam, 1998) residues in length can be found in different secondary structures in the proteins in the structural database.

Secondly, as demonstrated by the frequent confusion of alpha helices and beta sheets, and the fact that the chameleon sequence generally stumped most of the secondary structure prediction algorithms, it is clear that tertiary interactions are important, although not absolutely necessary, for accurate predictions of secondary structure. For instance, because Predator takes into account hydrogen-bonding propensities of amino acids, it is much better equipped to identify differences between alpha helices, parallel or anti-parallel beta sheets, and loop structures. Its success in predicting the secondary structure of TAFII-18 can be attributed to that ability.

What affect does helical capping have on secondary structure prediction?

C4-C3-C2-C1-Ccap-C''.....N''-N'-Ncap-N1-N2-N3-N4....

On average, we can deduce from the case studies that virtually all the secondary structure prediction programs predict the core of helices and strands more accurately than the caps, in particular the N- and C-terminal residues in α -helices. The helical capping problem arises because the first and last turns of α -helices contain residues that cannot form one of the backbone hydrogen bonds that hold the helix together. This problem is often addressed by the formation of “capping” hydrogen bonds from residues in the adjacent loops at both the C-terminus and N-terminus of an α -helix. In C-terminal capping, the backbone amide of the C'' residue forms a capping hydrogen bond to the backbone carbonyl of the C3 residue, and the amide of C' forms a second hydrogen bond with the carbonyl of C2. In addition, there is a hydrophobic interaction between the sidechain of C'' and that of C3, C2, or C4. Hydrophobic residues are required at C'' and at C3, C2 or C4, and a lysine or arginine residue is allowable at C'' because of the substantial hydrophobic character of their alkyl chains. “N-capping” occurs when the side chain and backbone amides of the Ncap and N3 residues make reciprocal hydrogen bonds. Residues that can participate in this motif are threonine, serine, asparagine, aspartic acid, glutamine, and glutamic acid. They help increase the stability of both proteins and helical peptides. Both of these motifs allow sharp turns in the direction of the peptide chain and are thought to prevent “fraying” of the helix ends.

Are buried helices predicted at lower accuracy than exposed helices?

None of the four case studies point to any differences in the ability of secondary structure prediction programs to predict the structures of buried or exposed helices. It has been argued in literature that buried helices would generally be predicted less accurately than exposed helices. Upon looking at the success rates among the various algorithms, we can conclude that the per-residue accuracy actually appears to be higher for buried residues than for residues in exposed, solvent-accessible helices. In fact, for TAFII-18, the exposed helices resulted in much greater confusion among the algorithms than the other buried helices.

Is the use of multiple sequence alignments in secondary structure prediction a great advantage?

Prediction from a multiple alignment of protein sequences rather than a single sequence has long been recognized as a way to improve prediction accuracy. Prediction methods that use multiple sequence alignments gain accuracy over single-sequence methods by exploiting the patterns of residue conservation that are seen in protein families. During evolution, residues with similar chemical properties are conserved if they are important to the fold or function of the protein. This makes patterns of hydrophobic residues characteristic of particular secondary structures easier to identify. Inclusion of more distantly related sequences in the alignment should improve the clarity of such patterns, but in an automated alignment building procedure, the risk is that unrelated protein sequences will pollute the alignment.

Generally, the more sequences in the alignment, the better. However, it is also important to realize that bias may arise from having highly redundant or identical sequences present in the alignment. Instead, you should leave out some family members in the high homology (>70%) region, in particular, when there are not many rather diverged sequences present.

Aligned sequences should have considerable variation with respect to the guide sequence. Ideally, it should contain sequences at least at a level of 30% pairwise sequence identity (with respect to the predicted

protein). In general, more diverged sequences contribute more to the information content than do very similar ones (> 80%). More distantly related sequences contribute more to the alignment diversity which is the base for an improved prediction accuracy. However, the more distant relative are difficult to align, causing alignment errors, which may adversely affect secondary structure prediction.

Does prediction accuracy increase by combining the results from multiple programs into a consensus sequence?

When predicting the secondary structure of a protein “blind”, without knowledge of the answer, it is useful to exploit the features of all available prediction algorithms rather than rely on one.

Averaging over many methods helps, on average, but it is possible to lose accuracy as well. For instance, generally the consensus sequence was not significantly better than most of the individual algorithms. If one of the programs consistently failed badly, such as Chou-Fasman, that would drag the average down, even though the other algorithms may be very successful and correct. Most often some methods are more accurate than the average, such as neural network algorithms. Furthermore, there are examples of proteins predicted poorly by all methods, such as predominantly beta-sheet proteins, for which all methods agree by mistake. Thus, trying to use many methods may not provide the answer to the question whether the prediction for your protein is more likely to be below or above average.

