Computational Molecular Biology - an overview

Torgeir R. Hvidsten*
torgeihv@idi.ntnu.no

Norwegian University of Science and Technology

Abstract

Genomic research is experiencing a shift of paradigm. Old reductionistic research methods are largely replaced by methods and technology that make possible conducting global studies of gene products and their interactions in complex networks in living organisms. Computational data analysis and modelling have become an important and necessary part of this research methodology. In this paper we look at some of the most commonly used methods for data analysis and modelling, and discuss some of the most important publications taking these methods into use in real-world biological studies.

Keywords: Computational Biology; Data Analysis; Class Discovery; Class Prediction; Gene Expression Data

1 Introduction

Stevenson [17] defines computational science as *an interdisciplinary approach to doing science on computers*. Many important challenges within science today imply computational modelling either because the experiments are impossible to conduct in the real world (such as in astronomy or meteorology) or because the experiments generate such a vast amount of data that interpretation becomes a problem. One example of the latter science is genomic research, where new experimental methods are generating such overwhelming amount of data that it seems impossible to neither handle nor interpret it without the help of computers. Consequently, the field of *computational biology* is growing rapidly. Computational biology can be viewed as *a general approach toward the solution of scientific problems through which advanced computational techniques are used to discover the hidden order in complex data sets and to decipher the languages of biology* [12].

One of the main goals of genomic research is to understand the multiple biological functions of gene products and their interaction in complex networks in living organisms. Another main goal is to understand the relation between this molecular world and various common diseases. With the scarce and fragmented status of the present knowledge, this is an enormous challenge. Microarray technology [14], however, give a view into the functioning of molecular biological systems by simultaneous readouts of tens of thousands of genes. The main problem has now become how to utilise this data together with the present knowledge to build models of biological interest.

2 Molecular Biology

The information macromolecule DNA is used to synthesis proteins which play important roles in many cellular functions as enzymes, receptors, storage proteins, transport proteins, transcription factors, signalling molecules and hormones. DNA serves as templates for the transcription of RNA. RNA transfers the genetic information from the nucleus where the DNA is stored to the ribosomes where RNA is translated into protein. A fragment of DNA used to synthesis a protein is called a *gene*. The DNA $\rightarrow$ RNA $\rightarrow$ protein flow of genetic information is called the *central dogma* of molecular biology.

A gene that is transcribed and used in protein synthesis is said to be expressed, and a key element in the dynamics of cellular processes is the regulation of gene expression. Since most changes in protein levels result from changes in RNA levels, quantitative measurements of RNAs provide important clues to how genes act together in complex biological systems. Microarray experiments provide us with these quantitative measurements.

- *Functional studies*: In functional microarray studies the goal is to study the function of gene products. Measuring temporal changes in gene expression throughout the course of a given biological response is especially relevant in these studies, since the collected data may reflect the temporal dynamics of gene interactions. Hence, a typical functional study aims to model the relationship between gene expression as a function of time and gene function.

- *Clinical studies*: In clinical microarray studies the goal is to compare the expression level of genes in for example healthy cells and cancerous cells. Hence, a typical clinical study aims to model the relationship between the gene expression levels in samples taken from patients and a set of clinical states assigned to these patients.

3 Methods for Analysing Microarray Data

A set of p different microarray experiments produce a vector $\boldsymbol{x} = [x_1, x_2, ..., x_p]'$ of measurements for each gene. If n genes are observed, the entire data set can be represented as the set:

$$X = \{\boldsymbol{x}_i \mid i = 1, 2, ..., n\} \tag{1}$$

* http://www.idi.ntnu.no/~torgeihv

Alternatively, a set of p genes determines a vector $\boldsymbol{x} = [x_1, x_2, ..., x_p]'$ of measurements for each experiment. In any case, we will assume that each observation $\boldsymbol{x}_i \in X$ is drawn from a population with m underlying classes $\boldsymbol{\pi} = \{\pi_i \mid i = 1, 2, ..., m\}$. We say that the observations are labelled if there exists a class vector $\boldsymbol{c} = [c_1, c_2, ..., c_n]'$ such that each observation $\boldsymbol{x}_i$ is associated with one class $c_i \in \boldsymbol{\pi}$.

