

Digital Signal Processing in Protein Secondary Structure Prediction

Debasis Mitra
dmitra@cs.fit.edu

Michael Smith
msmith@cs.fit.edu

Florida Institute of Technology
Department of Computer Science
150 West University Boulevard
Melbourne, Florida 32901

Keywords: *protein secondary structure prediction, homology modeling, digital signal processing techniques*

Abstract

Considerable research effort has been devoted to predicting the secondary structure of proteins from their amino acid sequences. Despite the plethora of prediction techniques, present methods typically have 76% approximate level of accuracy on an average. Thus, there is a considerable room for improvement. We present here a novel automated approach for the secondary structure prediction based on the Digital Signal Processing (DSP) techniques. DSP is an engineering discipline concerning the creation, manipulation and analysis of digital signals. Our technique involves two DSP operators, Convolution and Deconvolution, for the purpose of predicting secondary structures. We use some mappings between an amino acid sequences and the corresponding numerical time-series or “signals” that are processed. Convolution is a method of applying a *filter* on an *incoming signal*, producing an *outgoing signal*. Deconvolution is the inverse operation of convolution and permits the filter to be recovered if the outgoing signal and the incoming signal are known.

Our method predicts three states (helix, strand, and coil) for the secondary structure. We presume that each protein has a corresponding *filter*, which when convolved with the incoming signal (mapped from the primary structure) produces the outgoing signal (mapped from the secondary structure). It is our contention that the unknown secondary structure of a *target* protein can be predicted by using the appropriate digital *filter* for a *base* protein of a significant amino acid sequence-similarity and whose secondary structure is known. Our work presented here attempts to corroborate that contention and experiments with some methodologies towards that direction.

1. Introduction

Proteins are macromolecules that are responsible for a wide range of vital biochemical functions, which include acting as catalysts, oxygen transport, cell signaling, antibody production, nutrient transport and building up muscle fibers. More specifically, proteins are chains of amino acids, of which there are twenty different types, joined by peptide bonds. Proteins have a three-tiered structural hierarchy, typically referred to as primary, secondary and tertiary structure. [Brandon and Tooze 1999] [Rost 1998] Being able to determine the structures of proteins is of tremendous value to the biological community. This is because the higher-level structures determine the function of the protein and consequently, the knowledge of the structure provides insight into its function.

Databases of primary structures, for example SWISS-PROT [Boeckmann et al. 2003], are expanding tremendously, largely due in part to genome sequencing projects [Rost 1998]. However, databases of higher-level structures, for example the Protein Data Bank, are not expanding at such a rapid rate due to the inherent difficulties in determining these levels of structures. Experimental methods for structure determination-procedures can be expensive, very time consuming, labor intensive and may not be applicable to all proteins. [Brandon and Tooze 1999] [Rost 1998] Spurred by the importance of determining protein structure and the shortcomings of the laboratory approaches, significant research has been and continues to be devoted to the prediction of the higher-level structures via computational methods. From the pioneering works of Anfinsen [Anfinsen 1973], it is known that the higher-level structures of proteins are primarily determined by their amino acid sequences. Commonly referred to as the *Protein Folding Problem*, the ability to predict higher-level structures from the sequence remains one of the greatest challenges in bioinformatics. [Bourne and Weissig 2003]

Protein *secondary* structure describes the topology of the chain whereas the *tertiary* structure describes the three-dimensional arrangement of the amino acid residues in the chain. They influence each other, but could be predicted independent of each other. Also, both the secondary as well as the tertiary structures affect the functionality of the protein. For instance, secondary structure aids in the identification of membrane proteins, location of binding sites and identification of homologous proteins, to list a few of the benefits, and thus highlighting the importance, of knowing this level of structure [Rost 2001] This is the reason why considerable efforts have been devoted in predicting the secondary structure only. Knowing the secondary structure of a protein is extremely important and can also greatly enhance the accuracy of tertiary structure prediction [Brandon and Tooze 1999] [Rost 1998] [Rost 2001]. Furthermore, proteins can be classified according to their secondary structural elements, specifically their alpha helix and beta sheet content.

In this paper we have presented a novel method for predicting secondary structures by deploying the digital signal processing (DSP) technique. In section 2 we describe some related works in the secondary structure prediction. Section three introduces the DSP operators used for our purpose. The following two sections develop our main methodology and describe the preliminary experiments. Some enhanced techniques are further proposed in the section 6. The last section concludes the paper with future directions.

