Gene Expression—a New Challenge for Data Mining

Marko Salmenkivi

Department of Computer Science, University of Helsinki

P.O. Box 26 (Teollisuuskatu 23)
FIN-00014 University of Helsinki, Finland

E-mail: Marko.Salmenkivi@cs.helsinki.fi

Abstract

The need of automatic analysis methods is increasing rapidly in biosciences, e.g. genetics. Gene expression, i.e. the process by which the DNA information is transmitted to a cell, is one of the most challenging research topics of the coming decades. Research on gene expression means finding out the instructions of regulation mechanisms in cells. Data mining research develops methods for detecting regularities in large amounts of data. A short overview of computer intensive statistical methods, e.g. Monte Carlo, randomization and bootstrap, is given. These methods are powerful tools for many application areas. Their significance is still increasing, with the increasing performance of computers. Finally, applying randomization and Monte Carlo-testing to the analysis of a yeast gene expression and gene classification data is illustrated.

Contents

1 Introduction

Recent advances in genetics have led to a situation where a large amount of raw data is being produced using new techniques. In the near future the trend seems to be even stronger; automatic analysis methods are thus clearly becoming necessary. That means increasing the need for information technology and the methods of computer science in biosciences. For computer scientists, biosciences, e.g. genetics, mean a very interesting and challenging application area. The new field bioinformatics applies the methods of information science, computer science and computer engineering to biological, especially genetic, data.

One of the most significant research areas during the coming decades in the area of bioinformatics is the research on the regulatory mechanisms of a cell.

In Section 2 some basic concepts of genetics and gene expression are presented. Section 3 is a brief overview of data mining and in particular, computer-intensive statistical methods, a group of methods useful in data mining. Section 4 contains an example of research applying computer-intensive statistical methods to a yeast gene expression and gene classification data.

2 Genes—Basic Concepts

A gene is a unit of biological information. Human cells, for instance, contain 100,000 genes in 23 chromosomes. The information here basically means a set of instructions for synthesizing proteins at the correct time and in the correct place.

The information in a gene is coded in a DNA molecule. The structure of a DNA molecule is a double helix: two polynucleotide strands are wrapped around each other. The strands are sequences of four different bases: *guanine*, *cytosine*, *adenine*, *thymine*. Guanine and cytosine form one complementary pair, adenine and thymine another one. The parallel strands always contain the complementary bases.

Triplets of bases (*codons*) encode amino acids. Because there are only 20 different amino acids, most of them are encoded by several triplets. For example, triplets ACC, ACA and ACG all specify the amino acid *threonine* [14].

The order of the bases of a genome is a significant research task. The complete sequence of the human genome will be available in the near future. The next step is to try to find out the role of each gene. The functional analysis of the genome as a whole still remains the question, which will probably take a long time to be answered exhaustively. The functions of an individual gene and the whole genome are linked with gene expression; knowing when and where a gene is expressed provides clues as to its biological role [2].

3 Gene expression

Gene expression is the process by which the DNA information is transmitted to a cell. It can be divided into two stages: transcription and translation.

During transcription the strands of the DNA molecule separate from each other. The complementary RNA molecule is constructed by attaching the complementary bases to one of the separated strands. In this way the DNA information is copied. The messenger RNA molecule transports the information to a cell, where amino acids are produced according to the information. This is the translation stage. Amino acids are still used as elementary parts of polypeptide molecules in the synthesis process, which is directed by RNA molecules.

All the genes of an individual are present in all the cells, but only some of them are active, i.e. expressed. Expression patterns determine the characteristics of a cell; if different genes are expressed in two cells, the cells belong to different cell types. Abnormal expression patterns may be associated with the development of tumors. For more detailed information, see [14].

Microarray technique One of the advanced methods developed recently for the analysis of gene expression is the microarray technique [1, 3, 11]. Using this technique thousands of individual gene sequences can be printed in a high-density array on a glass microscope slide. The relative intensities of indicator chemicals are measured at each array element. In Finland the equipment and knowledge of the microarray technique are now available at Turku Centre for Biotechnology.

The investigation of gene expression means a very interesting and challenging long-term research process; finding out the regulation mechanisms of the cells. The amount of expression data is increasing rapidly and consequently automatic analysis methods are needed.

4 Data mining and computer-intensive statistical methods

Data mining (or knowledge discovery from databases, as the area is also called) aims at developing methods for discovering interesting regularities and exceptions in large data collections. Data mining combines viewpoints from database research, statistics and machine learning. This kind of approach became significant at the early 1990's when the amount of raw data available in business, as well as scientific fields, was increasing very quickly [9].