Moreover, combining programs may help improve on non-systematic errors. Any prediction method has two sources of errors: (1) systematic errors, such as through non-local effects, and (2) white noise errors caused by, for instance, the succession of the examples during training neural networks. Theoretically, combining any number of methods improves accuracy as long as the errors of the individual methods are mutually independent and are not only systematic

REFERENCES

[SIRT2]

Finnin MS, Donigian JR, Pavletich NP. (2001). Structure of the histone deacetylase SIRT2. *Nat Struct Biol*, 8, 621-5.

Min J, Landry J, Sternglanz R, Xu RM. (2001). Crystal structure of a SIR2 homolog-NAD complex. *Cell*, 105, 269-79.

Guarente L. (2000). Sir2 links chromatin silencing, metabolism, and aging. *Genes Dev*, 14, 1021-6.

[Chameleon]

Minor, D. L. J. & Kim, P. S. (1996). Context-dependent secondary structure formation of a designed protein sequence. *Nature*, 380, 730-734.

Rost WWW, B. (1996). 1D structure prediction for Chameleon (IgG binding domain of protein G). EMBL Heidelberg, Germany, WWW document (<http://www.embl-heidelberg.de/~rost/Res/96C-PredChameleon.html>) .

Dalal, S., Balasubramanian, S. & Regan, L. (1997). Protein alchemy: changing b-sheet into a-helix. *Nature Struct. Biol.*, 4, 548-552.

Kabsch W. and Sander C. (1983). On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations. *PNAS*, 81, 1075-1078.

Sudarsanam S. (1998). Structural diversity of sequentially identical subsequences of proteins: Identical octapeptides can have different conformations. *Proteins*, 30, 228-231.

[TAFII18]

Birck C, Poch O, Romier C, Ruff M, Mengus G, Lavigne AC, Davidson I, Moras D. (1998). Human TAF(II)28 and TAF(II)18 interact through a histone fold encoded by atypical evolutionary conserved motifs also found in the SPT3 family. *Cell*, 94, 239-49.

Goodrich, J.A., Cutler, G., and Tjian, R. (1996). Contacts in context: promoter specificity and macromolecular interactions in transcription. *Cell* 84, 825-830

[ADH]

Svensson S, Hoog JO, Schneider G, Sandalova T. (2000). Crystal structures of mouse class II alcohol dehydrogenase reveal determinants of substrate specificity and catalytic efficiency. *J Mol Biol*, 302, 441-53.

Ditlow CC, Holmquist B, Morelock MM, Vallee BL. (1984). Physical and enzymatic properties of a class II alcohol dehydrogenase isozyme of human liver: pi-ADH. *Biochemistry*, 23, 6363-8.

[DPM]

Deleage G, Roux B. (1987) DPM: An algorithm for protein secondary structure prediction based on class prediction. *Protein Eng*, 1, 289-294.

[Class prediction for DPM based off of the algorithm designed by]

Nakashima et al. (1986) *J. Biochem. Tokyo*, 99, 153-162.

[DSC]

King RD, Sternberg MJ. (1996) DSC: Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci*, 11, 2298-310.

[DSSP]

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22, 2577-2637.

[GOR]

Garnier, J., Gibrat, J.-F., & Robson, B. (1996). GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol*, 266, 540-553.

Garnier J, Gibrat J-F, Robson B. (1996). GOR secondary structure prediction method version IV. *Methods in Enzymology* R.F. Doolittle Ed., vol 266, 540-553.

[PHD]

B Rost. PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods in Enzymology*, 266, 525-539, 1996.

Rost, B., & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, 232, 584-599.

Rost, B., & Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins Struct. Funct. Genet.*, 19, 55-72.

Rost, B., Sander, C., & Schneider, R. (1994). Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.*, 235, 13-26.

Rost B, Sander C. (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, 19, 55-72.

[PREDATOR]

Frishman D, Argos P. PREDATOR: (1996) Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Eng*, 9, 133-142.

[SOPM]

Geourjon C, Deleage G. (1994) SOPM: a self-optimized method for protein secondary structure prediction. *Protein Eng*, 7, 157-164

[SOPMA]

Geourjon C, Deleage G. (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput Appl Biosci*, 11, 681-684.

[GENERAL]

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman., D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.

Jones, D. T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* 292: 195-202.

Mount, DW. (2001). *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press.

Salamov, A. A., & Solovyev, V. V. (1995). Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *J. Mol. Biol.*, 247, 11-15.

Salamov, A. A., & Solovyev, V. V. (1997). Protein secondary structure prediction using local alignments. *J. Mol. Biol.*, 268, 31-36.