Biclusters Evaluation Based on Shifting and Scaling Patterns

Juan A. Nepomuceno[1], Alicia Troncoso Lora[2], Jesús S. Aguilar–Ruiz[2]
Jorge García–Gutiérrez[1]

[1] Department of Computer Science,
University of Sevilla, Sevilla, Spain
{janepo, jgarcia}@lsi.us.es
[2] Area of Computer Science, School of Engineering,
Pablo de Olavide University, Sevilla, Spain
{ali, aguilar}@upo.es

Abstract. Microarray techniques have motivated the develop of different methods to extract useful information from a biological point of view. Biclustering algorithms obtain a set of genes with the same behaviour over a group of experimental conditions from gene expression data. In order to evaluate the quality of a bicluster, it is useful to identify specific tendencies represented by patterns on data. These patterns describe the behaviour of a bicluster obtained previously by an adequate biclustering technique from gene expression data. In this paper a new measure for evaluating biclusters is proposed. This measure captures a special kind of patterns with scaling trends which represents quality patterns. They are not contemplated with the previous evaluating measure accepted in the literature. This work is a first step to investigate methods that search biclusters based on the concept of shift and scale invariance. Experimental results based on the yeast cell cycle and the human B-cell lymphoma datasets are reported. Finally, the performance of the proposed technique is compared with an optimization method based on the Nelder-Mead Simplex search algorithm.

keywords: Gene expression data, biclustering, shifting and scaling patterns, unconstrained optimization.

1 Introduction

In the last few years microarrays techniques have generated a great amount of biological information. Microarray data can be represented by a numerical matrix with its columns corresponding to experimental conditions and rows associated with genes. Thus the element (i, j) is the expression level of the gene i under the specific condition j. Data mining techniques have been successfully applied to gene expression data in order to discover subtypes of diseases, identification of functional grouping of genes, etc. Clustering techniques have been applied to microarray data [1] in order to identify groups of genes that show similar

expression patterns. Most of clustering models have been focused on discovering clusters embedded in a subset of dimensions, because relevant genes are not necessary related to every condition [2]. This problem is known as *biclustering or subspace clustering*. Thus, the goal of biclustering techniques is to extract subgroups of genes with similar behavior under specific subgroups of conditions [3]. This is a vital task from a biomedical point of view, since it is the first step in order to discover networks of genes interaction.

Biclustering problem is a NP-hard problem [4], therefore different techniques use heuristics approaches in order to find biclusters, for example evolutionary algorithms [5–7]. These methods are based on a measure to evaluate the quality of biclusters, with the *Mean Squared Residue* (MSR) [8] the most important measure for assessing the quality of biclusters. For this reason, bicluster evaluation is a vital task for searching patterns in biological data. MSR evaluation measure is based on computing the arithmetic means of the values in each row, column, and the full matrix, and the numerical differences among the data. However, it have been proved that MSR is effective for recognizing biclusters with shifting patterns but not some patterns with scaling trends, in spite of representing quality patterns [9]. A bicluster has a shifting pattern when its values vary in the addition of a constant value, and scaling pattern when its values vary in the multiplication of a constant value. A *perfect bicluster* is considered as the one which follows exactly a perfect shifting and scaling pattern [4]. Consequently, it is interesting to study the behavior or tendencies in a bicluster in order to establish a new quality measure through the degree of similarity with its corresponding perfect bicluster.

This fact represents the main motivation of this work where a new measure for evaluating biclusters is proposed. We apply a classical optimization method to solve a least squared statistical estimation problem in order to build the perfect bicluster of a bicluster. After that, the value of the optimization function in the convergence point is the value for the measure of it. If a bicluster presents perfect shifting and scaling patterns, it will be a perfect bicluster itself and its measure will be zero. Although the main task about biclustering problem is to find good biclusters from a microarray, this work is relevant in order to investigate methods that search biclusters based on the concept of shift and scale invariance. First, the problem is formulated from a mathematical point of view leading to an unconstrained nonlinear optimization problem. Later, the problem is solved by a classical Quasi-Newton method. Finally, experimental results obtained from biclusters on the yeast cell cycle and the human B-cell lymphoma datasets are reported. The performance of the proposed method is compared with a search technique based on the Nelder-Mead Simplex algorithm.