Data mining is in practice an iterative and gradually sharpening process [8]. Data mining methods may produce a huge number of regularities that still need to be postprocessed. Different kinds of methods are thus applied at different stages of the process. Expert knowledge from the application area is needed at all stages of the research process: when planning suitable approaches, directing the search of regularities, and when interpreting the results.

In the following subsection a brief overview of a group of useful methods, called computer-intensive statistical methods, will be given.

4.1 Computer-intensive statistical methods

Increasing computational power has enabled new kind of approaches in statistics. The computer-intensive methods are challenging the traditional statistical methods and moving the emphasis in statistics by overcoming some restrictions and difficulties of the traditional approaches.

In classic statistical inference the essential task is to derive the sampling distribution for a statistic, and then to calculate the probability of a sample statistic. This procedure has some severe problems. Firstly, for many statistics there are no analytical distributions. Secondly, sampling distributions usually rely on restricting assumptions.

Computer-intensive statistical methods approach the problem from a different view. They *simulate* the sampling process and do not demand troubling assumptions about population distributions. An overview of computer-intensive statistical methods is given in [4].

Monte Carlo simulation If a population distribution is known, Monte Carlo simulation can be used to generate an empirical distribution for a test statistic in the population. This is useful in many situations, where the theoretical sampling distribution of a statistic is not known. In practice, Monte Carlo simulation typically proceeds as follows. A large number of sample sets from the population distribution are generated and the desired test statistic is calculated from each set. Using these values, an empirical distribution of the statistic is constructed. The original, real value of the statistic is compared with the empirical distribution. Finally, conclusions are drawn on the basis of the comparison.

Markov chain Monte Carlo (MCMC) Sampling from very complex distributions is usually not possible by using classic simulation methods, because independent values cannot be generated due to the high dimension and complex dependency structure of such distributions. Markov chain Monte Carlo methods are powerful tools for generating samples from distributions with even thousands of parameters dependent on each other. MCMC methods are Monte Carlo methods, simulation being used to construct an empirical approximation of the target distribution, as described above [6].

Unlike in the case of the classic methods, successive values generated by the MCMC methods are not independent. If the densities of all values, or at least the proportions of the densities of all the pairs of values, can be calculated, these proportions can be used to locally determine the "right" acceptance-rejection-ratio for the generated values, i.e. the accepted values are generated according to the target distribution. Because of the dependency of the values it is usually necessary to generate more values than when using classic simulation methods. For theoretical understanding of the MCMC methods, [12] gives a good overview.

MCMC methods have particularly significantly contributed to Bayesian statistics. Bayesian data analysis typically leads to complex integrals that cannot be solved analytically. Approximate integration using MCMC methods is usually possible in this kind of problems [7, 13].

Bootstrap Sometimes no other kind of information but one sample from a population may be available. If the sample is representative of the whole population, it is still possible to draw conclusions about the population parameters using bootstrap methods combined with Monte Carlo simulation.

Bootstrap methods are based on resampling from the available sample. Values from the sample are drawn using replacement as if the sample values were from the real population.

If the sample is representative of the population, using bootstrap methods gives almost as good results as the classic approaches, with some exceptions [5].

Randomization Unlike the bootstrap methods, randomization tests cannot be used to draw any conclusions about the populations behind the samples. Still they are very useful in many situations, where testing hypothesis about samples instead of populations is sufficient. A typical randomization procedure is described in the following section.

5 Example: analysis of yeast genome expression data

In this section an example of applying Monte Carlo simulation and randomization to a yeast genome expression data is given.

Two kinds of data were used: a functional classification of the genes of the yeast *Saccharomyces cerevisia* and a data set containing expression levels of the yeast genome resulted from a trial described below. The goal of the research was to consider the relevance of the classification in the light of this particular expression data.

In the following the data sets and the research process from a methodological point of view are described. The results and their biological interpretations are mostly passed.

5.1 Data

Expression data The data set was first introduced by [2] and it is publicly available. The data contains expression levels of 6,154 genes, i.e. almost the whole genome, of the yeast *Saccharomyces cerevisia.* The relative changes of the indicator chemicals were measured at seven time points with two hour intervals during a diauxic shift, that is, switch from anaerobic (fermentation) to aerobic metabolism (respiration). The microarray technique was used in producing the data.