2. Related Works in Secondary Structure Prediction

A simple goal in the secondary structure prediction is to predict whether an amino acid residue of a protein is in a helix, strand or in neither of the two, in which case the former is said to be in a coil (or loop). The first generation of secondary structure prediction techniques emerged in the 1960s and were based on single amino acid propensities and, for each amino acid, calculated the probability of it belonging each of the secondary structural elements. The secondary generation of prediction methods extended this concept by taking into account the local environment, of an amino acid, into consideration. Typically, in predicting the secondary structure for a particular amino acid, information gleaned from segments typically comprising of 3-51 adjacent residues were also used in the prediction process. Prediction accuracies with the second generation methods seemed to stall at around 60% accuracy, seemingly because these methods were local in that only information in a window of adjacent residues were used in predicting the secondary structure of an amino acid. [Bourne and Weissig 2003] Local information accounts for approximately 65% of secondary structure information. [Rost 1998] Since the early 1990s, third generation prediction methods achieved prediction accuracies around

70% and such methods incorporate *machine learning* techniques, evolutionary knowledge about proteins and with relatively more complex algorithms. [Pollastri et al. 2002] [Bourne and Weissig 2003] The PHD program [Rost and Sander 1993] [Rost and Sander 1994], which uses a system of neural networks, was the first prediction technique to surpass the 70% threshold. Similar performance was later achieved by other systems which include, JPred2, which combines results from various prediction methods; SAM-T99 [Karplus et al. 1998], which utilizes Hidden Markov Models and a simple neural network with two hidden layers; and SSPro, which uses bidirectional recurrent neural networks [Bourne and Weissig 2003] [Pollastri et al. 2002]

Techniques for secondary structure prediction include, but are not limited to, constraint programming methods [Krippahl and Barahona 1999], statistical approaches [Schmidler et al. 2000] to predict the probability of an amino acid being in one of the secondary structural elements, and Bayesian network models [Baldi et al. 2000]. Nearest neighbor techniques [Alexandrov and Solovyev 1996] attempt to predict the secondary structure of a central residue, within a segment of amino acids, based on the known secondary structures of homologous segments. In [Zhang et al. 1998], a technique based on multiple linear regression was presented to predict secondary structure. Published techniques for secondary structure prediction span over a period of three decades, with the early works of Lim [Lim 1974] and Chou & Fasman [Chou and Fasman 1978] in the 1970s.

To date, neural network techniques have exhibited the highest level of prediction accuracy [Rost 2001][Chandonia and Karplus 1999], and are similar to the nearest neighbor approaches in that the secondary structure targeting a central amino acid within a segment for prediction. Within the neural network-based approaches, various topologies and learning methods have been proposed. Such prediction methods are not used solely in isolation and often techniques are combined, for example, Shavlik and Maclin [Shavlik and Maclin 1993] use the Chou-Fasman algorithm in conjunction with their knowledge based neural network; Ouali and King combine neural networks with rule based statistics; and systems have been developed to combine the results of various neural networks, along with a jury process finally determining the secondary structure [Rost 2001][Chandonia and Karplus 1999][Shavlik and Maclin 1993].

Despite the existence of numerous and varying techniques, there are broadly three main approaches to structure prediction. Homology modeling [Abagyan et al. 1997] [Dunbrack 1999] bases the prediction for an unknown target protein, on the known secondary structures of proteins of similar amino acid sequence. *Threading* [Mirny and Shakhnovich 1998] [Xu et al. 1999] [Thiele et al. 1999], or the inverse folding technique, maintains a database of the models of known folds and attempts to ‘fit’ the amino acid sequence of the target protein, to a known model. The basis of *threading* is that a limited number of unique protein folds exist in nature and structure prediction of a target sequence can be performed by consulting a database of known folds and determining which fold-model best fits the sequence. The methodology of such an approach is not to predict the structure from a primary sequence, but rather to fit a known structural-model to a sequence. Typically, steps are taken to align the target sequence to a known set of folds and a scoring function is employed to determine the best fitting structure. Both homology modeling and threading rely on the existence of known structures and the disadvantage of such approaches is that accurate prediction relies on proteins of similar structure already being solved. The third approach, namely the *ab initio* techniques [Xia et al. 2000] [Bonneau et al. 2001], or prediction from first principles, bases structure prediction on known biochemical and biophysical facts related to the proteins. However, progress there has been relatively slow as the physical processes by which a protein folds are not completely understood. [Bourne and Weissig 2003] In general they are also computationally very expensive methods.

