Validating Clusterings of Gene Expression Data

Wim De Mulder, Martin Kuiper and René Boel

***Abstract*—We propose a measure for the validation of clusterings of gene expression data. This measure is also useful to estimate missing gene expression levels, based on the similarity information contained in a given clustering.
It is shown that this measure is an improvement over the figure of merit, an existing validation measure especially developed for clusterings of gene expression data.

***Keywords*—clustering, k-means, gene expression, validation, figure of merit.**

I. Introduction

One of the challenging fields in bioinformatics is the analysis of gene expression data, for example the inference of gene regulatory networks [1]. Typically the experimental data for gene expression analysis consists of the expression levels of a large number of genes for a small number of experimental conditions, due to the high cost of microarray experiments. This explains why clustering is a commonly used and important preprocessing step in gene expression analysis, since it ensures a reduction of dimensionality by grouping genes with similar expression behaviour together. The actual gene expression analysis can then be done on, for example, a set of representative genes, selected from each cluster.
Since there is no precise and workable definition of 'cluster' [2], different clustering algorithms can generate different clusterings, i.e. sets of clusters, and even a given clustering algorithm can produce different clusterings for different values of initial parameters (e.g. the number of initial centers). This necessitates the use of validation measures to compare different clusterings and to select one that is appropriate for the application under consideration.
An interesting validation measure in the context of gene expression analysis is the figure of merit [3]. In essence, this measure applies a clustering algorithm to all but one experimental condition in a data set, and the left-out condition is used to assess the predictive power of the clustering algorithm. This predictive power is then used as validation measure for the considered clustering algorithm. The figure of merit is more extensively explained in section II.
Instead of using the figure of merit to validate a clustering algorithm, we use it to validate a clustering. As explained above these are two different things, since a given clustering algorithm can still generate different clusterings. Thus having an appropriate clustering algorithm does not guarantee that an acceptable clustering will be produced. In our opinion it is therefore more relevant to compare different clusterings, generated by whatever clustering algorithm(s), if the final goal is not to compare clustering algorithms. In section III we describe the use of the figure of merit to validate clusterings of gene expression data, while in section IV some improvements are given. Our methodology is then applied to several data sets. These data sets are described in section V and the results are analyzed in section VI.
Two questions are addressed in this paper: 1. is the proposed measure an improvement over the figure of merit for estimating missing gene expression levels and 2. is it an improvement for selecting an appropriate number of clusters?

Wim De Mulder is with the Systems Research Group, University of Ghent, Ghent, Belgium, email: wim.demulder@ugent.be
Martin Kuiper is with the Systems Biology Group, NTNU, Trondheim, Norway, email: kuiper@bio.ntnu.no
René Boel is with the Systems Research Group, University of Ghent, Ghent, Belgium, email: rene.boel@ugent.be

II. Related Work

Over time a range of validation techniques for clusterings have been developed [4]. In this paper we concentrate on the figure of merit [3], which is motivated by the jackknife approach. Suppose a data set D consisting of the expression profiles of genes $g_1, \ldots, g_n$ is given. We shall restrict attention to the case where the expression profiles are time series, although the figure of merit is applicable to any kind of experimental conditions. Denote the time points by $t_1, \ldots, t_m$ and the expression level of g_i at time point t_j by $g_i(j)$. The data set D can thus be described as: $D = \{g_i(j) \,|\, i = 1, \ldots, n, j = 1, \ldots, m\}$. Define $D \setminus t_p = \{g_i(j) \,|\, i = 1, \ldots, n, j = 1, \ldots, p-1, p+1, \ldots, m\}$.
Now a clustering algorithm is applied to $D \setminus t_p$. Suppose that there are k clusters $C_1, \ldots, C_k$ and let $\mu_{C_l}(p)$ be the average expression level for time point t_p of genes in cluster C_l, i.e. $\mu_{C_l}(p) = 1/|C_l| \sum_{g_i \in C_l} g_i(p)$, where $|C_l|$ denotes the number of genes in C_l. If the clusters are meaningful, similar genes should be grouped together, implying that the average expression level $\mu_{C_l}(p)$ should be close to the expression levels $g_i(p), g_i \in C_l$ (where it is implicitly understood that the data set is normalized appropriately). The figure of merit, $FOM(p,k)$, for k clusters using time point t_p as validation is then defined as

$$FOM(p,k) = \sqrt{1/n \sum_{i=1}^{k} \sum_{g_i \in C_l} (g_i(p) - \mu_{C_l}(p))^2} \quad (1)$$

