Zafer Aydin and
Yucel Altunbasak

A Signal Processing Application in Genomic Research: Protein Secondary Structure Prediction

Bioinformatics is the use of computational techniques to understand and organize information associated with biomolecules. These biomolecules include genetic materials such as nucleic acids and proteins. Ongoing research is generating huge amounts of information about these biomolecules at a phenomenal rate. This requires efficient techniques and methodologies to organize, analyze, and interpret the results in a biologically meaningful manner. The high variability in the data acquisition process, the high dimensionality of the data space, and the high complexity of genetic signals call for sophisticated mathematical modeling, data processing, and information extraction methods [1].

DNA is a molecule of two helical strands where each strand is a long string of nucleotide units attached to one another. DNA contains four types of nucleotides (A, T, G, C) and can be represented by a string of characters. Coded in the DNA are instructions necessary for a cell's proper functioning. Those instructions are stored in specific units called genes. When a particular instruction becomes active, the corresponding gene is said to turn on or be expressed. Following the expression of a particular gene, the corresponding section of the DNA strand is copied into a less stable molecule called messenger RNA (mRNA). The process of producing mRNA is called *transcription*. The mRNA is then transferred to the ribosome, where the protein molecule is produced by interpreting the instruction in mRNA. This process of producing proteins is referred to as *translation*. Translation takes place according to the genetic code, which maps successive triplets of RNA bases to amino acids. Thus, a protein is a chain of amino acid units. There are 20 types of standard amino acids that are regularly found in nature as well nonstandard types that rarely appear. Similar to DNA, a protein molecule can be represented by a string of characters from an alphabet of size 20. Proteins are the molecules responsible for all cellular functions. The specific function of a protein is largely determined by the three-dimensional (3-D) structure into which it folds.

DNA molecules are packed into structural units called *chromosomes*. The total content of the DNA within the chromosomes is called the *genome* of an organism. All the cells within a living organism (with the exception of the sperm and egg cells) contain nearly identical copies of the entire genome and are capable of producing all proteins. Although the same instructions are present in almost every cell of an organism, their activation patterns show significant variation among different cell types. Cells differ radically in the proteins they actually produce. Moreover, any given cell produces different proteins at different stages within its cycle of operation. There is a control mechanism that regulates the protein production of the cell by its internal environment and by the signals received from other cells. Therefore, gene expression can be viewed as a complex network of interactions involving genes, proteins, and RNA as well as other factors such as temperature and the presence or absence of nutrients and drugs within the cell.

The major goals of the genomic research can be summarized as follows [2]:

1) Sequencing and comparison of genomes of different species. To sequence a genome means to determine its sequence of nucleotides. The sequencing of a new organism is often valuable for medical, agricultural, or environmental studies. In addition, it may be useful for comparative studies with related organisms.

2) Identifying genes and determining the functions of proteins they encode. Genes can be predicted using methods confined to a single genome or by comparative methods that use information about one organism to understand another related one.

3) Predicting the structural features of the protein from the amino acid sequence. One of the biggest challenges in today's biological science is the protein folding problem in which one attempts to predict the 3-D structure of a protein from its linear sequence of amino acids. Protein folding methods greatly benefit from secondary structure prediction, in which regular structural units with specific chemical bonding patterns are identified.

4) Understanding gene expression and gene protein interaction to control cellular processes. Here, instead of analyzing individual sequences, interactions between biomolecules are investigated at a systems level. Regulatory networks model how genes are jointly expressed in time, space, and amplitude. The design of such networks is greatly facilitated by microarray technology, in which array chips take a snapshot of the cellular activity by providing expression levels of thousands of genes simultaneously (Figure 1).