The paper is organized as follows: Section 2 presents basic concepts on patterns from gene expression data. A brief overview on unconstrained optimization techniques is shown in Section 3. The formulation of the problem is described in Section 4. Section 5 reports some results obtained from the application of two techniques to two real datasets. Finally, the main conclusions of the paper are outlined.

2 Shifting and Scaling Patterns

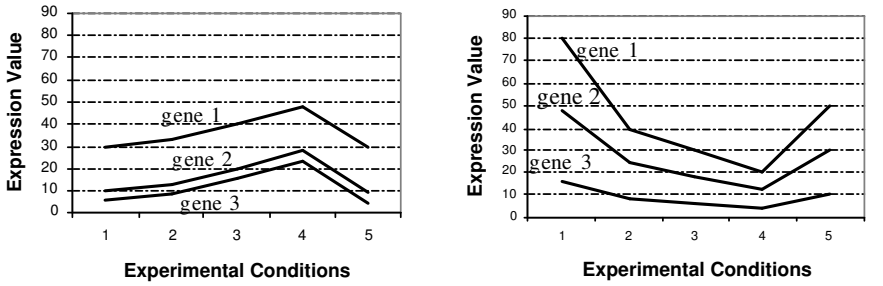
Given a bicluster, the shifting and scaling patterns can be formally defined. A microarray M is a real matrix composed by N genes and M conditions. The element (i, j) of the matrix is represented by $v_{i,j}$. A bicluster B is a submatrix of M composed by $n \leq N$ rows and $m \leq M$ columns. The element (i, j) of the bicluster B is represented by $w_{i,j}$.

A group of genes has a shifting pattern when the values $w_{i,j}$ vary in the addition of a value β_i. Analogously, a bicluster has a scaling pattern when the values $w_{i,j}$ vary in the multiplication of a value α_i. The values β_i and α_i are fixed for all the genes. Formally, a bicluster shows a shifting or scaling pattern respectively when it follows the expressions (1) or (2) respectively:

$$w_{i,j} = \pi_j + \beta_i \tag{1}$$

$$w_{i,j} = \pi_j \times \alpha_i \tag{2}$$

where π_j is a typical value for the gene j and fixed for all the conditions.



$$\begin{bmatrix} 30 & 10 & 5 \\ 33 & 13 & 8 \\ 40 & 20 & 15 \\ 48 & 28 & 23 \\ 29 & 9 & 4 \end{bmatrix} = \begin{bmatrix} 30+0 & 10+0 & 5+0 \\ 30+3 & 10+3 & 5+3 \\ 30+10 & 10+10 & 5+10 \\ 30+18 & 10+18 & 5+18 \\ 30-1 & 10-1 & 5-1 \end{bmatrix} \qquad \begin{bmatrix} 80 & 48 & 16 \\ 40 & 24 & 8 \\ 30 & 18 & 6 \\ 20 & 12 & 4 \\ 50 & 30 & 10 \end{bmatrix} = \begin{bmatrix} 10\times 8 & 6\times 8 & 2\times 8 \\ 10\times 4 & 6\times 4 & 2\times 4 \\ 10\times 3 & 6\times 3 & 2\times 3 \\ 10\times 2 & 6\times 2 & 2\times 2 \\ 10\times 5 & 6\times 5 & 2\times 5 \end{bmatrix}$$

Fig. 1. Bicluster with a) shifting patterns, b) scaling patterns.

Figure 1a) presents a bicluster that contains a shifting pattern. Shifting patterns represent related genes that show the same shape and slope. It can be observed that the genes start with different initial values. Thus, shapes of the graphs are similar, but values are not equal. Figure 1b) presents a bicluster that contains a scaling pattern. In this case, scaling patterns represent related genes showing the same shape, but different slopes. It can be noted that changes more abrupt for one gene than for the other are shown.

In a general case, an element of a bicluster showing both types of patterns can be defined as:

$$w_{i,j} = \alpha_i \times \pi_j + \beta_i + \varepsilon_{i,j} \tag{3}$$

where $\varepsilon_{i,j}$ is the error that the patterns have for the value $w_{i,j}$ of the bicluster considered.

A bicluster is a perfect bicluster when the value of $\varepsilon_{i,j}$ is equal to zero for all values $w_{i,j}$ of the bicluster.

3 Unconstrained Optimization Techniques

Unconstrained optimization techniques are used to search local minima in optimization problems whose objective function is not subject to equality and inequality constraints.

