# Integrating quality-based clustering of microarray data with Gibbs sampling for the discovery of regulatory motifs

Yves MOREAU[†] Gert THIJS[†] Kathleen MARCHAL[†]
Frank DE SMET[†] Janick MATHYS[†] Magali LESCOT[¥]
Stéphane ROMBAUTS[‡] Pierre ROUZÉ[‡*] Bart DE MOOR[†]

[†] Katholieke Universiteit Leuven ESAT-SCD (SISTA) – Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
[‡] Flemish Institute for Biotechnology (VIB) & Univ. of Ghent – K.L. Ledeganckstraat 35, B-9000 Gent, Belgium
[¥] LGPD, CNRS, Case 907 – Parc Scientifique de Luminy, F-13288 Marseille Cedex 9, France
[*] Laboratoire Associé de l'INRA

Email : yves.moreau@esat.kuleuven.ac.be

## Abstract

*In microarray experiments, genes exhibiting a similar expression profile are potentially coregulated. Clustering identifies such groups of coexpressed genes, whose upstream regions can then searched for putative regulatory elements. We present two algorithms and an interactive web-based user interface that integrate cluster analysis and motif finding for the analysis of microarray data. Starting from the expression, we present our adaptive quality-based clustering algorithm to define groups of tightly coexpressed genes. The upstream region is then retrieved based on the accession number and gene name. Once the upstream regions are identified, the sequences are analyzed using Gibbs sampling for motif finding to find the over-represented motifs. Our implementation (called Motif Sampler) allows the use of higher-order models for the sequence background. This methodology can be used through our INCLUSive web interface at the following URL: http://www.esat.kuleuven.ac.be/~dna/BioI/Software.html*

**Keywords:** *microarrays, clustering, motif finding, Gibbs sampling.*

## 1 Introduction

Microarray experiments generate a considerable amount of data, which analyzed properly help us gain a lot of biologically or medically relevant information about the global cellular behavior (e.g., normal versus tumor cells, dynamics of dividing cells in the cell cycle). One of the first steps in data analysis of high-throughput expression measurements (after preprocessing) is the clustering of genes to find groups of genes that have a similar behavior or expression profile. We describe here a new specialized algorithm, called adaptive quality-based clustering, which was developed for the specific task of finding highly coherent groups of genes. Once interesting groups of genes are identified we look at the DNA sequence around these genes. In this case we are particularly interested in the region upstream of the gene where many transcriptional regulators bind to the DNA. We present an extension of the Gibbs sampling method for motif finding that enables the use of higher-order models of the sequence background. Gibbs sampling makes it possible to identify, through a stochastic search method, possible motifs in upstream regions when the motif we are looking for has never been identified before. Finally, we describe shortly how these tools are integrated into a user-friendly web application.

## 2 Adaptive quality-based clustering

Currently, the analysis of microarray data from complex experiments starts mostly with clustering. One of the objectives of clustering algorithms is to detect groups of genes that exhibit a highly similar expression profile. Since gene expression profiles are encoded in real vectors, these algorithms intend to group gene expression vectors that are sufficiently close to each other according to a certain distance or similarity measure. However, most clustering algorithms originate from research fields outside biology. Therefore, although useful, the original implementations suffer from some drawbacks as has been highlighted by Sherlock [6]. These deficiencies can be summarized as follows. Firstly, algorithms such as *K*-means and Self-Organizing Maps require the predefinition of the number of clusters (parameter of the algorithm). When clustering gene expression profiles, the number of clusters present in the data is usually unknown. Changing this parameter usually affects the final clustering result considerably. Clustering, using for example *K*-means, therefore involves extensive parameter fine-tuning to detect the optimal clustering and the choice of the final parameter setting remains somehow arbitrary (e.g., based on visual inspection of the clusters). When using hierarchical clustering, the number of clusters is determined by cutting the tree structure at a certain level. The resulting cluster structure is therefore highly influenced by the choice of this level, which in turn is rather arbitrary. Secondly, the idea of

forcing each gene of the data set into a cluster is a significant drawback of these implementations. If genes are, despite a rather low correlation with other cluster members, forced to end up in one of the clusters, the average profile of this cluster is corrupted and the composition of the cluster becomes less suitable for further analyses (such as motif finding, see further).