Yeast genome classification data A functional classification of the yeast genes was received from Munich Information Center for Protein Sequences [10]. The classification is hierarchical, consisting of 14 main classes (e.g. Metabolism, Energy, Transcription, Protein synthesis, Transport facilitation) and several levels of subclasses. An individual gene can belong to several classes. There are more than 2,500 genes that are still unclassified or of which classification is not clear-cut.

5.2 Distance measures

Distance measures between two profiles and between two classes of genes are needed to allow comparison between expression profiles and between classes.

Let $x_i = x_i(0), \ldots, x_i(k)$ and $y_j = y_j(0), \ldots, y_j(k)$ be time series (expression profiles of two genes). The distance d between x_i and y_j was defined as the sum of the euclidean distances at each time point:

$$d(x_i, y_j) = \sum_{t=0}^{k} |x_i(t) - y_j(t)|^2 \ .$$

For comparison of the classes in respect to the expression, a distance measure is also needed between two classes.

Let $c_1 = \{x_1, \ldots, x_n\}$ and $c_2 = \{y_1, \ldots, y_m\}$ be clusters of time series.

The distance d_{inter} between two classes c_1 and c_2 (intercluster distance) was defined as the sum of distances between all the profile pairs divided by the number of the pairs:

$$d_{inter}(c_1, c_2) = \frac{1}{n \cdot m} \sum_{i=1}^{n} \sum_{j=1}^{m} d(x_i, , y_j)$$

and tightness d_{intra} of a class c_1 (intracluster distance), respectively:

$$d_{intra}(c_1) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} d(x_i, x_j) \ .$$

5.3 Randomization of classification

The relevance of classification was investigated by randomizing the classes according to the following procedure.

First, distances were calculated from the data between all the subclasses. The next stage was the randomization: as many genes were picked up randomly into each class as there were in the original classification. The randomization procedure was repeated 1,000 times and the distances were computed for each randomization. The final stage was the Monte Carlo testing; the values computed from the real classification were compared with the empirical distributions of the distance values calculated from the random classifications.

5.4 Results

Next some preliminary results of the analysis are given to illustrate the method used.

The empirical distributions of the intracluster distances for the subclasses 5 (Fermentation) and 8 (Oxidation of fatty acids) of the main class Energy, and the corresponding 95% confidence intervals are shown in Figure 1. The corresponding values calculated according to the real classification data are indicated by the points on the x-axes.

The results indicate that subclasses 5 and 8 are significantly tight. On the other hand these classes are also close to each other, the empirical distribution and the data value for the intercluster distance of the subclasses being shown in Figure 2.

For comparison the distributions of the intracluster distance of subclass 1 and the intercluster distance of subclasses 1 and 5 (Fig. 3) are shown. Here the distance values calculated from the data do not significantly deviate from the random values, i.e. the distributions of the values computed on the basis of the random classifications.

It is important to notice that discovering statistical significance does not straightforwardly guarantee the existence of biological significance. Hence, experts on genetics are necessarily needed when interpreting the results. Because the aim here is to illustrate the method, the question of the biological importance of the given results is ignored.

