Chapter VI
Deterministic Motif Mining in Protein Databases

Pedro Gabriel Ferreira
Universidade do Minho, Portugal

Paulo Jorge Azevedo
Universidade do Minho, Portugal

ABSTRACT

Protein sequence motifs describe, through means of enhanced regular expression syntax, regions of amino acids that have been conserved across several functionally related proteins. These regions may have an implication at the structural and functional level of the proteins. Sequence motif analysis can bring significant improvements towards a better understanding of the protein sequence-structure-function relation. In this chapter, we review the subject of mining deterministic motifs from protein sequence databases. We start by giving a formal definition of the different types of motifs and the respective specificities. Then, we explore the methods available to evaluate the quality and interest of such patterns. Examples of applications and motif repositories are described. We discuss the algorithmic aspects and different methodologies for motif extraction. A brief description on how sequence motifs can be used to extract structural level information patterns is also provided.

INTRODUCTION

Proteins are biological macromolecules involved in all biochemical functions in the life of the cell and therefore in the life of the being. Protein information is encoded in regions of the DNA helix, and these molecules are synthesized through a two step process: *translation* and *transcription* (Cooper, 1994; Hunter, 1993). Proteins are composed of basic unit molecules called *amino acids*.

Twenty different types of amino acids (AAs) exist, all with well differentiated structural and chemical properties.

After being synthesized, proteins acquire a complex 3-dimensional structure in a process called *folding*. The resulting 3D structure, which corresponds to a state of greatest stability (minimal energy), is essential for protein function. This structure is ultimately determined by the linear sequence of amino acids, also called primary structure. Therefore, a closer look at the primary sequence will certainly provide valuable insights about the protein structure and function.

When a set of functionally related sequences is closely analyzed, one can verify that parts of those sequences (subsequences) are common to several or all the analyzed sequences. These subsequences consist of a pattern and are called *sequence patterns* or *motifs*. These motifs occur in protein sequences because they have been preserved through the evolutionary history of the proteins. This suggests that they might play a structural and/or a functional role in the protein's mechanisms. On the other hand, AAs outside these critical regions tend to be less conserved. The discovery of these motifs can be used to support a better understanding of the protein's structure and function. This is due to the fact that the AAs that compose these motifs can be close in the tridimensional arrangement of the protein. Additionally, these motifs can be used to provide evidences and to determine relations with yet uncharacterized proteins.

At the time of this writing,[1] Swiss-Prot (Gasteiger, 2003), which is a comprehensive, annotated, and nonredundant protein knowledge base, contained approximately 208,000 sequences from 9,749 species, with an average length per sequence of 364 AAs. This volume of information demands intelligent and efficient sequence analysis techniques. These methods should look for similarities among the selected proteins and discriminate the regions that have been conserved among a significant number of proteins. These regions contain well-conserved positions, where the substitutions among different AAs for those positions are less frequent. Motifs can be used to capture the nature of those regions.

In this chapter, we present an overview on the subject of protein motif mining. The chapter has the following outline: First a characterization on the type of extracted patterns is given. Two main classes of motifs are introduced and briefly described (Motif Definition section). Since these two classes have different analysis and algorithmic requirements, we will focus our attention on the class of deterministic patterns. In the Deterministic Motifs section, details of the characteristics of this type of patterns are provided. Next, different ways to evaluate the interest of the motifs (Significance Measures section) are presented, followed by examples of the application of motifs in different contexts (Motif Applications section). In the Motif Databases section, several of the Internet databases that compile and manage protein motifs are surveyed. The motif mining algorithms section describes the algorithmic aspects of the motif extraction process, and some of the most well-known and successful methodologies for motif mining are presented. In the Structural Motifs section, the concept of structural motifs is introduced, and examples of motifs and algorithms are provided. To finish, some conclusions and final remarks are given.

Motif Definition

As previously introduced, a sequence motif describes a region of conserved elements from a set of related sequences. These motifs are eventually related to an important structural and/or functional role of the proteins. Two classes of motifs exist: *probabilistic* and *deterministic*.

Probabilistic motifs consist in a model that simulates the sequences or part of the sequences under consideration. When a given sequence is compared against the motif, the probability of that sequence matching the given motif can be easily