3. DSP-based Methodology

Various secondary structure prediction methods, particularly some neural network and nearest neighbor techniques, utilize a localized prediction methodology in the sense that a window, typically of less than 20 amino acids, is presented to the prediction system with the aim of predicting secondary structure of the central element, using only the information gleaned from the amino acids within the window. However, local information accounts for approximately 65% of secondary structure formation [Rost 1998]. Therefore, prediction can potentially be improved by incorporating a more global prediction scheme. It must be mentioned that this ideology has been documented and various prediction methods are adapting a more global view of structure prediction [Rost 2001].

We present a novel approach to secondary structure prediction using the digital signal processing operators *convolution* and *deconvolution*. We hypothesize that such methods can enhance structure prediction due to their inherent global nature. Secondly, we alleviate the need to perform any sort of training, as many of the best automated-methods do. We presume that the existing databases of secondary structures provide sufficient information for the purpose of predicting a protein's unknown secondary structure.

Our method is used in conjunction with the existing technology that locates similar proteins, such as PSI BLAST and FASTA. We also require sequence alignment operation, which is available through services, such as GeneStream [Person et al. 1997]. It is our contention, that secondary structure of a protein can be predicted by locating another appropriate protein of similar amino acid sequence (whose secondary structures are known), and then by deploying digital filtering techniques for transforming the primary sequence to the secondary structure. Thus, our approach can be viewed as a form of homology modeling [Skolnick and Kolinski 2001], in which the structures of closely related proteins, in terms of amino acid sequences, are used to predict unknown secondary structure. The details of the methodology are described in the sections (5 and 6) on the experiments.

4. Digital Signal Processing

Digital Signal Processing (DSP) is an area of science and engineering undergoing rapid development, largely due to the advances in computing and integrated circuits. In general terms, DSP is the mathematics, algorithms, techniques and methodologies employed in analyzing, manipulating and transforming digital signals or time-series [Proakis and Manolakis 1996]. We map a protein into a digital signal by assigning numeric values to each amino acid.

DSP techniques relating to protein structure analysis, such as [Hirakawa and Kuhara 1997], [Irback et al. 1996], [Irback and Sandelin 2000], [Veljkovic et al. 1985] assign numeric values - often their *hydrophobicity* values [Kyte and Doolittle 1982], to the amino acids, and analyze the resulting sequence via *Fourier analysis*, *wavelet processing* or some other DSP techniques. Although often employed in the analysis of protein structures, our literature review indicates an absence of DSP-based approaches in predicting secondary structures.

Our method employs two fundamental DSP operators for predicting secondary structure, namely, the *convolution* and the *deconvolution*. Convolution is a method of combining two signals (typically an *input signal* and a *filter*) to produce a third signal (*output*). Deconvolution is the inverse operation of convolution. Given the output signal and the filter, the input signal can be calculated by deconvolution, and conversely given the input and output signals, the filter can be calculated.

Mathematically, convolution, between signals x and h , is represented by the following formula, often referred to as the convolution sum,

$$y(n) = \sum_{k=-\infty}^{\infty} x(k)h(n-k)$$

Given input signal x and output signal y , the filter h can be calculated via deconvolution,

$$H(z) = \frac{Y(z)}{X(z)},$$

where $H(z)$, $X(z)$ and $Y(z)$ are z -transformations of the impulse response, input signal and output signal respectively. Mathematically, the z -transform of a discrete time signal, $x(n)$, is defined as the power series

$$X(z) \equiv \sum_{n=-\infty}^{\infty} x(n)z^{-n}$$

where z is a complex variable. This equation is often referred to as the direct z -transform as it transforms $x(n)$ into a complex representation. The inverse procedure, of calculating $x(n)$ from $X(z)$, is a detailed mathematical procedure, whose details will be omitted here.

5. Experiment I

We have developed a procedure and conducted a series of experiments to test the viability of the convolution and deconvolution operators in predicting protein secondary structure. The primary and secondary structures of a protein are modeled as input and output signals respectively. The “impulse response” (of the “system” transforming primary structure to secondary) is determined by *deconvolution* and can be considered as a *filter*. The *convolution* of the filter and primary structure of a protein t yields the secondary structure of t .