Techniques used to extract biological knowledge from expression matrices can conceptually be divided into *class discovery* methods and *class prediction* methods. In machine learning these methods are called unsupervised and supervised learning, respectively.

Class discovery methods consider unlabelled data and seek to discover the underlying classes by clustering genes or experiments with similar patterns of expression together. The hypothesis is that these clusters reflect the underlying unknown classes. In fact, it has been shown that genes coding for proteins with similar functions do tend to have similarities in their expression profiles [8, 5]. Correspondingly, it has been shown that samples from patients with the same type of cancer may be clustered together [6, 1].

Class prediction methods use a set of labelled genes or experiments to induce a model from these examples, defining the relationship between gene expression and classes.

3.1 Class Discovery Methods

Given a set of observations $X = \{\boldsymbol{x}_i \mid i = 1, 2, ..., n\}$, there are

$$\frac{m^n}{m!} \tag{2}$$

possible partitions of these observations into m nonempty subsets (clusters). Clearly, any exhaustive search procedure can be ruled out. For this reason a number of algorithms finding reasonable clusters without looking at all configurations has been developed.

Fundamental to all clustering algorithms are similarity measures. Given two observations $\boldsymbol{x} = [x_1, x_2, ..., x_p]'$ and $\boldsymbol{y} = [y_1, y_2, ..., y_p]'$, the most commonly used similarity measure is the Minkowski metric:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \left[\sum_{i=1}^{p} |x_i - y_i|^k\right]^{1/k} \tag{3}$$

For $k = 1$, (3) becomes the *city-block* distance, while for $k = 2$ it becomes the Euclidean (straight-line) distance. In general we want a similarity measure d such that if the class vector $\boldsymbol{c} = [c_1, c_2, ..., c_m]'$ is known, d would have the following property:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \begin{cases} \text{“large”} & \text{when } c_i \neq c_j \\ \text{“small”} & \text{when } c_i = c_j \end{cases} \tag{4}$$

Of course, the reason why we want to do class discovery in the first place is because the class vector $\boldsymbol{c}$ is unknown, and hence we can only hope that the underlying biological phenomena behave according to the assumption that similar observations belong to the same class.

Clustering algorithms are normally divided into *hierarchical* and *non-hierarchical* methods, although, in statistics, one also distinguish between *parametric* and *non-parametric* algorithms. In general, parametric methods make assumptions about the statistical distribution of the data, while non-parametric methods do not. The reader should consult e.g. [13, 9] for further reading on clustering.

Hierarchical Clustering Methods

Hierarchical clustering methods proceed by either a series of successive merges or a series of successive divisions. The former strategy is called *agglomerative hierarchical clustering*, while the latter is called *divisive hierarchical clustering*. Both methods produce a binary tree called a *dendrogram*. We shall view a dendrogram as a graph $D = (V, E)$, where $V = \{1, 2, ..., k\}$ is a set of nodes (vertices) each corresponding to a set of observations (i.e. a cluster) and E is a binary relation on V such that $(U, W) \in E$ if and only if $W \subset U$. An algorithm for agglomerative hierarchical clustering includes the following steps:

1. Add n nodes to V in $D = (V, E)$, each corresponding to a cluster with one observation $\boldsymbol{x} \in X$, and let E be empty. Initiate the similarity matrix $\underset{(n \times n)}{\boldsymbol{S}} = \{d_{ij} \mid 1 \leq i \leq j \leq n\}$ determined by the similarity measure d, i.e. $d_{ij} = d(\boldsymbol{x}_i, \boldsymbol{x}_j)$.
2. Search the similarity matrix for the most similar pair of clusters U and W.
3. Merge clusters U and W, and label the newly formed cluster (UW). Update the similarity matrix by (a) deleting the rows and columns corresponding to clusters U and W and (b) adding a row and a column that define the similarity between (UW) and the remaining clusters. Update the dendrogram D by adding (UW) to V and $\{((UW), U), ((UW), W)\}$ to E.
4. Repeat Steps 2 and 3 a total of $n - 1$ times.
5. Return D.

