Chapter 4.9
Data Mining in Proteomics Using Grid Computing

Fotis E. Psomopoulos
Aristotle University of Thessaloniki, Greece

Pericles A. Mitkas
Aristotle University of Thessaloniki, Greece

ABSTRACT

The scope of this chapter is the presentation of Data Mining techniques for knowledge extraction in proteomics, taking into account both the particular features of most proteomics issues (such as data retrieval and system complexity), and the opportunities and constraints found in a Grid environment. The chapter discusses the way new and potentially useful knowledge can be extracted from proteomics data, utilizing Grid resources in a transparent way. Protein classification is introduced as a current research issue in proteomics, which also demonstrates most of the domain – specific traits. An overview of common and custom-made Data Mining algorithms is provided, with emphasis on the specific needs of protein classification problems. A unified methodology is presented for complex Data Mining processes on the Grid, highlighting the different application types and the benefits and drawbacks in each case. Finally, the methodology is validated through real-world case studies, deployed over the EGEE grid environment.

INTRODUCTION

Although computational biology and bioinformatics are often confused as the same interdisciplinary field, they do have several distinguishing differences. Bioinformatics is mainly concerned with the analysis and processing of data, and therefore the advancement in both algorithmic and technical level of the techniques and theories to solve formal and practical data management problems. On the other hand, computational biology aims to solve specific biological problems, utilizing computers to test and evaluate hypotheses. The working definitions of these two fields, provided by National Institutes of Health (NIH, 2000), are the following:

DOI: 10.4018/978-1-4666-0879-5.ch4.9

*"***Bioinformatics:** *Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data."*

*"***Computational Biology:** *The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems."*

However, it is also emphasized that "*although bioinformatics and computational biology are distinct, there is also significant overlap and activity at their interface*". Proteomics is one of the key fields that exist in that overlapping area. In a nutshell, proteomics is the large-scale study of proteins, ranging from the structural and functional analysis to the construction of protein-protein interaction networks and phylogenetic trees. Proteins are large organic molecules composed of amino acids arranged in a linear chain and held together by peptide bonds. They are essential part of organisms, participating in all processes within cells; catalyzing biochemical reactions (enzymes), maintaining the cell shape serving as scaffolds, providing the means of signaling between cells, etc. The term proteome denotes the entire complement of proteins expressed by a genome at a given time and under defined conditions. The word itself is a portmanteau of "protein" and "genome".

There has been a recent shift in focus from genomics to proteomics, due to the fact that many consider proteomics to be the next step in the study of biological systems. The genome of an organism is fairly stable, showing little variation throughout its cells in comparison with the proteome, which is highly differentiated from cell to cell. One of the more significant insights that have emerged from proteomics is the nature of relationship between genes and proteins. The study of the mouse proteome (Gauss, 1999) has demonstrated that a protein can be considered as the expression of not one but many genes (Klose, 1999). Correspondingly, a single mutation in a gene can affect many proteins. Moreover, using the yeast proteome, the essential-essential protein interaction network has been proposed to form a generic scaffold around which organism-specific and taxon-specific proteins and interaction coalesce (Pereira-Leal, 2005).

BACKGROUND

In this section, some insight into the main data acquisition methods in proteomics will be provided, in order to present the common difficulties that may arise during data analysis. As far as the actual analysis is concerned, the main focus will be on the protein classification problem, due to the fact that it exhibits several of the issues common in other bioinformatics areas. Finally, after defining the concepts of Grid and Grid Computing, an overview of the current status concerning the symbiosis of bioinformatics and grid computing will be discussed.

Proteomics

There are several primary proteomics techniques for data acquisition; gel electrophoresis and mass spectrometry. Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) separates proteins in the first dimension according to charge and in the second dimension according to molecular mass (O'Farell, 1975). One of the many variations of this technique is two-dimensional difference in-gel electrophoresis (2D-DIGE), in which proteins from two samples (e.g., normal vs. diseased) are differentially labeled using fluorescent dyes and simultaneously electrophoresed (Unlu, 1997). Mass spectrometry (MS) on the other hand, is essential for protein identification (Glich, 2003; Honore, 2004). Utilizing ionization of the crystallized protein, and based on the applied voltage and