Much effort is currently being done to adapt clustering algorithms towards the specific needs of biological problems. In this context, Heyer *et al.* [4] proposed an algorithm that tries to identify clusters that have a certain quality (representing the minimal degree of coexpression needed) and where every cluster contains a maximal number of points. Genes not exhibiting this minimal degree of coexpression with any of the clusters are excluded from further analysis. A problem with the quality-based approach of Heyer *et al.*, however, is that this quality is a user-defined parameter that is hard to estimate (it is hard to find a good trade-off or optimal value: setting the quality too strictly will exclude a considerable number of coexpressed genes, setting it too loose will include too many genes that are not really coexpressed). Moreover, it should be noted that the optimal value for this quality is, in general, different for each cluster and data set dependent.

We have developed an adaptive quality-based clustering method [3] starting from the principles described by Heyer *et al.* [4] (quality-based approach; locating clusters, with a certain quality, in a volume where the density of points is maximal). The algorithm is in essence a heuristic, two-step approach that defines the clusters sequentially (the number of clusters is not known in advance, so it is not a parameter of the algorithm). The first step locates a cluster and the second step derives the size (quality) of this cluster from the data.

Clusters formed by our algorithm might be good 'seeds' for further analysis of expression data [8,9,10] since they only contain a limited number of false positives. When the presence of false positives in a cluster is undesirable, a more stringent value for the significance level $S$ might be applied (e.g., 99%; for noise-sensitive analyses such as motif finding) which will result in smaller clusters exhibiting a more tightly related expression profile.

The algorithm is a heuristic iterative two-step approach. There are two user-defined parameters: (1) the minimal number of genes in a cluster and (2) the significance level $S$. During each iteration, this algorithm first finds a cluster center using a preliminary estimate of the radius or quality of the cluster. Given a collection $G$ of gene expression profiles, the objective is to find a cluster center in an area of the data set where the 'density' (or number) of expression profiles, within a sphere with an initial radius or quality is locally maximal. After initialization of the cluster center (with the mean profile of all the expression profiles in the data set $G$), all the expression profiles within a sphere with radius RAD are selected. Iteratively, the mean profile of these expression profiles is calculated and subsequently the cluster center is moved to this mean profile. This approach moves the cluster in the direction where the 'density' of profiles is higher.

When the cluster center has been located, it remains fixed. The algorithm then determines a new estimate for the radius of the cluster. To estimate the radius of the cluster, first the distance of each gene in the data set to the cluster center is calculated. The distribution of these distances in the original data consists of two parts:

1. Genes belonging to the background of the current cluster center, which do not belong to any cluster (noise; are not significantly coexpressed with other genes) or belong to another cluster. The distribution of the distance of these genes does not differ significantly from the distribution of a randomized data set.
2. Genes belonging to the cluster have a distribution that is not present in the distribution of the randomized data set and are significantly coexpressed.

Finding the parameters of the proposed model is accomplished by fitting the model to the distance distribution with the Expectation-Maximization algorithm.

As an example, the algorithm was tested on the expression profiling experiment of Cho *et al.* [2] studying the yeast cell cycle in a synchronized culture on an Affymetrix chip. The cell cycle of yeast is a fundamental biological system as it reveals the core processes involved in cell replication and growth in general. This knowledge is essential to the understanding of the aberrant processes involved in tumorigenesis and carcinogenesis. This data set can be considered as a benchmark and contains expression profiles for 6,220 genes over 17 time points taken at 10-min intervals, covering nearly two full cell cycles. The majority of the genes included in the data set have been functionally classified, which makes this data set an ideal candidate to correlate the results of new clustering algorithms with the biological reality. Comparison with the results of Tavazoie *et al.* [7] showed that adaptive quality-based clustering provides a significantly higher enrichment in specific functional classes than $K$-means.