5) Tracing the evolutionary relationships among existing species and constructing *phylogenetic* trees. Understanding the genomic behavior in one organism will definitely contribute to the study of other

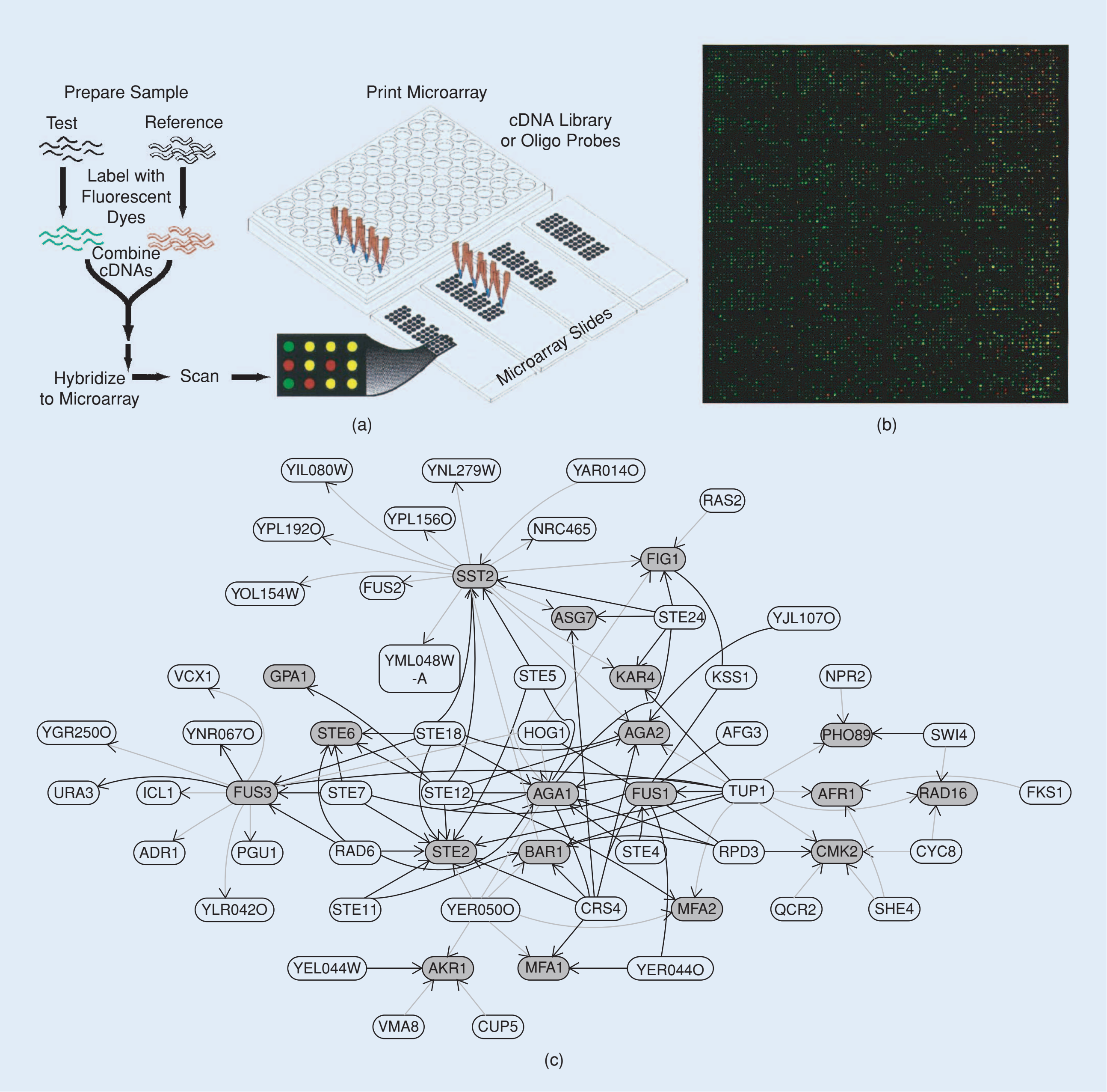
1053-5888/06/$20.00©2006IEEE

genetically similar organisms with similar DNA sequences. Information generated from simple model organisms such as mycoplasma, *E. coli*, and yeast helps to understand cellular mechanisms in more complex organisms such as humans.

6) Discovering associations between gene mutations and disease. Certain diseases, such as cystic fibrosis and Huntington's disease, are caused by a single mutation. Others such as heart disease, cancer, and diabetes are influenced by both genetic and environmental factors. In these cases, the genetic component involves a combination of influences from many genes. By studying the relationship between genetic endowment and disease states it is possible to develop efficient techniques to diagnose and cure genetically influenced diseases.