Note that this algorithm requires a strategy for calculating $d(U, W)$ when U and W are sets of observations and not single observations. If the similarity between two clusters U and W is defined as the similarity between the most similar pair of observations $(\boldsymbol{x}, \boldsymbol{y}) \in U \times V$, then the strategy is called *single linkage*. If the similarity between two clusters U and W is defined as the similarity between the most dissimilar pair of objects $(\boldsymbol{x}, \boldsymbol{y}) \in U \times V$, then the strategy is called *complete linkage*. Finally, if the similarity between two clusters U and W is defined as the average similarity between all pairs of objects $(\boldsymbol{x}, \boldsymbol{y}) \in U \times V$, then the strategy is called *average linkage*.

The agglomerative hierarchical algorithm is clearly an $O(n^2)$ method. However, performing divisive hierarchical clustering with a corresponding strategy would be equivalent to performing an exhaustive search. Note that such a strategy would require a $(2^{n-1} - 1) \times (2^{n-1} - 1)$ initial similarity matrix. Also, there would be no way of reusing the matrix from one iteration to the next. Consequently, non-hierarchical clustering methods are used to split each cluster in two when performing divisive hierarchical clustering.

Non-hierarchical Clustering Methods

Non-hierarchical clustering methods start with an initial set of seed points (centroids) and iteratively update these points until a stable configuration is reached. Consequently, the number of clusters is specified in advance.

k-means clustering is composed of the following simple steps:

1. Choose the number of classes m.
2. Select an initial set of m centroids $\{\boldsymbol{\mu}_i \mid i = 1, 2, ..., m\}$
3. Use the similarity measure d to assign each observation $\boldsymbol{x} \in X$ to the nearest centroid. Observations whose nearest centroid is $\boldsymbol{\mu}_i$ belong to cluster X_i.
4. Recompute the centroids $\{\boldsymbol{\mu_i} \mid i = 1, 2, ..., m\}$ using the clusters from step 3.
5. IF non of the centroids changed in step 4,
   THEN return the clusters $\{X_1, X_2, ..., X_m\}$,
   ELSE repeat step 3 and 4.

Typically, the non-hierarchical clustering algorithms like k-means and the related self-organising maps algorithm [10] are $O(kn)$ methods, where k is the number of iterations before a stable configuration is found. Also, these methods avoid storing a similarity matrix, and hence they have a small memory usage compared to the agglomerative hierarchical clustering method.

3.2 Class Prediction Methods

The goal of class prediction methods is to model the relationship between measured data X and class knowledge $\boldsymbol{c}$. Machine learning methods for this purpose include *decision trees*, *artificial neural networks*, *k-nearest neighbour learning*, *Bayesian learning*, *genetic algorithms*, *rule learning* (see [11] for an overview) and *support vector machines* [4]. In statistics, a parametric method called *Fisher classification* (see e.g. [9]) is quite commonly used.

Support Vector Machines

In its simplest linear form, a support vector machine is a hyperplane that separates a set of positive examples from a set of negative examples with maximum margin. Let $X = \{\boldsymbol{x}_i \mid i = 1, 2, ..., n\}$ be a set of observations, and let $\boldsymbol{c} = [c_1, c_2, ..., c_n]'$ be a class vector such that $c_i = 1$ if $\boldsymbol{x}_i$ is member of the class to be learned, and $c_i = -1$ otherwise. The goal is to learn a set of weights $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_n]$ such that the discriminant function

$$L(\boldsymbol{x}_i) = \sum_{j=1}^{n} c_j \alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{5}$$

is optimised over a set of labelled observations. $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is a kernel function which form depends on the application. The predicted class of a new observation $\boldsymbol{x}$ is given by the sign of the discriminant function $L(\boldsymbol{x})$ computed using the optimised weights.

Typical for supervised methods like support vector machines is their ability to induce models that can classify. By inducing a model from a subset of X (called the *training set*) and using it to classify the remaining observations (called the *test set*), this can be utilised to objectively evaluate the model. A systematic approach to dividing observations into training sets and test sets is k-fold cross validation. In this scheme X is divided into k disjoint equally sized subset. A model is induced from $k - 1$ subsets and tested on the remaining subset. This is repeated for each of the k subsets, allowing us to average the classification performance from each iteration to obtain an unbiased performance estimate for the algorithm on the relevant data. It is common to interpret the cross validation performance estimate as the performance we will get when using a model induced from the full set of data X to classify previously unclassified observations.

A performance estimate indicates the predictive capabilities of a supervised algorithm. If the performance is at least significantly better than what would be expected by chance, it follows that the model has captured something that is essential in order to understand which observations belong to which classes. Many machine learning methods, such as rule learners, induce models that are legible, and hence these methods can help humans better understand the data.