# 3 Motif finding by Gibbs sampling

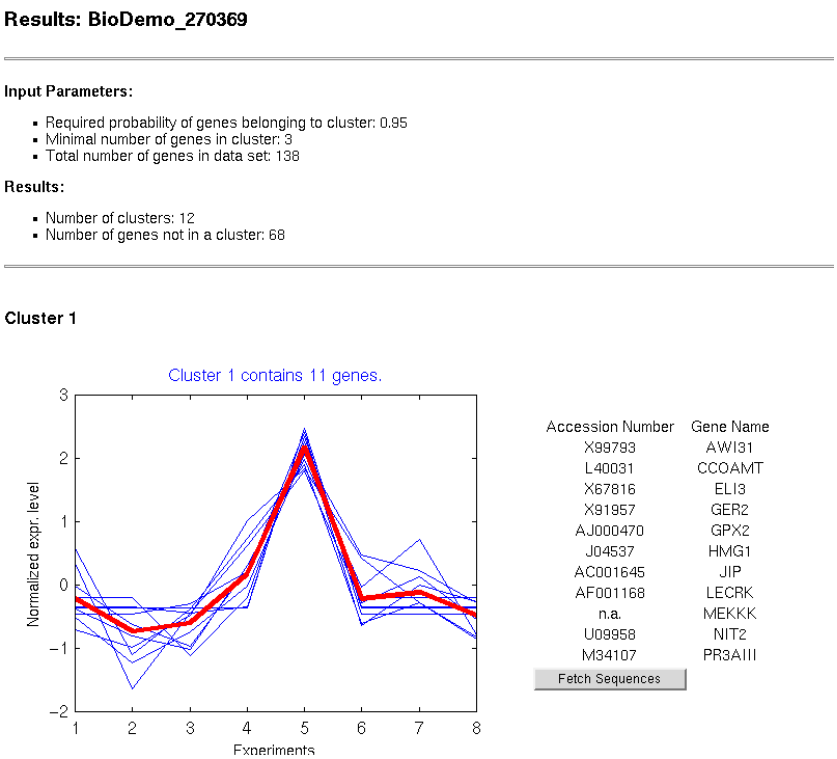
The use of microarrays has opened new direction in transcript profiling research. An interesting application of this technology is the study of the transcriptional regulatory mechanism that is responsible for the coordinated behavior of the coexpressed genes. The basic assumption states that coexpression frequently arises from transcriptional coregulation. As coregulated genes are known to share some similarities in their regulatory mechanism, possibly at the transcriptional level, their promoter regions might contain some common motifs that

are binding sites for transcription regulators. A sensible approach to detect these regulatory elements is to search for statistically over-represented motifs in the promoter region of such a set of coexpressed genes.
Algorithms to find regulatory elements can be divided into two classes: (1) methods based on word counting [11,12] and (2) methods based on probabilistic sequence models [1,5]. The word counting methods analyze the frequency of oligonucleotides in the upstream region and use intelligent strategies to speed up counting and to detect significantly over-represented motifs. These methods then compile a common motif by grouping similar words. The probabilistic methods represent the motif by a position probability matrix and they assume that the motif is hidden in a noisy background sequence. To find the parameters of this model these methods use maximum likelihood estimation in the form of Expectation Maximization (EM) and Gibbs sampling – EM is a deterministic algorithm and Gibbs Sampling is a stochastic equivalent of EM.
Within the probabilistic methods the motif is represented as a position probability matrix of dimensions 4x$W$, where $W$ is the length of the motif. Each column in the matrix is discrete distribution over the four nucleotides A, C, G and T and each entry in the column gives the probability of finding a given nucleotide at that position in the motif. Figure 2 shows the position probability matrix that is the output of our algorithm.
We have introduced two modifications [8,10] to the original Gibbs sampling algorithm by Lawrence *et al.* [5]. First, a probabilistic framework was used to estimate the expected number of copies of a motif in a sequence. Since both the microarray experiment and the clustering are subject to noise, only a subset of the coexpressed genes is actually coregulated. Furthermore, in higher organisms, regulatory elements can have several copies to increase the effect of the transcriptional binding factor in the transcriptional regulation.
When searching for possible regulatory elements in such a set of sequences we should take into account that the motif will only appear in a subset of the original data set or could have multiple copies. We therefore want to develop an algorithm that distinguishes between the sequences in which the motif is present and those in which it is absent. We reformulated the probabilistic sequence model in such a way that the number of copies of the motif in each sequence can be estimated.
Second, we introduced the use of a higher-order background model based on a Markov chain. The idea of using a higher-order background model is justified by the fact that Markov models have been used successfully in state-of-the-art gene detection software. Here, we developed a background model based on a Markov process of order $m$. This means that the probability of the nucleotide $b$ at position $l$ in the sequence depends on the $m$ previous bases in the sequence. A transition matrix describes such a model. Important to know is that the background model can be either constructed from the original sequence data or from an independent data set. The latter approach is more sensible if the independent data set is carefully created, which means that the sequences in the training set only come from the intergenic region and thus do not overlap with coding sequences. Nevertheless the algorithm can also be used for other organisms by building the background model from the input sequences.
Integrating the two proposed modifications into the original Gibbs sampling algorithm for motif finding lead to our implementation, called Motif Sampler. The input of the Motif Sampler is a set of upstream sequences. In the first step of the algorithm the higher-order background model is chosen. The background model can be pre-compiled or it can be calculated from the input sequences. The algorithm then uses this background model to compute, for each segment of length $W$ in every sequence the probability that the segment was generated by the background model. Second, the alignment vector and the corresponding motif model are initialized. In the next step, the actual core of the sampling procedure starts. The algorithm loops over all sequences and updates the alignment vector for each sequence. First, the motif model is calculated based on the current alignment vector. This estimated motif model is used to compute for each segment of length $W$ in the selected sequence a weight that is the ratio of the probability that the segment is generated by this motif model divided by the probability that the segment is generated by the background model. Finally, a new alignment vector is selected by taking samples from the normalized distribution of weights over all segments in the given sequence. This procedure is repeated until the motif model has converged.