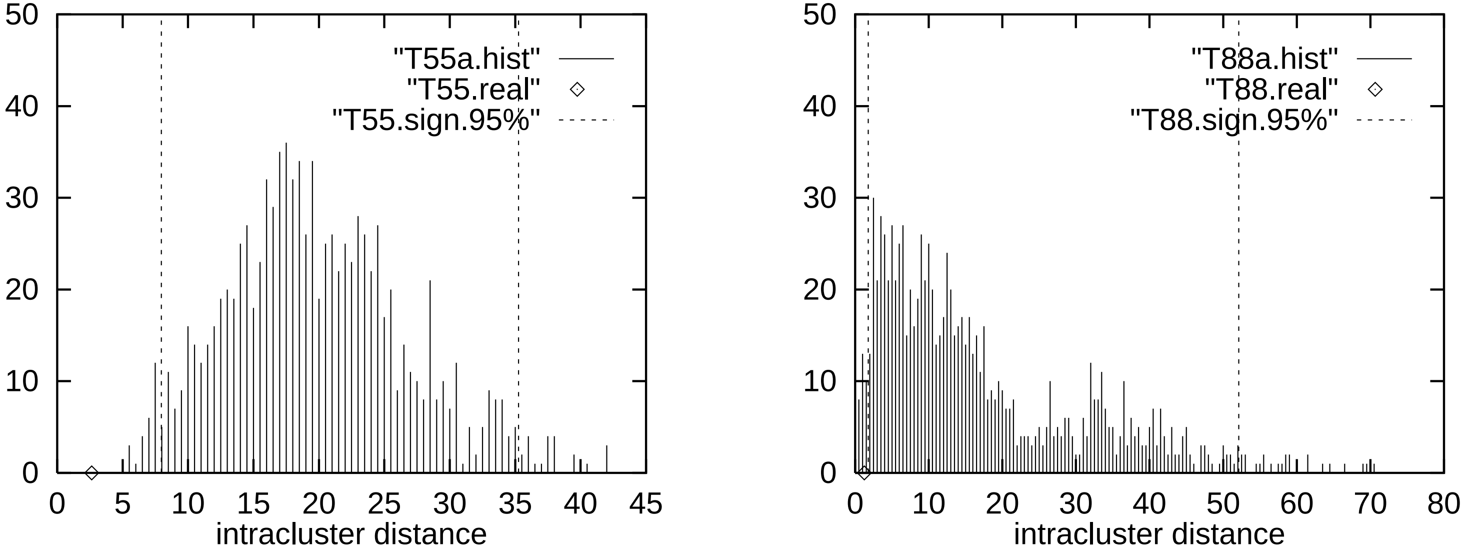

Figure 1: *Empirical distributions of the intracluster distances: subclasses 5 (left) and 8 (right)*

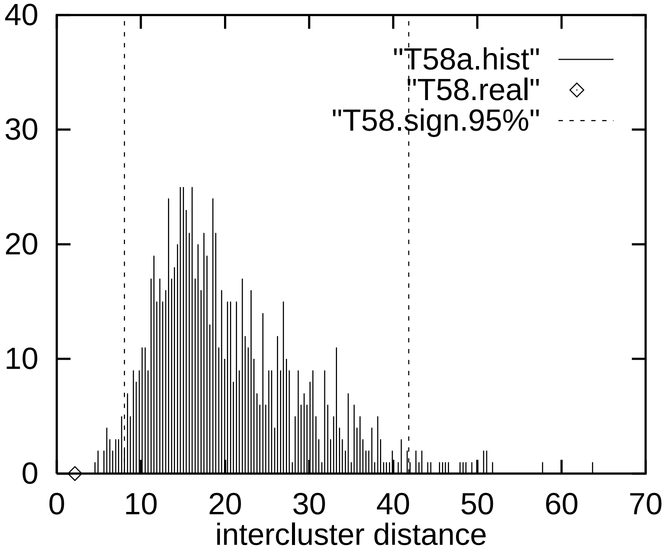

Figure 2: *Empirical distribution of the intercluster distance between subclasses 5 and 8*