An unconstrained optimization problem can be defined as:

$$min\ f(x)$$

where $x \in \mathbb{R}^n$ is a vector of real variables and $f : \mathbb{R}^n \mapsto \mathbb{R}$ is a linear or nonlinear scalar function.

There is a great number of methods to solve unconstrained optimization problems. Nowadays, unconstrained optimization problems can be classified in two groups: search methods and gradient methods. Search methods [10] use only function evaluations and these approaches are most suitable for problems that are very nonlinear or have a great number of discontinuities. Simplex search methods are based on searching the local minima inside a particular region or simplex. A simplex in n-dimension space is characterized by $n+1$ distinct vectors that are its vertices. At each iteration, the objective function is evaluated in a new point generated inside the simplex, which is compared with value of the function at vertices of the simplex. One of the vertices could be replaced by the new point. The process is repeated until the diameter of the simplex is less than a specified tolerance.

Gradient methods [11] are generally more efficient when the first derivative of the objective function is continuous. The search direction to locate the minimum is proportional to the gradient of the objective function as follows [3]:

$$x_{k+1} = x_k - \alpha_k \cdot \nabla f(x_k) \tag{4}$$

where α_k is the step–length parameter and x_k is the variable x at iteration k.

The parameter α_k is obtained by a line–search method. The line–search approach consists in solving a minimization problem in one dimension. This problem can be formulated as follows:

$$min\ \phi(\alpha)$$

[3] $\nabla f(x_k)$ is the value of gradient of function f in x_k point.
$\nabla f(x_k) = (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n})$, $\frac{\partial f}{\partial x_i}$ represents the derivative of f respect x_i variable.

where $\alpha \in \mathbb{R}$ and $\phi(\alpha) = f(x_k + \alpha \cdot \nabla f(x_k))$ with x_k and $\nabla f(x_k)$ fixed.

Newton methods are higher order gradient methods. This is due to the use of second order information. These methods are only really suitable when the second order information is readily and easily calculated, because calculation of such is computationally expensive. In this case, the search direction can be written as follows:

$$x_{k+1} = x_k - \alpha_k \cdot H_k^{-1} \cdot \nabla f(x_k) \tag{5}$$

where H_k^{-1} is the inverse of the Hessian matrix at point x_k.

Quasi-Newton methods are Newton methods, which use an approximation to the inverse of the matrix H_k as an alternative to calculate it directly. Different Quasi-Newton methods are based on different approximations of the inverse of the matrix H_k [12, 13].

4 Formulation of the problem

The objective of the problem is to determine the shifting and scaling patterns of a certain bicluster. It is supposed that biclusters are obtained previously by appropriate biclustering techniques.

The objective function is defined by the mean squared error (MSE) as follows:

$$MSE = \frac{1}{n \cdot m} \sum_{i=1}^{n} \sum_{j=1}^{m} \varepsilon_{i,j}^2 \tag{6}$$

where $\varepsilon_{i,j}$ is defined in Eq. 3. It can be noted that the error $\varepsilon_{i,j}$ depends on the shifting patterns β_i, scaling patterns α_i and the typical value for each gene π_j.

Thus, the problem can be formulated as the following unconstrained optimization problem:

$$min \quad f(\overrightarrow{\alpha}, \overrightarrow{\beta}, \overrightarrow{\pi}) \tag{7}$$

where $\overrightarrow{\alpha} = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$, $\overrightarrow{\beta} = (\beta_1, \ldots, \beta_n) \in \mathbb{R}^n$, $\overrightarrow{\pi} = (\pi_1, \ldots, \pi_m) \in \mathbb{R}^m$ and $f : \mathbb{R}^{2n+m} \longmapsto \mathbb{R}$ is defined by the MSE (Eq. 6).

The result of this optimization problem is the optimal point and the value of the function on it. Shifting and scaling patterns for the bicluster are built with $\overrightarrow{\alpha}, \overrightarrow{\beta}$ and $\overrightarrow{\pi}$ values. The quality of the bicluster is established with the value of the objective function on the solution.

This unconstrained optimization problem has been solved by using a Quasi–Newton method. The Hessian matrix has been approximated by using the formula of Shanno [13] and the step-length parameter has been determined by a line–search technique. Also, this optimization problem has been solved by using the Nelder-Mead Simplex search algorithm in order to establish a comparison.

5 Experiments

An unconstrained optimization technique based on the Quasi-Newton method has been applied to solve the proposal problem. Shifting and scaling patterns, that is to say, the perfect bicluster which approximates the original bicluster, and the value of objective function like a quality measure are obtained for each bicluster.

