Parallel Data Mining of microarray biological data

Roberto Innocente, Carlo Carloni Calame May 2002

Fundamental research has made huge efforts in order to analyze big amounts of data stored in electronic databases, developing the so-called "data mining" techniques. At the same time, in recent years, an enormous quantity of biological data has been produced, by means of some techniques recently developed in order to analyze cellular gene expression. The analysis of such data can not be perfermed with data mining tools and algorithms existing at the moment, because of the high "dimensionality" of these databases. It is therefore mandatory to take some efforts to realize a new parallel version of the data mining algorithms.

1 Data Mining

In the recent past, the amount of data stored in electronic form has became huge. Sales databases, the World Wide Web, biological and pharmaceutical databases are growing up at impressive rate. This is the reason why finding interesting information, which may be hidden, inside the databases has became a challenging task.

The words "Data mining", as well known as "Knowledge Discovery in Databases (KDD)", refer to the process of searching hidden information inside data. Usually, the information consists in a regular behaviour or meaningful sequences inside the data.

Searching for association rules (association rules mining (ARM)) is one of the techniques which was one of the more successful.

This technique allows the discovery of associative relations among some attributes of the data, which were previously unknown. Association rules are statements like: "30 per cent of customers who bought beer bought also wiskhy". More formally, an association rule can be written as the following: $X => Y$ *with support s and confidence c*, where X and Y are sets of binary attributes (itemsets), and the intersection of X and Y is empty. The support s is a measure of the statistical significancy of the rule, i.e. the percentage of rows in the database where both X and Y are contained. The confidence c is the percentage of rows where X and Y are present with respect to the one where only X is present. If any limitation on the data is not imposed, the complexity of the algorithms can grow as the number of the subsets (exponential grow). Furthermore, the dimension of the databases can reach hundreds of Terabytes (WalMart) and this is why a parallel implementation of the algorithms is mandatory.

2 Micro arrays

In the last few years, the technology of "micro-array" (as well known as DNA matrices or DNA chips) has been developed in order to identify the gene expression inside the cells. It exploits the preferential binding which can be formed between complementary single stranded sequences of nucleic acid.

Usually, a microarray is made of a few centimeters squared glass target, on which some nucleotide sequences are attached at fixed spots. A single microarray can hold tenth of thousands of such spots, each one containing a big amount of DNA, the length of which can go from twenty to hundreds of nucleotides. In principle, each DNA molecule should identify a gene of the examined genome.

The microarrays are usually used to compare the gene expression of two sample of cells, e.g. cells of the same kind grown under different environments, or ill and healthy cells.

In those applications, the mRNA coming from the two samples is extracted and it is labelled by means of different fluorescent dyes. The extracted mRNA's are then washed on top of the microarray: according to their own nucleotide sequence, the labelled mRNA's bind to the DNA complementary sequences which are present on the array, allowing thus to compare gene expression of the two cell samples. In fact, each spot on the chip, when excited with a laser, will be fluorescent at different levels, according to the amount and the kind of mRNA which is bounded in that spot. The analysis of microarray is producing huge amount of data, which are stored in high-dimension databases.

The techniques developed for data mining can play a fundamental role in order to search rules and interesting relations inside the data produced by microarray analyses.

3 Association Rules Mining (ARM)

This technique was born in the early 90's. In a fundamental paper by Agrawal, Imielinski e Swami(1993), the concept of association rule is formally defined.

The rules the support of which is greater than a minimum support (as known as minsupp) are said to be frequent. The one the confidence of which is greater than a minimum confidence (minconf) are said to be strong. The ARM (Association Rules mining) attempts to automatically find all the frequent and strong rules, once a minsup and a minconf are defined.

The first efficient algorithm, called "A priori algorithm", was then discussed by Agrawal e Srikant in 1994.

This algorithm belongs to the class of the so-called "breadh-first" algorithms, which find frequent rules which are longer and longer.

