Evolutionary Data Mining for Genomics

Laetitia Jourdan
University of Lille, France

Clarisse Dhaenens
University of Lille, France

El-Ghazali Talbi
University of Lille, France

INTRODUCTION

Knowledge discovery from genomic data has become an important research area for biologists. Nowadays, a lot of data is available on the web, but the corresponding knowledge is not necessarly also available. For example, the first draft of the human genome, which contains 3×10^9 letters, has been achieved in June 2000, but up to now only a small part of the hidden knowledge has been discovered. The aim of bioinformatics is to bring together biology, computer science, mathematics, statistics and information theory to analyze biological data for interpretation and prediction. Hence many problems encountered while studying genomic data may be modeled as data mining tasks, such as feature selection, classification, clustering, and association rule discovery.

An important characteristic of genomic applications is the large amount of data to analyze and it is, most of the time, not possible to enumerate all the possibilities. Therefore, we propose to model these knowledge discovery tasks as combinatorial optimization tasks, in order to apply efficient optimization algorithms to extract knowledge from large datasets. To design an efficient optimization algorithm, several aspects have to be considered. The main one is the choice of the type of resolution method according to the characteristics of the problem. Is it an easy problem, for which a polynomial algorithm may be found? If yes, let us design such an algorithm. Unfortunately, most of the time the response to the question is 'NO' and only heuristics, that may find good but not necessarily optimal solutions, can be used. In our approach we focus on evolutionary computation, which has already shown an interesting ability to solve highly complex combinatorial problems.

In this chapter, we will show the efficiency of such an approach while describing the main steps required to solve data mining problems from genomics with evolutionary algorithms. We will illustrate these steps with a real problem.

BACKGROUND

Evolutionary data mining for genomics groups three important fields: Evolutionary computation, knowledge discovery and genomics.

It is now well known that evolutionary algorithms are well suited for some data mining tasks and the reader may refer, for example, to (Freitas, 2008).

Here we want to show the interest of dealing with genomic data using evolutionary approaches. A first proof of this interest may be the book of Gary Fogel and David Corne on « Evolutionary Computation in Bioinformatics » which groups several applications of evolutionary computation to problems in the biological sciences, and in particular in bioinformatics (Corne, Pan, Fogel, 2008). In this book, several data mining tasks are addressed, such as feature selection or clustering, and solved thanks to evolutionary approaches.

Another proof of the interest of such approaches is the number of sessions around "Evolutionary computation in bioinformatics" in congresses on Evolutionary Computation. Man can take as an example, EvoBio, European Workshop on Evolutionary Computation and Machine Learning in Bioinformatics, or the special sessions on "Evolutionary computation in bioinformatics and computational biology" that have been organized during the last Congresses on Evolutionary Computation (CEC'06, CEC'07).

The aim of genomic studies is to understand the function of genes, to determine which genes are involved

in a given process and how genes are related. Hence experiments are conducted, for example, to localize coding regions in DNA sequences and/or to evaluate the expression level of genes in certain conditions. Resulting from this, data available for the bioinformatics researcher may not only deal with DNA sequence information but also with other types of data like for example in multi-factorial diseases the Body Mass Index, the sex, and the age. The example used to illustrate this chapter may be classified in this category.

Another type of data deals with the recent technology, called microarray, which allows the simultaneous measurement of the expression level of thousand of genes under different conditions (various time points of a process, absorption of different drugs…). This type of data requires specific data mining tasks as the number of genes to study is very large and the number of conditions may be limited. This kind of technology has now been extended to protein microarrays and generates also large amount of data. Classical questions are the classification or the clustering of genes based on their expression pattern, and commonly used approaches may vary from statistical approaches (Yeung & Ruzzo, 2001) to evolutionary approaches (Merz, 2002) and may use additional biological information such as the Gene Ontology - GO - (Speer, Spieth & Zell, 2004). A bi-clustering approach that allows the grouping of instances having similar characteristic for a subset of attributes (here, genes having the same expression patterns for a subset of conditions), has been applied to deal with this type of data and evolutionary approaches proposed (Bleuler, Prelié & Ziztler, 2004). Some authors are also working on the proposition of highly specialized crossover and mutation operators (Hernandez, Duval & Hao, 2007). In this context of microarray data analysis, data mining approaches have been proposed. Hence, for example, association rule discovery has been realized using evolutionary algorithms (Khabzaoui, Dhaenens & Talbi, 2004), as well as feature selection and classification (Jirapech-Umpai & Aitken, 2005).

MAIN THRUST OF THE CHAPTER

In order to extract knowledge from genomic data using evolutionary algorithms, several steps have to be considered:

1. Identification of the knowledge discovery task from the biological problem under study,
2. Design of this task as an optimization problem,
3. Resolution using an evolutionary approach.

In this part, we will focus on each of these steps. First we will present the genomic application we will use to illustrate the rest of the chapter and indicate the knowledge discovery tasks that have been extracted. Then we will show the challenges and some proposed solutions for the two other steps.

Genomics Application

The genomic problem under study is to formulate hypothesis on predisposition factors of different multifactorial diseases such as diabetes and obesity. In such diseases, one of the difficulties is that healthy people can become affected during their life so only the affected status is relevant. This work has been done in collaboration with the Biology Institute of Lille (IBL).

One approach aims to discover the contribution of environmental factors and genetic factors in the pathogenesis of the disease under study by discovering complex interactions such as [(gene A and gene B) or (gene C and environmental factor D)] in one or more population. The rest of the paper will take this problem as an illustration.

To deal with such a genomic application, the first thing is to formulate the underlying problem into a classical data mining task. This work must be done through discussions and cooperations with biologists in order to agree on the aim of the study. For example, in data of the problem under study, identifying groups of people can be modeled as a clustering task as we can not take into account non-affected people. Moreover a lot of attributes (features) have to be studied (a choice of 3652 points of comparison on the 23 chromosomes and two environmental factors) and classical clustering algorithms are not able to cope with so many features. So we decide to firstly execute a feature selection in order to reduce the number of loci into consideration and to extract the most influential features which will be used for the clustering. Hence, the model of this problem is decomposed into two phases: feature selection and clustering. The clustering is used to group individuals of same characteristics.