SIGNAL PROCESSING AND GENOMIC RESEARCH

The digital nature of genomic information makes it suitable for the application



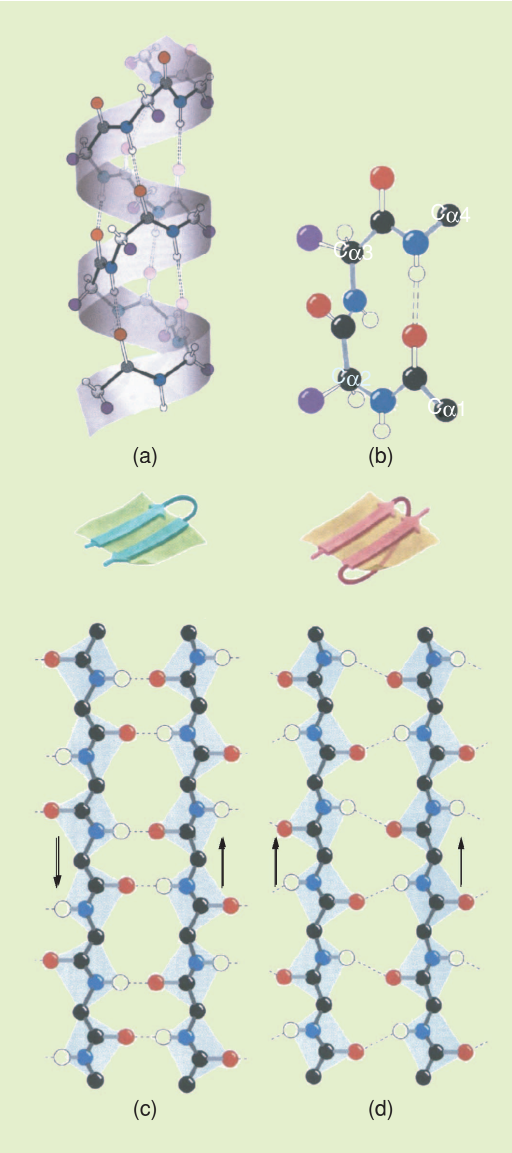
[FIG1] (a)–(b): DNA microarray technology. (a) Array fabrication and steps of a microarray experiment. (Reprinted from *Guide to Analysis of DNA Microarray Data,* Steen Knudsen, ©2004, John Wiley & Sons, Inc. Courtesy of Steen Knudsen with permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.) (b) A typical microarray image. (Reprinted from *Guide to Analysis of DNA Microarray Data,* Steen Knudsen, ©2004, John Wiley & Sons, Inc. Courtesy of Steen Knudsen with permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.) (c) A gene expression network. Nodes represent genes and edges represent the relationship between genes. (Courtesy of Thomas Schlitt and Alvis Brazma.)

of signal processing techniques to better analyze and understand the characteristics of DNA, proteins, and their interaction. Prediction of genes, protein structure, and protein function greatly utilize pattern recognition techniques, in which hidden Markov models, neural networks, and support vector machines (SVMs) play a central role. For instance, in hidden Markov models, predictions can be obtained using the Viterbi and BCJR (Bahl-Cocke-Jelinek-Raviv) decoding algorithms, which are widely used techniques in communications and speech recognition. Moreover, the subsequent analysis of microarray data (e.g., classification, gene selection, and clustering) seeks to extract meaningful results from the noisy measurements and reliably infer gene regulatory networks [3]. In that respect, genomic research greatly benefits from the signal processing theory for the detection of genes with strong or weak expression patterns; classification of genes according to their similarity in expression levels; and prediction, control, and statistical-dynamical modeling of gene networks. Therefore, signal processing offers a variety of methods from pattern recognition and network analysis for the diagnosis and therapy of genetic diseases [4].

In the next section, we will focus on protein secondary structure prediction and describe the problems in single sequence setting.

