Chapter 11
Analysis and Integration of Biological Data:
A Data Mining Approach using Neural Networks

Diego Milone
Universidad Nacional del Litoral & National Scientific and Technical Research Council, Argentina

Laura Kamenetzky
Institute of Biotechnology, INTA & National Scientific and Technical Research Council, Argentina

Georgina Stegmayer
Universidad Tecnologica Nacional & National Scientific and Technical Research Council, Argentina

Mariana López
Institute of Biotechnology, INTA & National Scientific and Technical Research Council, Argentina

Matías Gerard
Universidad Nacional del Litoral & Universidad Tecnologica Nacional & National Scientific and Technical Research Council, Argentina

Fernando Carrari
Institute of Biotechnology, INTA & National Scientific and Technical Research Council, Argentina

ABSTRACT

The volume of information derived from post genomic technologies is rapidly increasing. Due to the amount of involved data, novel computational methods are needed for the analysis and knowledge discovery into the massive data sets produced by these new technologies. Furthermore, data integration is also gaining attention for merging signals from different sources in order to discover unknown relations. This chapter presents a pipeline for biological data integration and discovery of a priori unknown relationships between gene expressions and metabolite accumulations. In this pipeline, two standard clustering methods are compared against a novel neural network approach. The neural model provides a simple visualization interface for identification of coordinated patterns variations, independently of the number of produced clusters. Several quality measurements have been defined for the evaluation of the clustering results obtained on a case study involving transcriptomic and metabolomic profiles from tomato fruits. Moreover, a method is proposed for the evaluation of the biological significance of the clusters found. The neural model has shown a high performance in most of the quality measures, with internal coherence in all the identified clusters and better visualization capabilities.

DOI: 10.4018/978-1-4666-2455-9.ch011

INTRODUCTION

Nowadays, the biology field is in the middle of a data explosion. A series of technical advances in recent years has increased the amount of data that biologists can record about different aspects of an organism at the genomic, transcriptomic and proteomic levels (Keedwell & Narayanan, 2005).

Nowadays, the discipline of computational biology has allowed biologists to make full use of the advances in computer science and statistics in understanding this information. Due to the amount and nature of the biological data involved (such as noisy and missing data), novel computational methodologies are needed for properly analysing it. Moreover, as the volume of data continues to grow at a high speed, new challenges appear, such as the need to extract information that was not previously known from these databases to supplement current knowledge. For example, the discovery of hidden patterns of gene expression in microarray and metabolite profiles from plants of economic importance to agro-biotechnology, is a current challenge because the use of any algorithm for pattern recognition suffers from the so-called curse of dimensionality. In addition, data integration is also gaining attention given the need for merging and extracting knowledge from signals of different sources and nature. Visualization of results is also an important issue for the understanding and interpretation of hidden relationships (Tasoulis, Plagianakos, & Vrahatis, 2008).

Bioinformatics has evolved over time, mainly from the development of data mining techniques and their application to automatic prediction and discovery of classes, two key tasks for the analysis and interpretation of gene expression data on microarrays (Polanski & Kimmel, 2007). The prediction of classes uses the available information on the expression profiles and the known characteristics of the sets of data or experiments to build classifiers for future data. On the contrary, in the case of classes discovery, data are explored from the viewpoint of the existence or not of unknown relations and a hypothesis to explain them is formulated (Golub et al., 1999). Among class discovery techniques, the Hierarchical Clustering (HC) algorithm is the most commonly used technique in biological data. It is a deterministic method based on a pairwise distance matrix. This algorithm establishes small groups of genes/conditions that have a common expression pattern and then constructs a dendrogram, sequentially, on the basis of the distances between feature vectors. Clusters are obtained by pruning the tree at some level, and the number of clusters is controlled by deciding at which level of the hierarchy of the tree the splitting is performed (Tasoulis et al., 2008). Regarding non-hierarchical algorithms, the distances are calculated from a predetermined number of clusters and the genes are iteratively placed in different groups until minimizing each cluster internal spread. The more representative algorithm of this type is the *k*-means (KM) algorithm (Duda & Hart, 2003).

NEW TRENDS

One of the current trends in the field is the integration of two types of biological data: metabolic profiles and transcriptional data from microarrays, with the objective of finding hidden relations among them and to infer new knowledge about the biological processes that involve them (Bino et al., 2004). For example, a problem of interest is how to evaluate the presence of genes associated with regulatory mechanisms in metabolic pathways. This is especially important in plants due to the availability of primary and secondary metabolites and the wide variety of genes associated with these pathways. In particular the integration of data of transcriptome and metabolome in plants, correlating gene transcription profiles with variations profiles of a large number of non-protein molecules, can be used for the identification of changes not reflected in the plant morphology (Carrari et al., 2006). This allows having a snapshot