It is not possible to use it in order to find association rules bigger than some tenths of attributes, because it requires to go through all the sub-rules of a frequent rule.

In 2000, Han, Pei and Yin proposed a new algorithm, called FP-growth, which exploits a new data structure, a FP-tree (Frequent Pattern tree), and which allows to obtain good results even with rules with some tenth of attributes.

Later, other algorithms have beeen proposed, which do not search for all the frequent sets, but search only for maximal frequent sets (Zaki, 1999). These algorithms have the drawback that, in order to calculate the confidence of the derived rules, they must read again the database to find the support of the sub-sets which are considered.

Finally, in 1999, Pasquier, Bastide et al. proposed an algorithm based on the application of the Formal Concept Analysis by R.Wille. This algorithm requires only to search for a restricted class of the frequent rules (the closed rules), and thus it allows to deal with databases having frequent rules with many attributes.

In 2000, Pei, Han and Mao proposed an algorithm merging the new data structure (FP-tree) and the search of the closed frequent rules.

Our research proposal is the implementation of a parallel algorithm based on the searching for closed frequent rules, using as data structure a prefix-tree.

This approach and eventually a pre-clustering, should fill the gap between the dimensionality of the microarray data and the ability of new algorithms for ARM.

4 Microarray technology

In the last fifty years, research in biology pointed out that life information of all beings is embedded into the DNA. DNA is a complex molecule, the fundamental components of which are the nucleotides, called also bases: A (adenine), T (thymine), C (citosine), G (guanine). This molecule is structured as an double helix, made of two complementary strands: one of the strands can be obtained from the other by replacing A with T and C with G. The information contained inside the DNA is coded by genes, which, over-simplifying, can be thought as strings belonging to a single DNA strand. The genes are responsible for the protein making (ammino-acid sequences), which are the fundamental building blocks and determine both cellular structure and functionality. The translation of genes into proteins proceeds through a complex machinery: a DNA segment necessary to produce a given protein is copied into the mRNA (a single stranded nucleotide sequence), which can move within the cell and is able to translate genes into proteins, with the help of ribosomes and tRNA.

Besides the recognition of the genes, modern biology faces the difficult problem of understanding the complex relations occurring among genes. Indeed, it has been experimetally shown that the activity of one gene is strictly correlated to the activity of the others, according to a dense network of interconnections binding gene expression of a gene (or a class of genes) to the activity of others.

In order these researches to be successful, microarrays are very useful tools. They are producing a real revolution in biological sciences: with old techniques available up to some year ago, it was possible to study the expression of only a gene at a time. The introduction of microarrays allowed to study the expression of thousands of genes at the same time and to get a global and genereal view of the examined cell status.

The biological principle on which microarrays are based is a simple: the preferential binding between complementary nucleotides strands is exploited. In order to understand how they work, it is useful to remind that the presence, inside the cell, of a given mRNA sequence is a "measure" of the activity of the corresponding gene.

Usually, a microarray is made by a glass support (or some plastic material support) as large as few squared inches, on which nucleotides sequences are attached at fixed spots. On a single microarray, there could be tenths of thousands spots, each one containing a big number of DNA molecules, the length of which can go from twenty to hundred of nucleotides. In principle, every DNA molecule should identify a gene of the examined genome.

There are mainly two technologies that have been developed in order to fix nucleotides chains on microarrays: one has been developed by Affymetrix (Santa Clara, California) and the other has been developed at the Stanford University. Affymetrix chips are made using photolitographic techniques to synthesize on the support nucleotides chains of about twenty elements. Indeed, according to the Stanford University standard, the mRNA sequences are directly extracted from organisms (multiplied by means of PCR) and then they are "printed" on the desided spots with a robot. The main advantage of Affymetrix microarrays is that it has a very accurate control on the synthesized nucleotide sequence. On the other hand, Stanford University microarrays have the advantage that they are cheaper and that are produced with an open standard: any lab can produce the microarray in its own according to its needs.