PROTEIN SECONDARY STRUCTURE PREDICTION

Proteins are large, complex molecules made up of smaller subunits called amino acids. Chemical properties that distinguish the 20 standard amino acids cause the protein chains to fold up into specific 3-D structures that define their particular functions in the cell. There are four levels of protein structure. The primary structure refers simply to the "linear" sequence of amino acids. The secondary structure is the "locally" ordered structure created by hydrogen bonding within the protein backbone. Tertiary structure refers to the "global" folding of a single polypeptide chain, and quaternary structure involves the association of two or more polypeptide chains into a multisubunit structure. Correspondingly, there are several levels at which protein structure prediction can be performed. Secondary structure prediction is concerned with the assignment of each amino acid to a secondary structure state. In tertiary structure estimation (i.e., protein folding), the goal is to predict the conformation assumed by a protein molecule in 3-D space.



[FIG2] (a)-(b): Local interactions in α-helix and loop segments. (c)-(d): Nonlocal interactions in β-strand segments. (Reprinted from *Biochemistry, 3rd Edition*, Donald Voet, Judith G. Voet, ©2004 John Wiley & Sons, Inc. Courtesy of Irving Geis. Rights owned by Howard Hughes Medical Institute. Not to be used without permission.)

Prediction of the secondary structure is important as it provides insights into the function of the protein. By jointly comparing amino acid and secondary structure sequences, it is possible to improve the prediction of protein function. In addition, secondary structure prediction is a step toward the prediction of the 3-D structure of a protein. For instance, secondary structure predictions can be included in fold recognition methods, in which a target amino acid sequence with unknown structure is compared against a library of structural templates (folds) and the best scoring fold is assumed to be the one adopted by the sequence.

The three major secondary structure states are the α-helix {H}, the β-strand {E}, and the loop {L}. α-helices are strengthened by hydrogen bonds between every fourth amino acid so that the protein backbone adopts a helical configuration as shown in Figure 2(a). Likewise in loops (e.g., turns or bends), the hydrogen bonding is mostly local. For example, the turn segment in Figure 2(b) has a hydrogen bond between the first and the fourth amino acids. The hydrogen bonding structure in β-strands is slightly different, where both local and nonlocal interactions are observed. In β-strands, the most common local hydrogen bonding is between every two amino acids, and nonlocal interactions are due to hydrogen bonds between amino acid pairs positioned in interacting β-strand segments. A β-sheet is a set of such segments, in which the interacting segment pairs adopt either a parallel or an antiparallel conformation as shown in Figure 2.

Secondary structure prediction assigns to each amino acid a structural state from a three-letter alphabet {H, E, L}. Secondary structure prediction methods often employ neural networks (NNs) [5], SVMs [6], and hidden Markov models (HMMs)

[7], [8]. In HMM methods, hidden states generate segments of amino acids that correspond to the nonoverlapping secondary structure segments, and the goal is to find the most likely hidden-state sequence representation under the probabilistic model defined by the HMM. On the other hand, neural networks and SVMs utilize an encoding scheme to represent the amino acid residues by numerical vectors. To convert the amino acids into vectors, the amino acid sequence is partitioned into overlapping segments by a sliding window of size n (typically between 13 and 17). Then each segment is represented (as a vector) in the $21 \times n$-dimensional input space to predict the secondary structure class of the central residue. Here the first 20 dimensions are allocated for the amino acid types, and the 21st dimension is added to be able to extend the window over the sequence ends. The secondary structure prediction problem then becomes the classification of points in a multidimensional vector space. This is achieved by partitioning the space into disjoint regions of secondary structure classes. NN methods perform the classification in the space of the input vector by defining decision boundaries. On the other hand, SVM methods first map the input vectors into a higher dimensional Hilbert space by a transformation kernel and then perform the classification in that space by finding separating hyperplanes.

There are two types of protein secondary structure prediction algorithms. A single sequence algorithm does not use information about other similar proteins. The algorithm should be suitable for a nonhomologous sequence with no sequence similarity to any other protein sequence. Algorithms of another type explicitly use sequences of homologous proteins, which often have similar structures. Prediction accuracy of such an algorithm should be higher than one of a single sequence algorithm due to incorporation of additional evolutionary information from multiple alignments. The accuracy (sensitivity) of the best current single sequence prediction methods is below 70% [8]. The prediction accuracy of the best prediction methods that employ information from multiple alignments is close to 82.0% [5]. The theoretical limit of the accuracy of secondary structure assignment from experimentally determined 3-D structure is around 88% [9].