# 4 INCLUSive: a web application for adaptive quality-based clustering and Gibbs sampling for motif finding

The implementations of adaptive quality-based clustering and of our motif finding algorithm are part of our INCLUSive web site [9] and are accessible through a web interface:
http://www.esat.kuleuven.ac.be/~dna/BioI/Software.html.

Figure 1 is a screenshot of part of the results page. First an overview of the parameters is given. For each cluster a plot of the expression profiles is given together with a list of all the genes present in the cluster. In this example the genes are identified by their accession number and gene name.



**Results: BioDemo_270369**

**Input Parameters:**

- Required probability of genes belonging to cluster: 0.95
- Minimal number of genes in cluster: 3
- Total number of genes in data set: 138

**Results:**

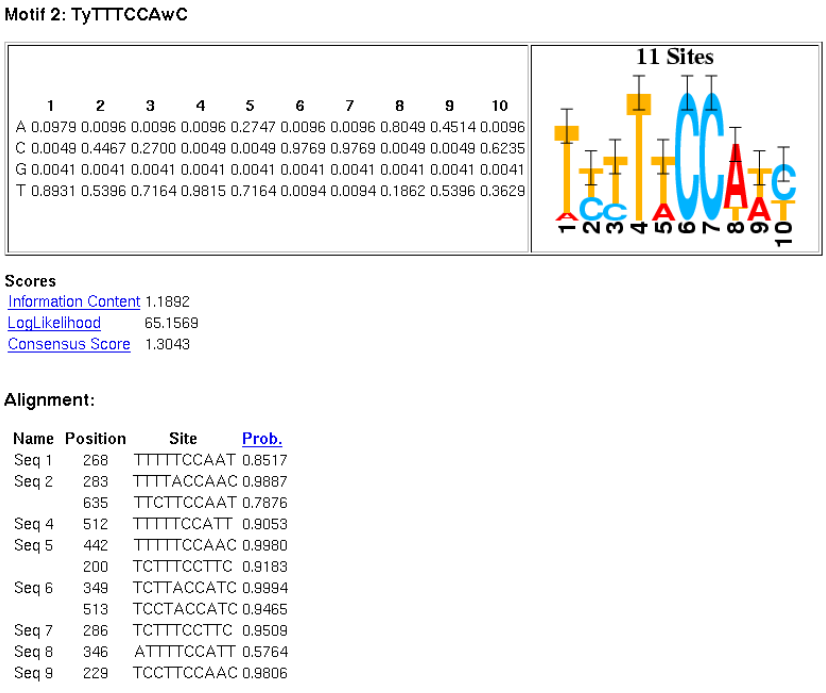
- Number of clusters: 12
- Number of genes not in a cluster: 68

**Cluster 1**



FIG.1 - Screenshot from the results page of the clustering web interface. The normalized expression profiles are shown together with the identifiers of the genes present in the cluster.

Figure 2 gives a screenshot of the part of the results page of the Motif Sampler, which displays the position probability matrix, the corresponding logo of the motif, the motif scores, and the positions of the motif instance in the sequence set.



**Motif 2: TyTTTCCAwC**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.0979 | 0.0096 | 0.0096 | 0.0096 | 0.2747 | 0.0096 | 0.0096 | 0.8049 | 0.4514 | 0.0096 |
| C | 0.0049 | 0.4467 | 0.2700 | 0.0049 | 0.0049 | 0.9769 | 0.9769 | 0.0049 | 0.0049 | 0.6235 |
| G | 0.0041 | 0.0041 | 0.0041 | 0.0041 | 0.0041 | 0.0041 | 0.0041 | 0.0041 | 0.0041 | 0.0041 |
| T | 0.8931 | 0.5396 | 0.7164 | 0.9815 | 0.7164 | 0.0094 | 0.0094 | 0.1862 | 0.5396 | 0.3629 |