An important component of our method involves developing appropriate mapping technique between the biological sequences (primary or secondary structures) and numerical time-series. To transform the amino- acid sequence into a discrete time signal, the numeric *hydrophobicity* values (an amino acid residue's attraction to water molecules), calculated by Kyte and Doolittle [Kyte and Doolittle 1982] of each amino acid, were substituted into the protein.

The secondary structures of the proteins are being obtained from the Protein Data Bank (PDB) [Berman et al. 2000]. PDB uses the DSSP classification of Kabsch and Sander [Kabsch and Sander 1983] to represent secondary structure, which defines secondary structural elements as (H) helix, (B) residue in isolated beta bridge, (E) extended beta strand, (G) 3_{10} helix; (I) pi helix, (T) hydrogen bonded turn and (S) bend. Our experiment follows the CASP [Skolnick and Kolinski 2001] classification, which identifies the main secondary structural elements, namely, the *alpha helix*, the *beta strand* and the *coil*. In converting the secondary structure states to three classes, the typical convention was utilized of grouping elements G, H and I into the alpha helix class, elements B and E into beta strands, and other elements are defined as coils. [Bourne and Weissig 2003] For the purpose of mapping from a sequence to its corresponding time-series we arbitrarily use the numerical values of 100, 300 and 500 for *alpha helices*, *beta strands* and *coils* respectively.

Our experiment is based on the hypothesis that the secondary structure prediction of an unknown protein can be based upon the known secondary structure of a protein of similar amino acid sequence. By modeling proteins as discrete-time signals and utilizing the convolution and deconvolution operators, we adopt a rather global view in the secondary structure prediction.

5.1 Experimental Steps

- Select the target protein T, whose secondary structure is to be predicted.
- Perform a PSI BLAST search, using the primary amino acid sequence T_p of the target protein T. The objective is being to locate a set of proteins, $S = \{S_1, S_2, \dots\}$ of similar sequence
- Select from S the primary structure B_p of a base protein, with a significant match to the target protein. A PSI BLAST search produces a measure of similarity between each protein in S and the target protein T. Therefore, B_p can be chosen as the protein with the highest such value
- Obtain the base protein's secondary structure, B_s , from the PDB
- Using B_p , create an input signal I_b (corresponding to the base protein) by replacing each amino acid in the primary structure with its hydrophobicity value. The output signal O_b is created by replacing the secondary structural elements in B_s with the values, 100, 300, 500 for helix, strand and coil respectively
- Solve the system identification problem, by performing deconvolution with the output signal O_b and the input signal I_b to obtain the impulse response, or the sought after filter F
- Transform the amino acid sequence of T_p into a discrete time signal I_t , and convolve with F; thereby producing the predicted secondary structure ($O_t = I_t * F$) of the target protein
- The result of this calculation O_t is a vector of numerical values. For values between 0 and 200, a *helix* is predicted, and between 300 and 500, a *strand* is predicted. All other values will be predicted as a *coil*. This produces mapping for the required target secondary structure T_s of the target protein T