A typical microarray experiment compares the gene expression of two different cell samples: they could be ill and healthy cells or cell grown under different environment conditions, or again cells coming from different tissues. The experiments aim to discover differences and similarities in the gene expression of one sample, taken as reference, and the other. The steps of a typical experiment could be the following:

- the mRNA is extracted from the two cellular samples

- the mRNA coming from the samples is labelled with different chemical dyes. Usually, fluorescent dyes are used, which emit light at different wave length when excited with a laser (for example red and green light)

- then the mRNA's are washed on the microarray: the mRNA sequences bind to the nucleotides chains which are complementary and are present on the array (hybridization). The microarray is then cleaned from the exceeding mRNA which did not hybridize.

- the microarray is excited with a laser. Each spot will appear differently coloured according to the mRNA bound to it: for example, it will be red if only a kind of mRNA has hybridized, or yellow if both mRNA hybridized.

- the measure of the intensity ratio between red and green light for each spot is a measure of the degree of the activity of the gene corresponding to that spot, in the two cell samples.

The data coming from microarray experiments consist then in the value of the light intensity ratio for each spot.

In the last five or six years, microarrays gained an enormous interest and consensus in the research community, as demonstrated by thousands of papers published on the most important scientific magazines (see the references below).

The applications of microarrays in the biology research cover a large range of problems, going from functional biology to medical diagnosis: they are used to study the cell response to drugs, to determine the genetic signature of cancerous cells, to study the time evolution of gene expression during cell life. For example, recently (DeRisi et al.) the whole genome of *Saccaromyces Cerevisiae* (yeast), consisting of about 6000 genes, has been synthesized on a single microarray. Its time evolution from the anaerobic to the aerobic phase was also studied, pointing out the relations among genes and finding classes of genes which follow an analogous behaviour.

One of the most widely used method for the analysis of data produced by means of microarrays is the so-called "hierarchical clustering": it allows to identify in a series of microarrays, containing the same genes, which are the genes behaving in a similar way. The microarray series could represent a series of experiments performed at different times during cell evolution or cell samples processed with different drugs.

In order to extract the most complete and meaningful information from biological data produced by means of microarrays, data mining techniques could be very helpful. Indeed, the problem shows an high "dimensionality" (each microarray contains tenth of thousands of spots), and finding interesting information and relations in the database becomes very soon prohibitive.

5 Research originality

To the best of our knowledge, this should be the first real application of ARM techniques to microarrays data.

The ARM could be used in order to:

1. find rules being able to strongly distinguish ill cells from healthy ones (finding frequent rules in ill and healthy cells and finding the maximal differences)
2. find episodes characterizing the cells transition from one condition to another (in the case a time series of microarray is considered for the same population)
3. find rules which identify cellular classes.

Furthermore, this implementation of the ARM algorithm should be the first parallel implementation of an ARM algorithm for searching closed rules.

References:

The papers published about the microarray technology are very numerous, as the ones showing scientific results obtained with this technology. As an example, we give the exact reference to a volume of Nature Genetics devoted to microarrays and the paper by DeRisi et al. about the aerobic evolution of the *Saccaromyces Cerevisiae*:

1) Nature Genetics, volume 21 supplement, 1999

2) DeRisi et al., Science, volume 278, October 1997

The web is of course a huge source of information. Among many web sites devoted to the microarrays and to biological results obtained with them, very interesting are the following:

1) genome-www5.stanford.edu/MicroArray/SMD/

2) www.ebi.ac.uk

3) www.microarrays.org

4) cmgm.stanford.edu/pbrown/

5) www.affymetrix.com

For Data Mining, recently many books have been published and a lot of material is available on the web. Concerning books, it may be useful to cite:

1. Principles of data mining, *Mannila,Smyth,Hand*
2. Data mining : Concepts and techniques,*Han, Kamber*

and concerning web sites :

1. www.idagroup.com
2. www.kdunggets.com
3. http://www.almaden.ibm.com/cs/quest/