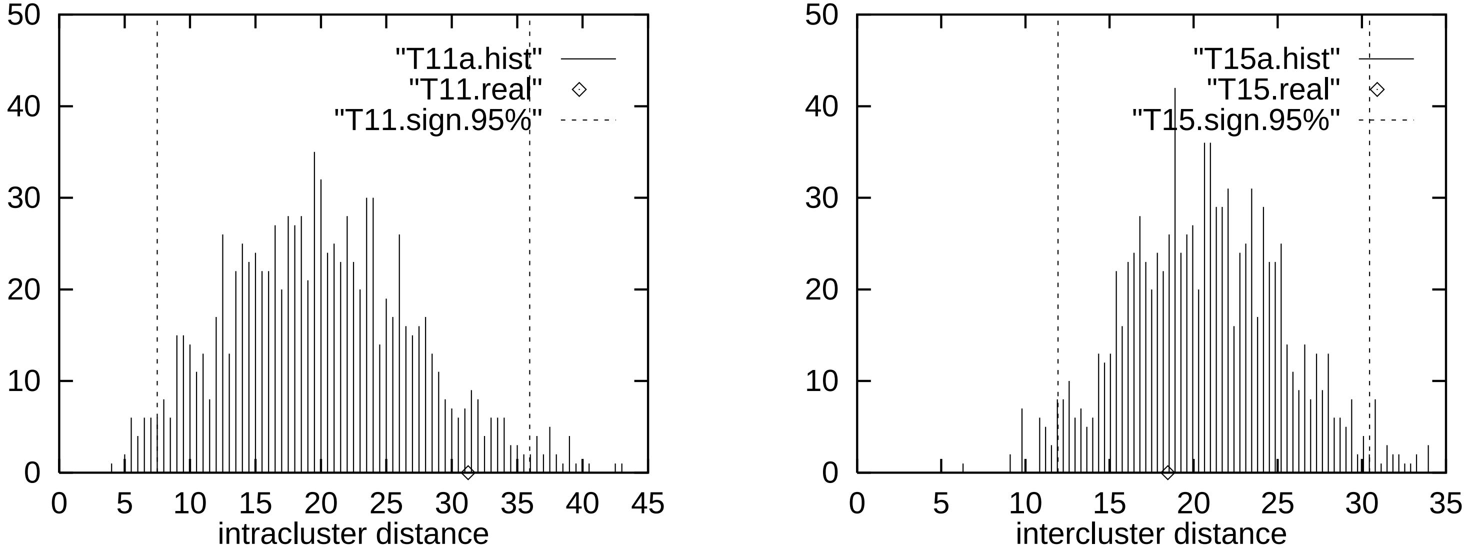

Figure 3: *Empirical distributions of the intracluster distance of subclass 1 (left) and intercluster distance between subclasses 1 and 5 (right)*

6 Conclusions

Co-operation of experts in computer science and biosciences is needed as the amount of data in biosciences is significantly increasing due to new research techniques. The process of decoding and making use of genetic information code (gene expression) is one of the most fascinating and demanding future challenges and in the future a huge amount of raw data from the area will be available.

Data mining is an area of computer science playing an essential role in the development of methods for these application areas. Computer-intensive statistical methods provide a powerful set of tools for the modern data analysis; they can avoid some crucial problems of the traditional statistical inference by simulating the sampling process.

The example research process on the relevance of a yeast functional classification in respect to a yeast genome expression data set illustrates how randomization and Monte Carlo methods can be applied. Empirical distributions for the interesting statistics can be generated and used to test the significance of the values from the data without assumptions about the population distributions. Statistical significance does not, however, guarantee biological significance. Hence, expert knowledge is needed to interpret the results.

Acknowledgements

I am grateful to Professor Heikki Mannila for the example in Section 4. Most of the ideas presented were planned in co-operation with him.

References

[1] P. O. Brown and D. Botstein, *Exploring the new world of the genome with DNA microarrays.* Nature Genetics, 21, pp. 33–37, 1999.

[2] J. L. DeRisi, V. R. Iyer, and P. O. Brown, *Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale.* Science, 278, pp. 680–686, 1997.

[3] V. G. Cheung, M. Morley, F. A, A. Massimi, R. Kucherlapati, and G. Childs, *Making and reading microarrays.* Nature Genetics, 21, pp. 15–19, 1999.

[4] P. R. Cohen, *Empirical Methods for Artificial Intelligence.* The MIT Press: 1995.

[5] B. Efron, *An introduction to the bootstrap.* Chapman and Hall: 1993.

[6] W. K. Hastings, *Monte Carlo sampling methods using Monte Carlo computation and Bayesian model determination.* Biometrica, 57, pp. 97–109, 1970.

[7] A. Gelman, J. Carlin, H. Stein, D. Rubin, *Bayesian Data Analysis.* Texts in Statistical Science. Chapman & Hall, 1995.

[8] M. Klemettinen, H. Mannila, and H. Toivonen, *A data-mining methodology and its application to semi-automatic knowledge acquisition.* Proceedings of the 8th International Conference and Workshop on Database and Expert Systems Applications (DEXA'97), pp. 670–677, Toulouse, 1997.

[9] H. Mannila, *Data mining: machine learning, statistics, and databases.* Eight International Conference on Scientific and Statistical Database Management, pp. 1–8, Stockholm, 1996.

[10] Munich Information Centre. `http://www.mips.biochem.mpg.de`

[11] S. G. Penn, D. R. Rank, D. K. Hanzel, and D. L. Parker, *Mining the human genome using microarrays of open reading frames.* Nature Genetics, 26 (3), pp. 315–318, 2000.

[12] L. Tierney, *Markov chains for exploring posterior distributions.* Annals of Statistics, 22 (4), pp. 1701-1728, 1994.

[13] H. Toivonen, H. Mannila, M. Salmenkivi, and K-P. Laakso, *Bassist—a tool for MCMC simulation of statistical models.* Proceedings of the Eurosim '98 Simulation Congress, pp. 590–595, Helsinki, 1998.

[14] P. C. Winter, G. I. Hickey, and H. L. Fletcher, *Instant Notes in Genetics.* Bios Scientific Publishers: 1998.