The *aggregate* figure of merit is an estimate of the total predictive power of the algorithm over all time points:

$$FOM(k) = \sum_{p=1}^{m} FOM(p,k) \quad (2)$$

In [3] it is noticed that the FOM shows a declining figure with the number of clusters, for all examined data sets. To overcome this bias a compensating factor is introduced, giving

the *adjusted* figure of merit $FOM_A(k)$:

$$FOM_A(p,k) = \sqrt{n/(n-k)}\, FOM(p,k) \quad (3)$$

$$FOM_A(k) = \sum_{p=1}^{m} FOM_A(p,k) \quad (4)$$

However this adjusted FOM still shows a significant bias towards the number of clusters and the authors of [3] state that "it is not safe to compare clustering results with different numbers of clusters".

III. Use of the figure of merit to validate clusterings

Instead of using the ideas behind the figure of merit to compare different clustering algorithms, we use them to compare different clusterings. This implies that given clustering algorithms are applied to the complete data set D. Given a set of clusterings, possibly produced by different clustering algorithms, the question is then how to choose among them.
To validate these clusterings we rely on the basic idea behind the figure of merit: if the clusters of a considered clustering are of high quality, meaning that each cluster contains very similar genes, it should be possible to make a reliable estimate of a missing gene expression level $g_i(p)$ based on the expression levels of the other genes in the same cluster at the same time point. Notice that to validate a given clustering it is only *supposed* that gene expression levels are missing. In fact the true gene expression levels are needed to define the FOM as the root mean square deviation of the estimated expression levels from the true expression levels, see (2).
Practically speaking, the only difference with the FOM outlined in section II is that the clustering algorithm (or algorithms) is applied to the complete data set. This means that formulas (1)-(4) remain valid.
Since the FOM represents the error in estimating gene expression levels, based on the similarity information contained in the given clustering, it is clear that the lower the FOM the higher the quality we ascribe to this clustering.

IV. Improvements on the figure of merit

A. Improvement 1

A first improvement on the figure of merit we propose is based on the observation that the most common measure for coexpressed genes is the correlation; in [6] it is argued that this measure is better suited to detect coexpressed genes than, for example, Euclidean distance. For this reason, we define the distance between two genes g_i and g_j as $1 - c(g_i, g_j)$, where $c(g_i, g_j)$ denotes the correlation between the time series of g_i and g_j:

$$c(g_i,g_j) = \frac{\sum_{t=1}^{m}(g_i(t)-\bar{g_i})(g_j(t)-\bar{g_j})}{\sqrt{\sum_{t=1}^{m}(g_i(t)-\bar{g_i})^2}\sqrt{\sum_{t=1}^{m}(g_j(t)-\bar{g_j})^2}} \quad (5)$$

where $\bar{g_i}$ and $\bar{g_j}$ denote the average expression level of g_i and g_j over time. This distance measure is used in clustering the data sets from section V.
Now, the estimate for $g_i(p), g_i \in C_l$, is given by $\mu_{C_l}(p)$ (see section II), the average expression level at time point t_p of all genes belonging to C_l. However, this estimate is not consistent with the correlation distance used to cluster the genes. For two genes can have a high correlation, while their expression levels are very different, since the correlation distance takes the shape of the genes into account and not their expression levels. Differently stated: if given genes are similar in terms of correlation or 'behavior', they need not be similar in terms of expression levels.
Normalizing the data, for example using the z score transformation, can only partly solve this problem, since it is still possible that similar genes have incomparable expression levels. This follows from the fact that an outlier in the original data set is still an outlier in the z-transformed data set.
Our idea is to make the expression levels of two genes similar, while the correlation between them, i.e. their shape, is unaltered. Observe that from formula (5) it follows that $c(g_i, a_j\, g_j + b_j) = c(g_i, g_j)$, where $a_j \in \mathbb{R}$ and where we define $b_j = (b_j, \ldots, b_j)^T \in \mathbb{R}^m$. Given that we want to estimate $g_i(p), g_i \in C_l$, we transform the genes $g_j \in C_l$ such that their expression levels are as close as possible to the expression levels of g_i, while the correlation between g_j and g_i is unaltered. Thus the following is to be minimized:

$$E(a_j,b_j) = 1/2 \sum_{t=1}^{m}(a_j\, g_j(t) + b_j - g_i(t))^2$$

The necessary condition for this is that the partial derivatives equal zero:

$$\frac{\partial E}{\partial a_j} = a_j \sum_{t=1}^{m} g_j^2(t) + b_j \sum_{t=1}^{m} g_j(t) - \sum_{t=1}^{m} g_i(t) g_j(t) = 0 \quad (6)$$

$$\frac{\partial E}{\partial b_j} = a_j \sum_{t=1}^{m} g_j(t) + m\, b_j - \sum_{t=1}^{m} g_i(t) = 0 \quad (7)$$