Results

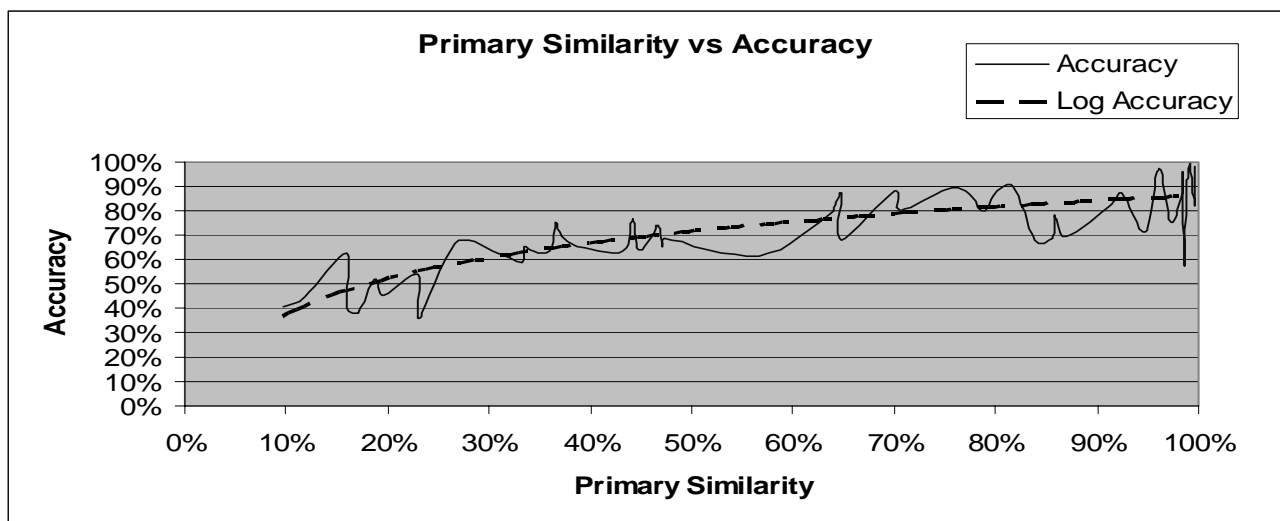


Figure 1 – Results of 58 experiments

Primary similarity - Sequence alignment programs produce a measure of similarity of the aligned primary structures

Accuracy – measures the accuracy of the secondary structure prediction. The predicted secondary structure is compared to the actual one, which is obtained from the PDB. Expressed as a percentage, accuracy is calculated by summing the number of correct individual secondary structure predictions and dividing this value by the length of the sequence.

These results provide indication that the convolution/deconvolution technique offers promise in the prediction of secondary structure. Figure 1 illustrates that the accuracy of prediction tends to be logarithmically increasing with respect to the primary sequence similarity between the target and the base proteins. This is crucial as these results support our contention that by identifying a base protein of significant amino acid sequence similarity to that of the target, our technique can produce accurate predictions. At the lower end of the scale, primary similarity values of 10% tend to produce accuracy of around 40%. However, at primary similarity values of greater than 30%, the prediction accuracy increased to over 65%, which is certainly comparable to the existing prediction methods.

6. Enhanced Methodologies, Experiment II

The preceding set of experimental data indicates that the technique using convolution /deconvolution operators is a viable method in predicting protein secondary structure. We further refine our methodology (1) by investigating the numbering scheme of the secondary structural elements, and (2) by addressing a drawback of the homology modeling in some cases, where the base protein of sufficient similarity to that of the target is currently not available.

In our previous set of experiments, the secondary structural elements of the base protein, have been assigned values of 100, 300 and 500 for helix, strand and coil respectively. We devised an alternate secondary structure-numbering scheme that is Boolean in nature and that separately predicts helices, strands and coils. In order to predict the occurrence of *helices* in the target protein, the secondary structure of the base protein is converted to an output signal by assigning a value of 1 for each occurrence of a helix and 0 for all other structures. The DSP-based prediction procedure as described in the previous section is then executed, and each occurrence of 1 in the resulting output, is predicted to be a *helix*. In a similar fashion, the next phase of the experimentation over the same target protein involves predicting only *strands*. The secondary structure of the base protein is converted to an output signal by assigning a 1 for each occurrence of a *strand* and 0 for the other secondary structural elements. As in the case of helix, the prediction scheme involving the convolution /deconvolution procedure is executed. This time a 1 in the resulting output is predicted to be a *strand*. In our current experimentation, we did not encounter a case in which an amino acid is being predicted to be in both a helix and a strand. However if such collisions arise, they can be resolved by employing a simple heuristic of looking into the predicted structures over a small window of surrounding amino acids (of the one with a conflict) and predicting the secondary structure of the amino acid in question accordingly, based on the dominant predicted structure in the surrounding window. For example, if the surrounding secondary structures are predicted to be mostly helices, then the conflicting amino acid can be predicted to belong to a helix structure. Amino acids that were not predicted to be belonging to either *helices* or *strands* were predicted to be belonging to the *coil* structures.

Secondly, we address the drawback of the homology approach by a proposing *partitioning* method. If the PSI BLAST search does not yield a base protein with at least, say, 30% similarity, to that of the target, a restricted localized prediction methodology could be

employed. In this situation, the target protein is divided into partitions and steps are taken to find base proteins with regions of sufficient similarity to that of the target partitions. The PSI BLAST can be of benefit in partitioning the target protein, because the search results often list such segment regions of similarity. Each segment of the target protein is isolated for the purpose of predicting the corresponding region of similarity in the base protein. The convolution /deconvolution prediction scheme is executed on each segment and the individual results are concatenated to give the prediction for the entire target protein.