4 Selected Functional Studies

CLASS DISCOVERY: The Transcriptional Program in the Response of Human Fibroblasts to Serum

Iyer et al. [8] studied the human fibroblast's response to serum. These cells have a pivotal structural role in connective tissue and in important processes such as wound healing. The mRNA level of 8613 human genes probes were measured at 12 time points from 0 minutes to 24 hours after serum stimulation. A subset of 517 gene probes whose expression changed substantially in response to serum was selected for further analysis.

This study is typical for functional microarray studies in that an agglomerative hierarchical clustering method is used. Ten clusters were selected from the resulting dendrogram and expression profiles of genes from the same cluster were plotted together to show that they in fact had similar expression profiles. Genes which are known to code proteins with the same function were also plotted together, and some similarities within these groups were apparent. However, no mapping between the functional classes and the clusters were attempted, and hence nothing could be said neither about the quality of the clusters nor about the function of previously unclassified genes.

CLASS PREDICTIONS: Knowledge-based Analysis of Microarray Gene Expression Data by Support Vector Machines

Very few studies have been published using supervised methods in functional microarray studies. Brown et al. [3], however, used 2467 yeast genes to train support vector machines that could recognise six different functional classes containing 230 genes.

The study used 3-fold cross validation for performance estimation. Support vector machines with four different kernels were compared with a Fisher classification method and two different decision tree methods. Good performance was reported for five of the six classes. These five classes were already known to cluster well using hierarchical clustering. The study includes a discussion of the genes which are wrongly classified by more than one method, and suggests that some of these genes should undergo further biological experiments. The study also includes a list of predictions for 15 previously unclassified genes.

5 Selected Clinical Studies

CLASS DISCOVERY: Distinct Types of Diffuse Large B-cell Lymphoma Identified by Gene Expression Profiling

Despite the large variety of parameters used to classify human malignancies today, patients receiving the same diagnosis can have quite different treatment responses. Alizadeh et al. [1] studied diffuse large B-cell lymphoma (DLBCL) where 40% of the patients respond well to current therapy and have prolonged survival, whereas the remainder succumb to the disease. 128 microarray experiments were conducted, measuring the gene expression levels of over 18 000 human gene probes in 96 normal and malignant patient samples.

Alone et al. [2] introduced the notion of two-dimensional clustering, in which both gene vectors and experiment vectors are organised by clustering. Alizadeh et al. performed hierarchical clustering on both patients and genes. The clinical samples were examined with respect to six gene clusters reflecting relevant biological knowledge. Three branches of the patient dendrogram captured all but three of the DLBCL samples. By re-clustering the DLBCL samples using only the genes from a cluster called 'Germinal Centre B cell', two sub-clusters emerged. These two sub-clusters where named 'GC B-like DLBCL' and 'activated B cell DLBCL', and it was shown that these patients had significantly different expected survival time. The study therefore concludes that based on molecular differences these two sub-clusters should be regarded as distinct diseases.

CLASS PREDICTION: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

Golub et al. [6] studied acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) of 38 bone marrow samples (27 ALL, 11 AML) using 6817 human gene probes. Distinguishing ALL from AML is critical for successful treatment.

The study used neighbourhood analysis to select the 50 genes most correlated with the ALL-AML classes. These genes were denoted *informative genes*. New samples were classified by using these informative genes to cast a 'weighted vote' for one of the classes. The magnitude of each vote depended on the expression similarity between the new sample and the informative gene and the degree of correlation between the informative gene and the ALL-AML classes. The votes were summed to determine the winning class. Using cross validation, 36 of the 38 samples were correctly classified.

Golub et al. also used the class discovery method of self-organising maps to test whether the two classes could have been found automatically if the AML-ALL distinction were not already known. Using two initial seed points the algorithm found one cluster in which 24 of 25 samples were ALL and another cluster in which 10 of 13 samples were AML. Furthermore, using four seed point the algorithm found four clusters largely corresponding to AML and three known subclasses of ALL. The authors conclude that it would have been possible to discover these subclasses of leukemias without the present knowledge, and hence it should also be possible to discover presently unknown subclasses using the same methodology.