The solution is given by

$$a_j = \frac{m\sum_{t=1}^{m} g_i(t)g_j(t) - \sum_{t=1}^{m} g_j(t)\sum_{t=1}^{m} g_i(t)}{m\sum_{t=1}^{m} g_j^2(t) - (\sum_{t=1}^{m} g_j(t))^2} \quad (8)$$

$$b_j = \frac{\sum_{t=1}^{m} g_i(t) - a_j \sum_{t=1}^{m} g_j(t)}{m} \quad (9)$$

Denote the transformed genes by $\tau_i(g_j) = a_j g_j + b_j$ with a_j and b_j given by (8) and (9). The subscript i in τ_i stresses the point that the transformation is done with g_i as 'reference'.
We then estimate $g_i(p)$ as $\tau_i(\mu_{C_l}(p))$, defined as:

$$\tau_i(\mu_{C_l}(p)) = \frac{1}{|C_l|} \sum_{g_j \in C_l} \tau_i(g_j)(p)$$

The transformation τ_i thus ensures that the similarity between g_j and g_i in terms of expression levels is maximized, while their similarity in terms of correlation remains constant. In particular a given clustering is thus not influenced by this transformation, since the distances between each pair of genes is the same before and after this transformation.

B. Improvement 2

An important cause of the bias towards the number of clusters is probably the fact that for the estimation of $g_i(p)$ with $g_i \in C_l$ *all* expression levels at time t_p of genes belonging

to C_l are used, thus including $g_i(p)$ itself. Since for the purpose of validation it is supposed that $g_i(p)$ is unknown, it is in fact not allowed to use $g_i(p)$ in the estimation of it. Notice that as the number of clusters grow, the number of genes per cluster declines (on average), implying that the relative weight of $g_i(p)$ in the estimation of it increases. This follows from the fact that $g_i(p)$ is estimated as $\mu_{C_l}(p)$, defined as $1/|C_l| \sum_{g \in C_l} g(p)$. In the most extreme case $|C_l| = 1$, i.e. $C_l = g_i$, and thus $\mu_{C_l}(p) = g_i(p)$ implying a perfect estimate. This is of course an unfair method to make estimations and it explains how the bias towards the number of clusters arises.
Thus we redefine $\mu_{C_l}(p)$ as follows: $\mu_{C_l}(p) = 1/(|C_l| - 1) \sum_{g \in C_l \setminus \{g_i\}} g(p)$. If $|C_l| = 1$ we can estimate $g_i(p)$ as, for example, the average of $g_i(p-1)$ and $g_i(p+1)$ if $1 < p < m$, and as $g_i(2)$ if $p = 1$ and as $g_i(m-1)$ if $p = m$.

C. Improvements 1 and 2 together

The improvements proposed in sections IV-A and IV-B can be combined by defining $\tau_i(\mu_{C_l}(p))$ as follows:

$$\begin{aligned}\tau_i(\mu_{C_l}(p)) &= \frac{1}{|C_l| - 1} \sum_{g_j \in C_l \setminus g_i} \tau_i(g)(p) \quad \text{if } |C_l| > 1 \\ &= \frac{g_i(p-1) + g_i(p+1)}{2} \quad \text{if } |C_l| = 1,\ 1 < p < m \\ &= g_i(2) \quad \text{if } |C_l| = 1,\ p = 1 \\ &= g_i(m-1) \quad \text{if } |C_l| = 1,\ p = m\end{aligned}$$

