Data Mining in Bioinformatics

André C.P.L.F. de Carvalho

Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Departamento de Ciências de Computação e Estatistica
Caixa Postal 668, 13560-970 - São Carlos, SP, Brasil
andre@icmc.usp.br

Abstract Molecular Biology laboratories have gathered a very large amount of data in sequence and functional genome projects. It is frequently not possible to analyze these data manually. Sophisticated computing techniques are necessary to extract new, meaningful and useful information from these data. Data Mining techniques have been successfully applied in such analysis. Examples of these applications are analysis of gene expression data, recognition of genes in DNA sequences and protein structure prediction. This tutorial will present the main issues on the use of Data Mining techniques in Bioinformatics. The tutorial will start with the introduction of the key aspects of Data Mining, with special emphasis on Machine Learning. Next, the necessary issues of molecular biology for the understanding of the Data Mining applications in Bioinformatics will be described. Usually, biological data needs to be pre-processed before they can be used in a Data Mining process. The main techniques for data pre-processing will be presented. Later, a few applications of data Mining techniques, mainly classification and clustering, to bioinformatics problems will be presented.