For example, in predicting for protein with PDB Id 1PFC (see Table 3): it was divided into two partitions containing amino acid sub-sequences 1 – 60 and 61 – 113. Amino acids 323-384 of PDB Id: 1LGY and amino acids 403-455 of PDB Id: 2IG2 were chosen as the bases for the first and second partitions respectively. The secondary structure of each partition was predicted separately and the results were combined to give the prediction for the whole protein.

Results

Target : 1PFC	Prediction Accuracy	Target: 1PP2	Prediction Accuracy	Target: 1QL8	Prediction Accuracy
Exp 1	64%	Exp 5	82%	Exp 9	88%
Exp 2	63%	Exp 6	92%	Exp 10	96%
Exp 3	68%	Exp 7	80%	Exp 11	83%
Exp 4	70%	Exp 8	89%	Exp 12	92%

Table 3 – Experimental results

Table 3 presents the results of additional experiments to compare the effectiveness of the two secondary structural numbering schemes and the partitioning and non-partitioning methodologies in predicting secondary structure.

Three proteins, an immunoglobulin protein found in guinea pigs (PDB Id: 1PFC), a hydrolase protein found in western diamond back rattlesnakes (PDB Id: 1PP2) and a serine protease protein found in cows (PDB Id: 1QL8) were arbitrarily chosen as the target proteins.

- In the first row of Table 3, for experiments 1, 5 and 9, lists the predictions accuracies when a single base protein is chosen and secondary structural elements are assigned values of 100, 300 and 500 for helix, strand and coil respectively.
- The second row, experiments 2, 6, and 10: lists the prediction accuracies when a single base protein is chosen and the Boolean secondary structure numbering is being employed.
- The third row, experiments 3, 7 and 11: lists the prediction accuracies using the secondary structural scheme of assigning 100, 300 and 500 to helix, strand and coil, but instead of a single base protein being chosen, the target is split into a series of segments. The secondary structure of each segment is predicted and the results are combined to give the prediction for the entire protein.
- The fourth row, experiments 4, 8 and 12: lists the prediction accuracies utilizing the partitioning method of selecting base proteins and Boolean secondary structural numbering scheme.

The relatively high prediction accuracies yield further indication of the viability of using convolution and deconvolution operators in predicting the secondary structures. The results

indicate that a specific approach performs better than the others under certain circumstances. In selecting the secondary structure numbering scheme and deciding whether to use the partitioning method for selecting multiple base protein segments, we conjecture that the class, family, superfamily, and classification or biochemical properties of the target and base protein(s) might be important in determining the exact method to be deployed. For instance, in predicting the secondary structure of 1PFC, the highest prediction accuracy was achieved using partitioning in conjunction with the Boolean methodology. However in predicting 1PP2, the highest prediction accuracy was achieved using the non-partitioning and Boolean approach. Certain properties of the target and base protein(s) may determine which method is most effective in predicting secondary structure. Therefore, instead of relying on a single prediction technique, our results indicate that the best technique to use in a particular situation is dependent upon the properties and characteristics of the target and base protein(s). Of course, when a significantly similar base protein is found in the PDB the simpler method (non-partition plus non-Boolean) may be sufficient.

The following Table 4 compares the highest accuracy achieved with our DSP-based approach with those of some other secondary structure-prediction systems, PHD, SAM-T99 and SS Pro. Our method is comparable with and, in certain cases, surpasses the prediction accuracies of the other techniques.

Comparison with Other Methods

Prediction Method	Prediction Accuracy for 1PFC	Prediction Accuracy for 1PP2	Prediction Accuracy for 1QL8
DSP	92%	70%	96%
PHD	70%	68%	84%
SAM-T99	68%	77%	87%
SS Pro	70%	73%	81%

Table 4 – Comparison with other Secondary Structure Prediction Techniques

7. Conclusion and Future Work

To provide a more thorough analysis of the viability of our proposed technique more experiments will be conducted. Existing secondary structure-prediction methods are, at best, 76% [Rost 2001] accurate in general. Our preliminary results indicate that such a level of accuracy is attainable, and can be potentially surpassed with our method.