We thus redefine (1) as a new measure H:

$$H(p) = \sqrt{1/n \sum_{i=1}^{k} \sum_{g_i \in C_l} [g_i(p) - \tau_i(\mu_{C_l}(p))]^2} \quad (10)$$

$$H = \sum_{p=1}^{m} H(p) \quad (11)$$

V. Data sets

A. Artificial data

Since the validation measure is intended to use for clusterings of gene expression data and since attention in the previous sections was directed towards time series data, our goal here is to generate artificial data that mimics gene expression time series. In particular we simulate periodically expressed genes by introducing some noise to the sine function $f(x) = \sin(x)$. Suppose that we want to generate k clusters of n genes with m time points. Predefining the number of clusters allows us to test whether the FOM and the proposed measure H are able to detect this number of clusters.
We then define k possible translations of the independent variable: $T_x = \{(j2\pi)/k \,|\, j \in \{0, \ldots, k-1\}\}$. For each artificial gene g a translation is randomly selected from T_x, denoted as $T_x(g)$. Each artificial gene g is also randomly translated in the direction of the Y axis. This translation is denoted as $T_y(g)$ and is chosen to be uniformly distributed between 1 and 3. Furthermore for each gene g a random amplitude $A(g)$ is generated, also uniformly distributed between 1 and 3.
Thus the time series for an artificial gene g is given by $f_g(x) = A(g)\sin(x + T_x(g)) + T_y(g)$.
Finally, from a given time series f_g we have to select m time points. For this selection we introduce two levels of randomness: one on the level of the gene and one on the level of the time points. For each gene g a random number $R_1(g)$ is generated, Gaussian distributed with mean 0 and standard deviation σ_1. For each gene g *and* each pth time point, $p = 1, \ldots, m$, a second Gaussian number $R_2(g, p)$ is generated with mean $p - 1$ and standard deviation σ_2. The value of the pth time point, $t_p(g)$, is then defined as $t_p(g) = R_1(g) + R_2(g, p) + T_x(g)$. Whereas $R_1(g)$ is fixed for each time point, the value of $R_2(g, p)$ is dependent on the time point.
The value $g(p)$ is given by $g(p) = A(g)\sin(t_p(g)) + T_y(g)$.

B. Biological data set: S. pombe

We considered also a real gene expression data set: the Schizosaccharomyces pombe cell cycle data set reported in [5]. The data set consists of three elutriations and 20 time points. We randomly selected 150 genes for which all expression levels over all time points and all elutriations are non missing. For each gene the expression level at a certain time point is defined as the average of its time points over the elutriations.

VI. Results and discussion

Figures 1 and 2 show H (11), the FOM (2) and the adjusted FOM (4) as a function of the number of clusters.
Clusterings were generated for the data sets described in section V. All these data sets were normalized to have mean 0 and variance 1, as in [3].
The clusterings were generated with the k-means algorithm [2], using the implementation of Forgy [7]. The distance measure used to cluster the data is the correlation distance, see section IV-A. K-means requires to predefine the number of initial centers; we vary this number from 2 to 15. Forgy k-means allows for empty clusters, implying that the number of clusters does not necessarily equal the number of initial centers. Thus in figures 1 and 2 the number of times that a certain number of clusters was constructed is also shown, displayed as the percentage of the total number of clusterings generated. It is thus possible that a certain number of clusters is never generated, even if the corresponding number of initial centers lies in the range from 2 to 15. If this is the case, we arbitrarily give H the maximum value of all the available values for H (i.e. for which the corresponding percentage of clusters is non zero). The same is done for the FOMs.

A. Artificial data

For all artificial data sets we choose $n = 60, m = 18$ and $k = 3$. The values for the parameters σ_1 and σ_2 were chosen from $\{0.025, 0.075\}$, giving four possible combinations. An example of an artificial data set with $\sigma_1 = 0.025, \sigma_2 = 0.075$ is shown in Fig 3. For each combination of σ_1 and σ_2, 10 data sets were generated and the resulting values for H, the FOM and the adjusted FOM were averaged over these data sets. The results are shown in figure 1.
The most obvious observation is that the value of H is much