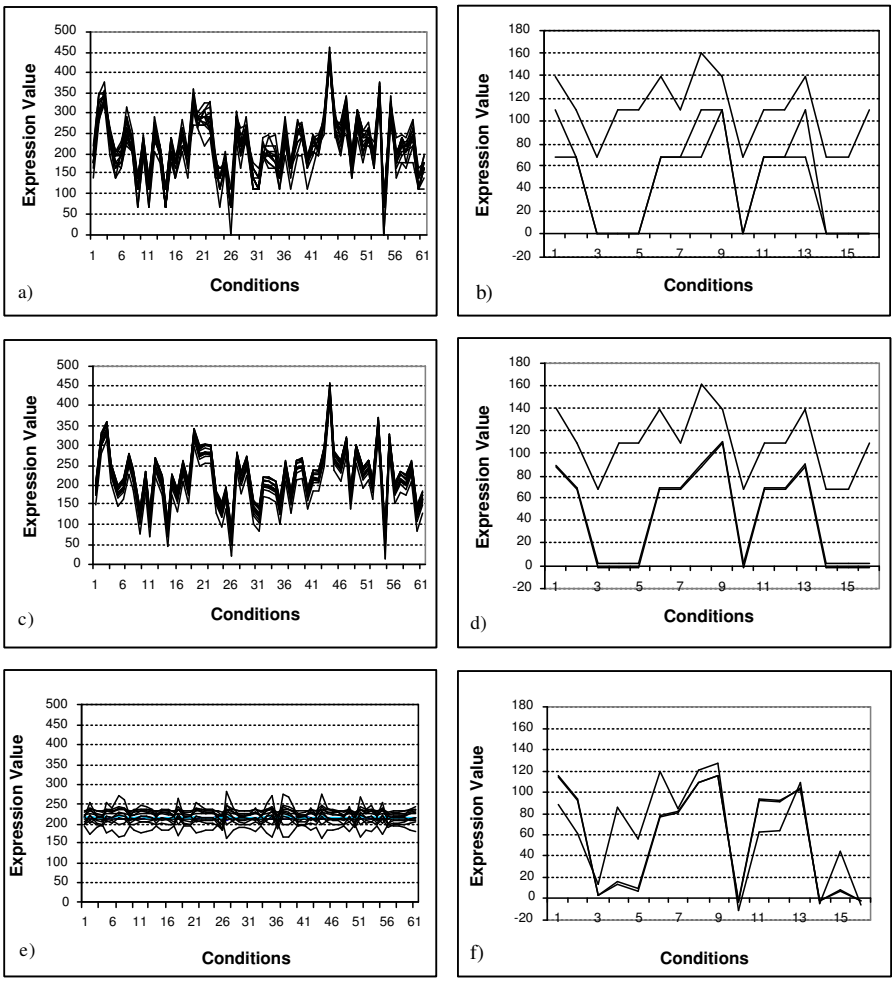


Fig. 2. a), b) Two yeast bicluster leading to the worst and best patterns respectively, c), d) patterns obtained with Quasi-Newton method and e), f) patterns obtained with Nelder-Mead Simplex algorithm.

Original biclusters have been obtained from the work recently published in [5]. These biclusters have been built from a biclustering technique based on an

