Chapter 13
Deterministic Pattern Mining on Genetic Sequences

Pedro Gabriel Ferreira
Centre for Genomic Regulation, Spain

Paulo Jorge Azevedo
Universidade do Minho, Portugal

ABSTRACT

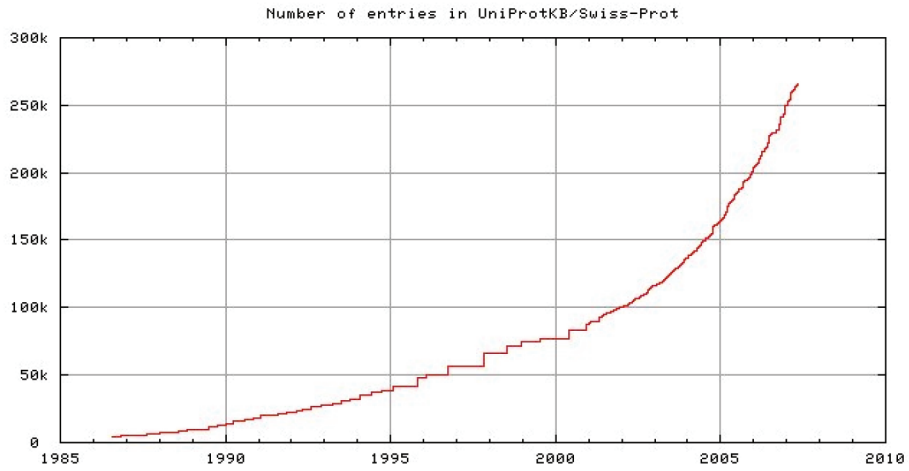
The recent increase in the number of complete genetic sequences freely available through specialized Internet databases presents big challenges for the research community. One such challenge is the efficient and effective search of sequence patterns, also known as motifs, among a set of related genetic sequences. Such patterns describe regions that may provide important insights about the structural and functional role of DNA and proteins. Two main classes can be considered: probabilistic patterns represent a model that simulates the sequences or part of the sequences under consideration and deterministic patterns that either match or not the input sequences. In this chapter a general overview of deterministic sequence mining over sets of genetic sequences is proposed. The authors formulate an architecture that divides the mining process workflow into a set of blocks. Each of these blocks is discussed individually.

INTRODUCTION

The Human Genome Project (NIH, 2007; Cooper, 1994) has led to the development of a set of new biological techniques. These new technological breakthroughs have in turn resulted on an exponential growth of biological data. In particular, the development of sequencing techniques culminated in a massive accumulation of genetic (DNA and protein) sequence data, which have then become freely available through specialized internet databases. For instance, the GenBank database contains approximately 85 759 586 764 nucleotides in 82 853 685 sequences (February 2008). The UniProtKB/Swiss-Prot contains 392 667 sequence entries, comprising 141 217 034 amino acids (July 2008). Figure 1 depicts the growth of UniProtKB/Swiss-Prot in the last years. Such amount of data raises big challenges both at the

DOI: 10.4018/978-1-60566-766-9.ch013

Figure 1. Growth in the number of protein sequences entered in the UniProtKB/Swiss-Prot database.



organizational and analysis level. In fact, the way that molecular biology research is done has changed significantly, being now a much more data-oriented and computationally based discipline. This fact, combined with the lack of robust theories to interpret these data opens a new and vast research area. The focus is now not on how to obtain the data, but on how to understand it. In the particular case of sequence data, this poses the question of how to retrieve biologically relevant patterns.

In the next two sections we briefly motivate the use of sequence mining to better understand DNA and protein mechanisms. More information about these topics can be easily found in any biology text book, for instance (Alberts, 2002; John, 1995). We also refer the reader to the following introductory articles (Hunter, 1993; Koonin, 2003; Brazma, 2001; Lesk, 2002).

DNA Sequence Mining

The DNA sequence contains all the necessary genetic information for the life of the being. Along this extremely long sequence, different regions encode different biological functional units. The function for some of these regions still needs to be determined. One of these regions is called *gene* and contains all the information necessary to create a protein. For a gene to be expressed, i.e. to result into a protein, it is necessary that a large number of conditions are fulfilled. One of these conditions is the existence of a certain types of sequences signals upstream and downstream the gene region. An important problem in bioinformatics is the identification of these signals in sequence segments found nearby genes. Such signals, called sequence patterns or motifs, are of major importance since they can provide insights into gene expression regulation.

Protein Sequence Mining

The analysis of a set of biologically related protein sequences may reveal regions of amino acid residues that occur highly conserved in several of those sequences. Such fact is certainly related to an evolutionary, chemical, structural or functional role of the protein. *Domains*, which consist of relatively small