**Scores**

Information Content 1.1892
LogLikelihood 65.1569
Consensus Score 1.3043

**Alignment:**

| Name | Position | Site | Prob. |
|---|---|---|---|
| Seq 1 | 268 | TTTTTCCAAT | 0.8517 |
| Seq 2 | 283 | TTTTACCAAC | 0.9887 |
| | 635 | TTCTTCCAAT | 0.7876 |
| Seq 4 | 512 | TTTTTCCATT | 0.9053 |
| Seq 5 | 442 | TTTTTCCAAC | 0.9980 |
| | 200 | TCTTTCCTTC | 0.9183 |
| Seq 6 | 349 | TCTTACCATC | 0.9994 |
| | 513 | TCCTACCATC | 0.9465 |
| Seq 7 | 286 | TCTTTCCTTC | 0.9509 |
| Seq 8 | 346 | ATTTTCCATT | 0.5764 |
| Seq 9 | 229 | TCCTTCCAAC | 0.9806 |

FIG.2 - Screenshot of the results page of the Motif Sampler.

# 5 Conclusions

We have briefly reviewed an integrated approach to motif discovery from clusters of gene expression profile. We first introduced a new clustering algorithm called adaptive quality-based clustering. This algorithm guarantees the creation of tight clusters of expression profiles and comparison with *K*-means on yeast cell cycle data showed higher enrichment in functional classes. Second, we discussed an extension of the Gibbs sampling algorithm for motif finding where higher-order models of the sequence background are incorporated, which results in a better sensitivity. Finally, these two methods are available through a user-friendly web interface called INCLUSive at http://www.esat.kuleuven.ac.be/~dna/BioI/Software.html.

# 6 Acknowledgements

Gert Thijs is research assistant with the IWT. Yves Moreau and Kathleen Marchal are post-doctoral researchers of the FWO. Prof. Bart De Moor is full time professor at the K.U.Leuven. Pierre Rouzé is Research Director of INRA (Institut National de la Recherche Agronomique, France). This research is supported by grants from several funding agencies and sources: (1) Research Council KUL: GOA-Mefisto 666, IDO (IOTA, Genetic networks); (2) Flemish Government: FWO Vlaanderen G.0115.01, G.0407.02, research communities (ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary/ Poland), IWT STWW-Genprom, GBOU-McKnow; (3) Belgian Federal Government: DWTC (IUAP IV-02 (1996-2001) and IUAP V-10-29 (2002-2006)

# Bibliography

[1] Bailey,T.L. and Elkan,C. (1995). The value of prior knowledge in discovering motifs with MEME. Proc. Int. Conf. Intell. Syst. Mol. Biol. *3*, 21-29.

[2] Cho,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D., Lockhart,D.J., and Davis,R.W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell. *2*, 65-73.

[3] De Smet,F., Mathys,J., Marchal,K., Thijs,G., De Moor,B., and Moreau,Y. (2002). Adaptive quality-based clustering of gene expression profiles. Bioinformatics, in press.

[4] Heyer,L.J., Kruglyak,S., and Yooseph,S. (1999). Exploring expression data: identification and analysis of coexpressed genes. Genome Res. *9*, 1106-1115.

[5] Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F., and Wootton,J.C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science *262*, 208-214.

[6] Sherlock,G. (2000). Analysis of large-scale gene expression data. Curr. Opin. Immunol. *12*, 201-205.

[7] Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J., and Church,G.M. (1999). Systematic determination of genetic network architecture. Nat. Genet. *22*, 281-285.

[8] Thijs,G., Lescot,M., Marchal,K., Rombauts,S., De Moor,B., Rouze,P., and Moreau,Y. (2001). A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics *17*, 1113-1122.

[9] Thijs,G., Moreau,Y., De Smet,F., Mathys,J., Lescot,M., Rombauts,S., Rouze,P., De Moor,B., and Marchal,K. (2002). INCLUSive: INtegrated CLustering, Upstream Sequence retrieval and motif Sampling. Bioinformatics *18*, 331-332.

[10] Thijs,G., Marchal,K., Lescot,M., Rombauts,S., De Moor,B., Rouzé,P., and Moreau,Y. (2002). A Gibbs sampling method to detect over-represented motifs in the upstream regions of coexpressed genes. J. Comp. Biol., in press.

[11] van Helden,J., Rios,A.F., and Collado-Vides,J. (2000). Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res. *28*, 1808-1818.

[12] Vanet,A., Marsan,L., Labigne,A., and Sagot,M.F. (2000). Inferring regulatory elements from a whole genome. An analysis of Helicobacter pylori sigma(80) family of promoter signals. J. Mol. Biol. *297*, 335-353.