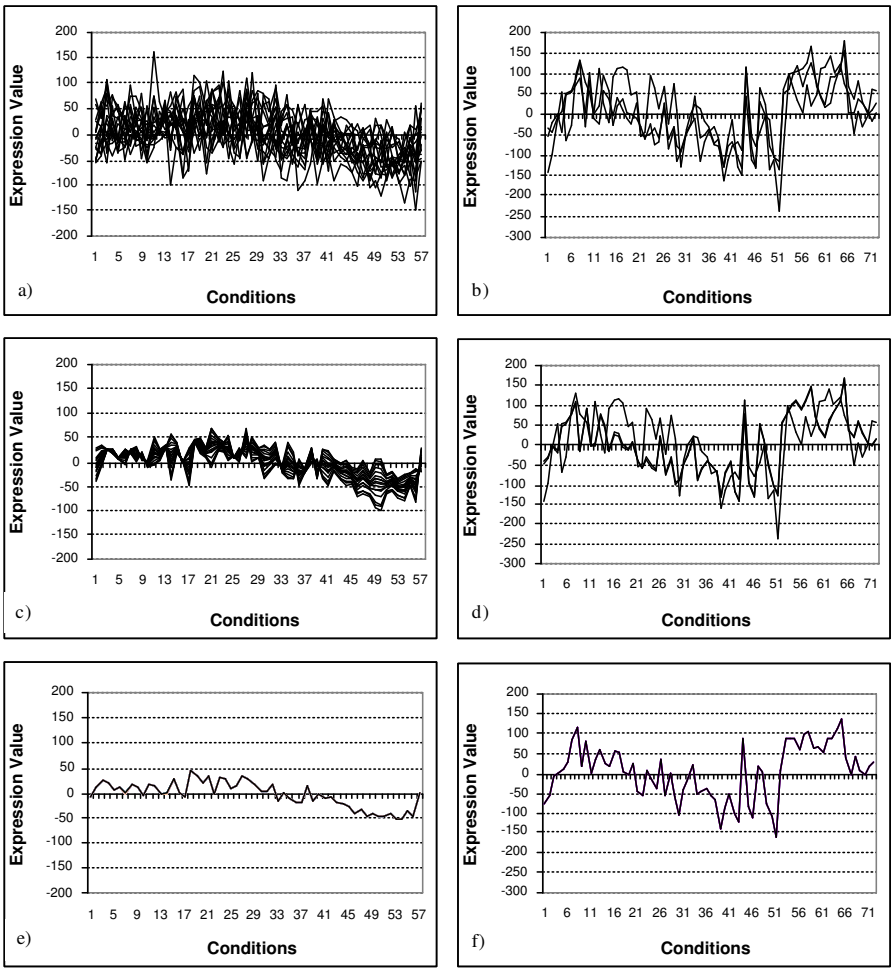


Fig. 3. a), b)Two human bicluster, leading to the worst and best patterns respectively, c), d) patterns obtained with Quasi-Newton method and e), f) patterns obtained with Nelder-Mead Simplex algorithm.

evolutionary algorithm applied to two well-known datasets: yeast *Saccharomyces cerevisiae* cell cycle expression dataset [14]; and the human B-cells expression data [15]. The Yeast dataset contains 2884 genes and 17 experimental conditions and the Human dataset consists of 4026 genes and 96 conditions. The proposed technique has been applied over the one hundred biclusters obtained in [5]. Results obtained for the biclusters leading to the worst and the best shifting and scaling patterns in both datasets are reported.

Figures 2a) and 2b) present two biclusters from Yeast dataset leading to the worst and the best shifting and scaling patterns obtained by the Quasi-Newton optimization method (Figures 2c) and 2d)), in the sense of the highest and the

lowest value for the evaluation function. These two biclusters are composed by 14 and 3 genes and 61 and 16 conditions, respectively. Notice that the expression values for several genes over certain conditions are the same (black thick lines). Its corresponding shifting and scaling patterns are shown in Figures 2c) and 2d). The final value of the error function defined by the MSE (Eq. 6) is equal to 14.51 for bicluster on the left and 7.16 for bicluster on the right, respectively. A good quality of the discovered patterns can be observed in both biclusters. Figures 2e) and 2f) show the patterns obtained by the optimization method based on the Nelder-Mead Simplex algorithm for two biclusters. Notice that these shifting and scaling patterns obtained are worst than the first ones, in the sense that they adjust the shape of the original bicluster in a worst way.

Figures 3a) and 3b) present two biclusters from Human dataset leading to the worst and the best shifting and scaling patterns, left and right respectively. These biclusters are constituted for 17 genes and 57 conditions and 3 genes and 72 conditions, respectively. A bad quality of the built patterns can be observed in the Figure 3d), in spite of the bicluster leads to the best patterns. It is due to the low number of genes and the irregular behaviour of this bicluster. Obviously, biclusters with an uniform behaviour provide better shifting and scaling patterns than those with no inherent tendency. Figures 3e) and 3f) present the patterns obtained from the application of the Nelder-Mead Simplex method for two biclusters. These patterns have 17 and 3 genes, as original bicluster, but expression values for genes over all the conditions are the same, as only a black thick line can be distinguished. Obviously, these shifting and scaling patterns are worst than the previous ones obtained with the proposed method based on the Quasi-Newton method.

