Chapter 8

ANALYZING PROTEIN SEQUENCES USING SIGNAL ANALYSIS TECHNIQUES

Karen M. Bloch¹ and Gonzalo R. Arce²
¹DuPont Company, Wilmington, Delaware, USA
²Department of Electrical & Computer Engineering, University of Delaware, Newark, Delaware, USA

1. Introduction

Genomes carry all information of life from one generation to the next for every organism on earth. Each genome, which is a collection of DNA molecules, can be represented as a series of strings comprised of four letter symbols. Less than 10 years ago, determining the sequence of these letters to read a single gene was a slow tedious process. But today, through the use of new strategies, genome sequencing is a billion-dollar worldwide effort in both industry and academia.

At the end of 1998, researchers had completely read the genome of only one multicellular organism, a worm known as C. elegans. Now, sequences exist for the fruit fly, the human and for the weed important to plant geneticists known as Arabidopsis. Drafts of the genomes of the mouse, rat, zebrafish, and pufferfish are soon to be completed. Researchers have also been working on simpler organisms. Several dozen microbial genomes are now available, including those that cause cholera and meningitis. Most of these data are accessible free of charge, encouraging the exploration of this data. However, it is not the genes, but the proteins they code for that actually do all the work. The search for protein function has lead to the era of proteomics, the identification and characterization of each protein and its structure, and of every protein-protein interaction (Pennisi, 2000).

Proteins are the molecules that accomplish most of the functions of living cells. All proteins are constructed from linear sequences of smaller molecules called amino acids. There are twenty naturally occurring amino acids and they can be represented in a protein sequence as a
string of alphabetic symbols. Protein molecules fold to form specific three dimensional shapes which specify their particular chemical function (Hunter, 1993).

Analysis of protein sequences can provide insights into function and can also lead to knowledge regarding biologically active sites of the protein. While analysis of protein sequences is often performed directly on the symbolic representation of the amino acid sequence, patterns in the sequence are often too weak to be detected as patterns of symbols. Alternative sequence analysis techniques can be performed by assigning numerical values to the amino acids in a protein. The numerical values are derived from the physicochemical properties of the amino acid and are relevant to biological activity. It has been shown that the EIIP, as one such measure, correlates with certain biological properties (Veljkovic, 1985). Once a numerical mapping for a protein sequence is achieved, the sequence can be treated as a signal.

From a mathematical point of view, a signal can be described in a variety of ways. For example, a signal can be represented as a function of time which shows how the signal magnitude changes over time. Alternatively, a signal can be written as a function of frequency by performing a Fourier transform. This tells how quickly a signal’s magnitude changes (Qian, 1996). For many real world applications, it is useful to characterize a signal in the time and frequency domains simultaneously. Such signal analysis methods can provide fingerprints which indicate the existence of some event of importance. In the case of protein sequences represented as numerical signals, such an event might be the existence of a binding site.

This chapter illustrates the use of frequency and time-frequency signal analysis techniques with two classes of proteins, fibroblast growth factors and homeodomain proteins. Fibroblast growth factors constitute a family of proteins that affect the growth, migration, differentiation, and survival of certain cells. Homeodomain proteins contain a single 60-amino acid DNA binding domain. It is the numerical representation of these amino acid sequences, along with various frequency and time-frequency analysis methods which we describe herein.

2. Frequency Analysis of Proteins

The Resonant Recognition Model (RRM) (Cosic, 1994) is a physico-mathematical model that analyses the interaction of a protein and its target using signal processing methods. One application of this model involves prediction of a protein’s biological function. In this technique, a Fourier transform is applied to a numerical representation of a protein