The DSP-based approach involving convolution/deconvolution is strictly mathematical. To improve the level of accuracy, we are researching ways of integrating known biological facts in the prediction process. For example, certain amino acids, such as *Ala*, *Glu* and *Leu* are more likely to form a *helix* structure as opposed to a *strand*, and the secondary structural elements like *helices* or *strands* often have their own characteristic sizes. [Brandon and Tooze 1999] These and other facts can be included in the numbering process to improve the accuracy.

Determining a more appropriate scheme for translating amino acids and secondary structural elements to numerical values can result in more representative discrete time signals. In addition to investigating such schemes for translating primary and secondary structures into signals, our current research is also focused on analyzing experimental data to ascertain which prediction technique, e.g., partitioning vs. non-partitioning and Boolean vs. non-Boolean, will yield the highest prediction accuracy given certain properties and/or characteristics of the target

and base protein(s). Also, further improvements and systematization of the partitioning techniques is our near future goal. In this regard, there exists a tempting possibility of combining our approach with the *threading* technique mentioned before.

Finally, we envisage extending the DSP-based method toward prediction of tertiary structures as well. Tertiary structures are described using numbers. This will reduce one aspect of the mapping problem since the “output” signal is already in the number domain.

References

[Abagyan et al. 1997] Abagyan, R., Batalov S., Cardozo,T., Totrov, M., Webber, J., Zhou, Y. 1997. **Homology Modeling With Internal Coordinate Mechanics: Deformation Zone Mapping and Improvements of Models via Conformational Search**. PROTEINS: Structure, Function and Genetics. 1:29-37

[Alexandrov and Solovyev 1996] Alexandrov, N., Solovyev, V., 1996. **Effect of secondary structure prediction on protein fold recognition and database search**. Genome Informatics 7, 119-127

[Anfinsen 1973] Anfinsen, C. B., 1973. **Principles that govern the folding of protein chains**. Science. 181, 223-230.

[Baldi et al. 2000] Baldi, P., Brunak, S., Frasconi, P., Pollastri, G., Soda, G., 2000. **Bidirectional Dynamics for Protein Secondary Structure Prediction**. Sequence Learning: Paradigms, Algorithms and Applications. Springer, 80-104

[Boeckmann et al. 2003] Boeckmann B., Bairoch A., Apweiler R., Blatter M.-C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S., Schneider M., 2003. **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003** Nucleic Acids Res. 31:365-370.

[Bonneau et al. 2001] Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, C., Baker, D. 2001. **Rosetta in CASP4: Progress in Ab Initio Protein Structure Prediction**. PROTEINS: Structure, Function and Genetics. 5:119-126

[Bourne and Weissig 2003] Bourne, Philip E., Weissig, Helge, 2003. **Structural Bioinformatics**. John Wiley & Sons.

[Brandon and Tooze 1999] Brandon C., Tooze J., 1999. **Introduction to Protein Structure**. Garland Publishing.

[Chandonia and Karplus 1999] Chandonia, J., Karplus M., 1999. **New Methods for Accurate Prediction of Protein Secondary Structure**. PROTEINS: Structure, Function and Genetics, 35, 293-306

[Chou and Fasman 1978] Chou, P., Fasman G., 1978. **Prediction of the secondary structure of proteins from their amino acid sequence**. Advanced Enzymology, 47, 45-148

