Bio-Inspired Algorithms in Bioinformatics I

José Antonio Seoane Fernández
University of A Coruña, Spain

Mónica Miguélez Rico
University of A Coruña, Spain

INTRODUCTION

Large worldwide projects like the Human Genome Project, which in 2003 successfully concluded the sequencing of the human genome, and the recently terminated Hapmap Project, have opened new perspectives in the study of complex multigene illnesses: they have provided us with new information to tackle the complex mechanisms and relationships between genes and environmental factors that generate complex illnesses (Lopez, 2004; Dominguez, 2006).

Thanks to these new genomic and proteomic data, it becomes increasingly possible to develop new medicines and therapies, establish early diagnoses, and even discover new solutions for old problems. These tasks however inevitably require the analysis, filtration, and comparison of a large amount of data generated in a laboratory with an enormous amount of data stored in public databases, such as the NCBI and the EBI.

Computer sciences equip biomedicine with an environment that simplifies our understanding of the biological processes that take place in each and every organizational level of live matter (molecular level, genetic level, cell, tissue, organ, individual, and population) and the intrinsic relationships between them.

Bioinformatics can be described as the application of computational methods to biological discoveries (Baldi, 1998). It is a multidisciplinary area that includes computer sciences, biology, chemistry, mathematics, and statistics. The three main tasks of bioinformatics are the following: develop algorithms and mathematical models to test the relationships between the members of large biological datasets, analyze and interpret heterogeneous data types, and implement tools that allow the storage, retrieve, and management of large amounts of biological data.

BACKGROUND

The following section describes some of the problems that are most commonly found in bioinformatics.

Interpretation of Gene Expression

The expression of genes is the process by which the codified information of a gene is transformed into the necessary proteins for the development and functioning of the cell. In the course of this process, small sequences of ARN, also called ARN messengers, are formed by transcription and subsequently *translated* into proteins.

The amount of expressed mARN can be measured with various methods, such as gel electrophoresis, but large numbers of simultaneous expression analyses are usually carried out with microarrays (Quackenbush, 2001), which make it possible to obtain the simultaneous expression of tens of thousands of genes; such an amount of data can only be analyzed with the help of an informatic process.

Among the most common tasks in this type of analysis is the task to find the differences between, for instance, a patient and a test that determines whether a gene is expressed or not. These tasks can be divided into classical problems of classification and clustering. Clustering is used not only in experiments of microarrays (to identify groups of genes with similar expressions), but also suggests functional relationships between the members of the cluster.

Alignment of ADN, ARN, and Protein Sequences

Sequences alignment consists in superposing two or more sequences of both nucleotides (ADN and ARN) and amino acids (proteins) in order to compare them and analyze the sequence parts that are alike and unalike.

The optimal alignment is that which mainly shows correspondences between the nucleotides or amino acids and is therefore said to have the highest score. This alignment may or may not have a biological meaning. There are two types of alignment: the global alignment, which maximizes the number of coincidences in the entire sequence, and the local alignment, which looks for similar regions in large sequences that are normally highly divergent. The most commonly used technique to implement alignments is dynamic programming by means of the Smith-Waterman algorithm (Smith, 1981), which explores all the possible comparisons in the sequences.

Another problem in sequences alignment is multiple alignment (Wallace, 2005), which consists in aligning three or more sequences of ADN, ARN, or proteins, and is generally used to search for evolutive relationships between these sequences. The problem is equivalent to that of simple sequences alignment, but takes into consideration the n sequences that are to be compared. The complexity of the algorithm increases exponentially with the number of sequences to compare.

Identification of the Gene Regulatory Network

All the information of a living organism's genome is stored in each and every one of its cells. Whereas the genome is used to synthesize information on all the body cells, the regulating network is in charge of guiding the expression of a given set of genes in one cell rather than another so as to form certain types of cells (cellular differentiation) or carry out specific functions related to spatial and temporal localization; in other words, it makes the genes express themselves when and where necessary. The role of a gene regulatory network therefore consists in integrating the dynamic behaviour of the cell and the external signals with the environment of the cell, and to guide the interaction of all the cells so as to control the process of cellular differentiation (Geard, 2004). Inferring this regulating network from the cellular expression data is considered to be one of the most complex problems in bioinformatics (Akustsu, 1999).

Construction of Phylogenetic Trees

A phylogenetic tree (Setúbal, 1999) is a tree that shows the evolutionary relationships between various species of individuals that are believed to have common descendence. Whereas traditionally morphological characteristics are used to carry out such analyses, in the present case we will study molecular phylogenetic trees, which use sequences of nucleotides or amino acids for classification. The construction of these trees is initially based on algorithms for multiple sequences alignment, which allows us to classify the evolutive relationships between homologue genes present in various species. In a second phase, we must calculate the genetic distance between each pair of sequences in order to represent them correctly in the tree.

Gene Finding and Mapping

Gene finding (Fickett, 1996) basically consists in identifying genes in an ADN chain by recognizing the sequence that initiates the codification of the gene or *gene promoter*. When the protein that will interpret the gene finds the sequence of that promoter, we know that the next step is the recognition of the gene.

Gene mapping (Setúbal, 1999) consists in creating a genetic map by assigning genes to a position inside the chromosome and by indicating the relative distance between them. There are two types of mapping. Physical or *cytogenetic* mapping, on the one hand, consists in dividing the chromosome into small labelled fragments. Once divided, they must be ordered and situated in their correct position in the chromosome. Link mapping, on the other hand, shows the position of some genes with respect to others. The latter mapping type has two inconveniences: it does not provide the distance between the genes, and it is unable to provide the correct order if the genes are very close to each other.

Prediction of DNA, RNA, and Protein Structure

The DNA and RNA sequences are folded into a tridimensional structure that is determined by the order of the nucleotides within the sequence. Under the same environmental conditions, the tridimensional structure of these sequences implies a diverging behaviour. Since the secondary structure of the nucleic acids is a factor that affects the link of both DNA molecules and RNA molecules, it is essential to know these structures in order to analyze a sequence.

The prediction of the folds that determine the RNA structure is an important factor in the understanding of