Table 1. Comparison between two optimization methods used to build shifting and scaling patterns in biclusters: Quasi–Newton method and Nelder-Mead Simplex algorithm.

	number of iterations		**time in seconds**		**obj. function**	
	Yeast	Human	Yeast	Human	Yeast	Human
Quasi-Newton method	67,59	193,67	5,08	83,73	13,31	30,52
N-M Simplex algorithm	49264,20	50000	69,15	226,06	66,431	35,05

Finally, in Table 1 a comparison is made between the two used techniques, Quasi–Newton algorithm and Nelder-Mead Simplex search method. Table 1 shows the most representative parameters of optimization process for two datasets. It can be observed the average of the iterations number, the CPU time and the value of the objective function on one hundred biclusters obtained from yeast cell cycle microarray and one hundred biclusters from human B-cells microarray. Notice the highest cost in time (in seconds) and in number of iterations and the worst patterns are obtained by Nelder-Mead Simplex search approach.

6 Conclusions

An unconstrained optimization technique has been applied in order to build shifting and scaling patterns from biclusters. The method has been tested over biclusters obtained from two different real datasets: yeast cell cycle and human B-cells. Results have shown that the proposed approach has a good performance for finding shifting and scaling patterns of a given bicluster. The proposed technique has been compared with an optimization method based on the Nelder-Mead Simplex algorithm showing better results with regarding to the patterns found and the CPU time.

Future works will be focused on the comparison between different biclustering algorithms using the proposal measure in order to establish which one is the best. Some more actual microarrays taken from PNAS journal will be used. Biclusters obtained by others biclustering algorithms such as Cheng-Church, ISA, OPSM, etc, will be also used for this proposal. On the other hand, we will also study the possibility of new biclustering tecniques based on the concept of shift and scale invariance.

Acknowledgments: Thanks are due to the Spanish CICYT (TIC2004-00159) and Junta de Andalucía (P05-TIC-00531) for sponsoring this research.

References

1. A. Ben-Dor, R. Shamir and Z. Yakhini. Clustering Gene Expression Patterns. Journal of Computational Biology, 6, pp. 281–297, 1999.
2. H. Wang, W. Wang, J. Yang and P. S. Yu. Clustering by Pattern Similarity in Large Data Sets. ACM SIGMOD International Conference on Management of Data, pp. 394-405, 2002.
3. A. Tanay, R. Sharan and R. Shamir. Discovering Statistically Significant Biclusters in Gene Expression Data. Bioinformatics, 18, pp. 196–205, 2002.
4. S. C. Madeira, A. L. Oliveira. Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1, pp. 24-45, 2004.
5. F. Divina and J. S. Aguilar-Ruiz. Biclustering of Expression Data with Evolutionary Computation. IEEE Transactions on Knowledge & Data Engineering, Vol. 18, no. 5, pp. 590–602, 2006.
6. K. Bryan, P. Cunningham and N. Bolshakova. Biclustering of Expression Data Using Simulated Annealing. IEEE Symposium on Computer-Based Medical Systems, pp. 383-388, 2005.
7. Sushmita Mitra, Haider Banka Multi-objective evolutionary biclustering of gene expression data. Pattern Recognition, vol. 39 - n°12, pp. (2464-2477). 2006.
8. Y. Cheng and G. M. Church. Biclustering of Expression Data. In Proceedings of the 8th International Conference on Itelligent. Systems for Molecular Biology, pp. 93–103, La Jolla, CA, 2000.
9. J. S. Aguilar-Ruiz. Shifting and Scaling Patterns from Gene Expression Data. Bioinformtics, Vol. 21, no. 20, pp. 3840–3845, 2005.
10. J.A. Nelder and R. Mead. A Simplex Method for Function Minimization. Computer J., Vol.7, pp. 308-313, 1965.

11. M. S. Bazaraa, H. D. Sherali and C. M. Shetty. Nonlinear Programing: Theory and Algorithms. John Wiley and sons, 1993.
12. R. Fletcher. A New Approach to Variable Metric Algorithms. Computer Journal, Vol. 13, pp. 317-322, 1970.
13. D. F. Shanno. Conditioning of Quasi-Newton Methods for Function Minimization. Mathematics of Computing, Vol. 24, pp. 647-656, 1970.
14. R. Cho et al. A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle. Molecular Cell, Vol. 2, pp. 65–73, 1998.
15. A. A. Alizadeh et al. Distinct Types of Diffuse Large b-cell Lymphoma Identified by Gene Expression Profiling. Nature, Vol. 403, pp. 503–511, 2000.