- [Dunbrack 1999] Dunbrack, R.. 1999. **Comparative Modeling of CASP3 Targets Using PSI-BLAST and SCWRL**. *PROTEINS: Structure, Function and Genetics* 3:81-87
- [Hirakawa and Kuhara 1997] Hirakawa, H., Kuhara, S., 1997. **Prediction of Hydrophobic Cores of Proteins Using Wavelet Analysis**. *Genome Informatics*, 8, 61-70
- [Irback and Sandelin 2000] Irback, A., Sandelin, E., 2000 **On Hydrophobicity Correlations in Protein Chains**. *Biophysical Journal*, 79, 2252-2258
- [Irback et al. 1996] Irback, A., Peterson, C., Potthast, F., 1996. **Evidence for nonrandom hydrophobicity structures in protein chains**. *Proc. Natl. Acad. Sci.*, 93, September, 9533-9538
- [Kabsch and Sander 1983] Kabsch W., Sander C., 1983 **Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features**. *Biopolymers*, 3, 2577-2638
- [Karplus et al. 1998] Karplus. K., Barrett, C., Hughey, R. 1998. **Hidden Markov Models for Detecting Remote Protein Homologies**. *Bioinformatics*, vol. 14, no. 10, 846-856
- [Krippahl and Barahona 1999] Krippahl, L., Barahona, P., 1999. **Applying Constraint Programming to Protein Structure Determination**. *Proceedings from Fifth International Conference on Principles and Practice of Constraint Programming*, 289-302
- [Kyte and Doolittle 1982] Kyte, J., Doolittle, R., 1982. **A Simple Method for Displaying the Hydrophobic Character of a Protein**. *Journal of Molecular Biology*, 157, 105-132
- [Lim 1974] Lim, V, 1974. **Algorithms for prediction of α -helical and β -structural regions in globular proteins**. *Journal of Molecular Biology*, 88, 1974, 873-894
- [Mirny and Shakhnovich 1998] Mirny, L., Shakhnovich, E. **Protein Structure Prediction by Threading. Why it Works and Why it Does Not**. *Journal of Molecular Biology* 283: 507, 526
- [Person et al. 1997] Person, W.R., Wood, T., Zhang, Z., and Miller, W., 1997. **Comparison of DNA sequences with protein sequences**, *Genomics* 46: 24-36
- [Pollastri et al. 2002] Pollastri, G., Przybylski, D., Rost, B., Baldi, P., 2002. **Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks**. *Protein: Structure, Function and Genetics*. 47:228-235
- [Proakis and Manolakis 1996] Proakis, J., Manolakis, D., 1996. **Digital Signal Processing, Principles, Algorithms, and Applications**. Prentice-Hall
- [Rost and Sander 1993] Rost, B., Sander, C., 1993. **Prediction of protein structure at better than 70% accuracy**. *Journal of Molecular Biology*. 232:584-599
- [Rost and Sander 1994] Rost, B., Sander, C. 1994. **Combining evolutionary information and neural networks to predict protein secondary structure**. *Proteins*, 19:55-72.

[Rost 1998] Rost B., 1998. **Protein Structure Prediction in 1D, 2D, and 3D**. The Encyclopedia of Computational Chemistry (eds. PvR Schleyer, NL Allinger, T Clark, J Gasteiger, PA Kollman, HF Schaefer III and PR Schreiner), 3, 1998, 2242-2255

[Rost 2001] Rost, B., 2001. **Review: Protein Secondary Structure Prediction Continues to Rise**. Journal of Structural Biology, 134, 204-218.

[Schmidler et al. 2000] Schmidler, S. C., Liu, J. S. and Brutlag, D. L., 2000. **Bayesian segmentation of protein secondary structure**. Journal of Computational Biology Vol. 7, No 1/2., 2000, 233-248

[Shavlik and Maclin 1993] Shavlik, J., Maclin, R., 1993. **Using Knowledge-Based Neural Networks to Improve Algorithms: Refining the Chou-Fasman Algorithm for Protein Folding**. Machine Learning Journal, 11, 195-215

[Skolnick and Kolinski 2001] Skolnick, J., Kolinski, A., 2001. **Computational Studies of Protein Folding**. Computing in Science and Engineering. September/October, Vol. 3, No. 5, 40-49

[Thiele et al. 1999] Thiele, R., Zimmer, R., Lengauer, T. **Protein Threading by Recursive Dynamic Programming**. Journal of Molecular Biology 290, 757-779

[Veljkovic et al. 1985] Veljkovic V, Cosic I, Dimitrijevic B, Lalovic D., 1985. **Is It Possible To Analyze DNA and Protein Sequences by the Methods of Digital Signal Processing?** IEEE Transactions on Biomedical Engineering, Vol. BME-32, No. 5, 337-341, 1985

[Xia et al. 2000] Xia, Y., Huang, E., Levitt, M., Samudrala, R. 2000. **Ab Initio Construction of Protein Tertiary Structures Using a Hierarchical Approach**. Journal of Molecular Biology 300: 171-185

[Xu et al. 1999] Xu, Y., Xu, D., Crawford, O., Einstein, J., Larimer, F., Uberbacher, E., Unseren, M., Zhang, G. 1999. **Protein Threading by PROSPECT: a prediction experiment in CASP3**. Protein Engineering vol.12 no. 11 899-907

[Zhang et al. 1998] Zhang, C., Lin, Z.S., Zhang, Z.D., Yan, M., 1998. **Prediction of the helix/strand content of globular proteins based on their primary sequences**. Protein Engineering, Vol. 11, No. 11, 971-979