6 Discussion and Conclusions

Functional Studies

Shatkay et al. [15] argue that clustering analysis cannot solve the core issues of functional microarray studies. Genes that are functionally related often show a strong anti-correlation in their expression profiles and hence will not be clustered together using ordinary similarity measures. Clustering genes in disjoint clusters will not capture the fact that many gene products have more than one function. Sherlock [16] also points out that most studies using clustering techniques do not report any measure of whether the overlap between the genes in a functional class and the genes in a particular expression cluster is greater than what would be expected by chance. Tavazoie et al. [18] address this problem.

The ultimate goal of functional microarray studies is to classify previously unclassified genes. These classifications can be used as hypotheses in wet-lab experiments and can greatly reduce the number of options the biologists have to consider. Classifying previously unclassified genes, however, requires a model of the relationship between gene expression and function. Supervised methods seem to meet this requirement better than unsupervised methods. Although Brown et al. [3] showed that gene function indeed can be learned, a lot of issues are still not treated. These include the development of methods for determining which functional classes to learn and methods for handling genes with more than one function. Some solutions are suggested in Hvidsten et al. [7].

Clinical Studies

Clinical microarray studies have up until now mostly evolved around cancer classification. However, a lot of different diseases will probably benefit from the new opportunities offered by microarrays. As a rule of thumb one should use supervised methods when the classes are known. This offers the opportunity to investigating the properties of different subclasses, e.g. which genes are responsible for a certain type of disease. In the future, classifiers may also offer decision support in daily medical practice. Unsupervised methods will probably prove very important in finding new subclasses of diseases and hence help distinguish diseases that up until now have been treated similarly.

Acknowledgements

I want to thank Jan Komorowski for introducing me to the field and for continuous guidance. I would also like to thank Astrid Lgreid and Arne Sandvik for much needed biological guidance, and for providing me with many of the articles cited in this paper. Thanks to Herman Midelfart for interesting discussions on machine learning and performance estimates and to Lasse Natvig for introducing me to articles discussing computational science in general.

References

[1] Ash A. Alizadeh and et. al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[2] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, 96:6745–6750, 1999.

[3] M. P. S. Brown, W. N. Grundy, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA*, 97(1):262–267, 2000.

[4] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines (and other kernel-based learning methods)*. Cambridge University Press, 2000.

[5] M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression pattern. *Proc. Natl. Acad. Sci. USA*, 95(25):14863–14868, 1998.

[6] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

[7] Torgeir R. Hvidsten, Jan Komorowski, Arne K. Sandvik, and Astrid Lægreid. Predicting gene function from gene expressions and ontologies. In Russ B. Altman, A. Keith Dunker, Lawrence Hunter, Kevin Lauderdale, and Teri E. Klein, editors, *Pacific Symposium on Biocomputing*, pages 299–310, Mauna Lani, Hawai'i, January 2001. World Scientific Publishing Co.

[8] V. R. Iyer, M. B. Eisen, D. T. Ross, G. Schuler, T. Moore, J. C. F. Lee, J. M. Trent, L. M. Staudt, J. Dudson Jr., M. S. Boguski, D. Lashkari, D. Shalon, D. Botstein, and P. O. Brown. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283(5398):83–87, 1999.

[9] Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice hall, Upper Saddle River, New Jersey, Upper Saddle River, New Jersey, fourth edition, 1998.

[10] Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[11] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, Singapore, 1997.

[12] Center for Computational Biology, Montana State University, Bozeman. [http://nervana.montana.edu/general/overview.html]. 28.03.2001.

[13] Robert Schalkoff. *Pattern Recognition: Statistical, Structural and Neural Approaches*. John Wiley & Sons, 1992.

[14] M. Schena, D. Shalon, R. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.

[15] H. Shatkay, S.W. Edwards, W.J. Wilbur, and M. Boguski. Genes, themes, and microarrays. In *ISMB2000 Proceedings*, 2000.

[16] Gavin Sherlock. Analysis of large-scale gene expression data. *Current Opinion in Immunology*, 12:201–205, 2000.

[17] D.E. Stevenson. Science, computational science, and computer science: At a crossroads. *Communications of the ACM*, 37(12):85–96, 1994.

[18] Saeed Tavazoie, Jason D. Hughes, Michael J. Campbell, Raymond J. Cho, and George M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22:281–285, 1999.