Single-sequence algorithms for protein secondary structure prediction are important because a significant percentage of proteins identified in genome sequencing projects have no detectable sequence similarity to any known protein. Also, many of these hypothetical proteins do not have detectable similarity to any protein at all. Such “orphan” proteins may represent a sizable portion of a proteome. For an orphan protein, any method of secondary structure prediction performs as a single sequence method. Developing better methods of protein secondary structure prediction from single sequence has a definite merit as it helps in improving the functional annotation of orphan proteins.

RESEARCH PROBLEMS IN SECONDARY STRUCTURE PREDICTION FOR A SINGLE SEQUENCE

The main goal of a secondary structure prediction algorithm should be to design a classifier having a feature set (dependency structure) that is comprehensive enough to capture the essential correlations yet simple enough to allow reliable parameter estimation from available training data. In single-sequence prediction, one issue limiting the prediction accuracy is the small sample size. The latest available single-sequence data set for evaluating the prediction performance of a single-sequence method contains approximately 3,200 proteins, which significantly limits the number of features. As new sequences are added to the database, it will be possible to augment the dependency structure and obtain even higher accuracy.

The second and more crucial problem arises when predicting β-strands. Typically, protein secondary structure prediction methods suffer from low accuracy in β-strand predictions (40–50%) where nonlocal correlations resulting from long-distance chemical interactions play a significant role. To the best of our knowledge, there have not been any significant attempts to model nonlocal dependencies for single sequence predictions. There is a considerable need to model such long-distance interactions that significantly contribute to the stabilization of the protein molecule.

The estimation of model parameters (training) is as important as the model derivation and feature selection. A well-trained model is more likely to succeed on new examples. Therefore, improved training methods will significantly contribute to the secondary structure prediction accuracy. (Interested readers can refer to the extended version of this column at http://www.ece.gatech.edu/research/labs/MCCL/ pubs/journal.html.)

AUTHORS

Zafer Aydin (aydinz@ece.gatech. edu) and ***Yucel Altunbasak*** (yucel@ece.gatech. edu) are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta.

REFERENCES

[1] X.Y. Zhang, F. Chen, Y.T. Zhang, S.C. Agner, M. Akay, Z. H. Lu, M.M.Y. Waye, and S.K.W. Tsui, “Signal processing techniques in genomic engineering,” *Proc. IEEE*, vol. 90, no. 12, pp. 1822–1833, 2002.

[2] R.M. Karp, “Mathematical challenges from genomics and molecular biology,” *Notices AMS*, vol. 49, no. 5, pp. 544–553, 2002.

[3] J. Chen, H. Li, K. Sun, and B. Kim, “How will bioinformatics impact signal processing research,” *IEEE Signal Processing Mag.*, vol. 20, no. 6, pp. 16–26, 2003.

[4] E.R. Dougherty and A. Datta, “Genomics signal processing: Diagnosis and therapy,” *IEEE Signal Processing Mag.*, vol. 22, no. 1, pp. 107–112, 2005.

[5] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri, “Exploiting the past and the future in protein secondary structure prediction,” *Bioinformatics*, vol. 15, no. 11, pp. 937–946, 1999.

[6] J. Guo, H. Chen, Z. Sun, and Y. Lin, “A novel method for protein secondary structure prediction using dual-layer SVM and profiles,” *Proteins*, vol. 54, no. 4, pp. 738–743, 2004.

[7] S.C. Schmidler, J.S. Liu, and D.L. Brutlag, “Bayesian segmentation of protein secondary structure,” *J. Comp. Biol.*, vol. 7, no. 1/2, pp. 233–248, 2000.

[8] Z. Aydin, Y. Altunbasak, and M. Borodovsky, “Protein secondary structure prediction with semi Markov HMMs,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing 2004 (ICASSP’04)*, 2004, vol. 5, pp. 577–580.

[9] B. Rost, “Rising accuracy of protein secondary structure prediction,” in *Protein Structure Determination, Analysis, and Modeling for Drug Discovery*, D Chasman, Ed. New York: Marcel Dekker, 2003, pp. 207–249.

[10] J. Rung, T. Schlitt, A. Brazma, K. Freivalds, and J. Vilo, “Building and analysing genome-wide gene disruption networks,” *Bioinformatics*, vol. 18, Suppl. 2, pp. S202–210, 2002.

[SP]