lower than that of the FOMs. Thus if the purpose is to estimate the value of missing expression levels, based on the information of a given clustering, the measure H is of much more use than the FOM.
Secondly, although the FOMs do not show a bias for the artificial data sets, they do not show a clear minimum either. Thus if a number of clusters is to be chosen, solely based on the FOMs in figure 1, it would not be a reassuring choice. Furthermore, the number of clusters at the minimum FOM and minimum adjusted FOM is for all data sets different from 3. Even for the data set with the lowest level of noise ($\sigma_1 = \sigma_2 = 0.025$), both FOMs give an unexpected (and probably undesired) optimal number of clusters of 9.
On the other hand, the measure H is minimal at three clusters for $\sigma_1 = \sigma_2 = 0.025$ and $\sigma_1 = 0.025, \sigma_2 = 0.075$. Notice that for these two data sets, H shows a sharp decline from 2 to 3 clusters, where it reaches a minimum and then shows a ascending behavior. This 'convex behavior' would be much more pronounced if we would display only H on the figure.
Although the minimal H for $\sigma_1 = 0.075, \sigma_2 = 0.025$ and $\sigma_1 = \sigma_2 = 0.075$ appears at 5 clusters, the value at 3 clusters is also low. An interesting observation is that for these two data sets H does not show a clear minimum. This can possibly indicate that there is either no number of clusters that is significantly better than other numbers of clusters or that either the clustering algorithm is not able to detect the most appropriate number of clusters. This last possibility follows from the fact that a given clustering algorithm does not necessarily recognize the 'true' number of clusters. If this is the case, i.e. if our Forgy k-means constructs clusterings consisting of three clusters of low quality, it is not to be expected - and even undesired - that the considered validation measures consider these clusterings as superior. Further research with several clustering algorithms and more data sets is needed to hypothesize whether the behavior of H for the last two artificial data sets is due to a deficiency in the considered clustering algorithm, due to a deficiency in H or that too much noise is present to have a number of clusters that is significantly better than others. This last possibility is unlikely, since by visual inspection we concluded that for these noise levels there are still three well-separated clusters.

B. Biological data

The results for the biological data set are shown in Fig 2. For each number of initial centers the clustering algorithm was applied ten times, since the resulting clustering is dependent on the exact choice of initial centers. The averaged values of H and the FOMs are shown.
Since the real number of clusters is not known, it is impossible to state whether the validation measures perform well in this case. However, it is interesting to notice that all three validation measures have their optimum in the same number of clusters, 14.

VII. Conclusion

We present a validation measure to compare different clusterings, produced by whatever clustering algorithm(s), in case of gene expression data. Our methodology is an improvement over the figure of merit, a measure that assess the predictive power of a clustering or clustering algorithm, based on the information contained in a given clustering. The most obvious improvement of the proposed measure over the figure of merit is that it gives much better estimates of missing expression levels. For the examined artificial data sets it also gives the optimal number of clusters, if the level of noise is low.

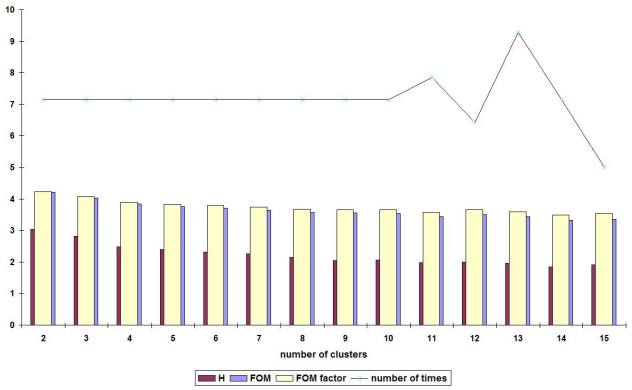


Fig. 2. Validation measures on real biological data

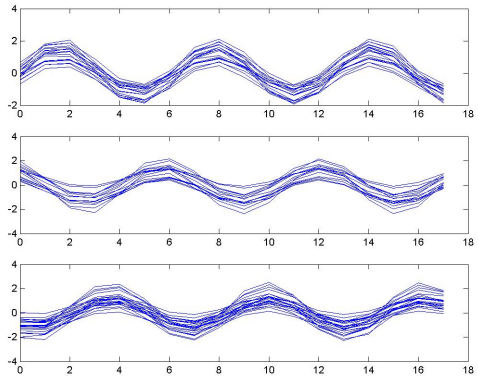


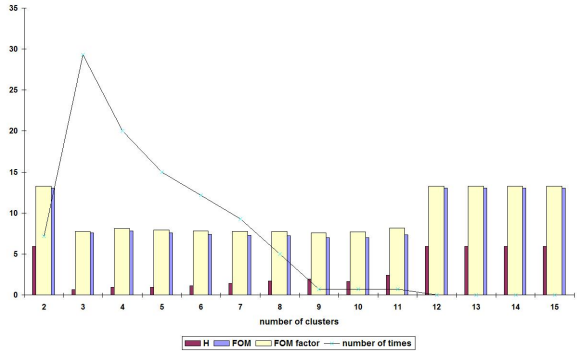
Fig. 3. Example of artificial data set with $\sigma_1 = 0.025, \sigma_2 = 0.075$

Acknowledgments

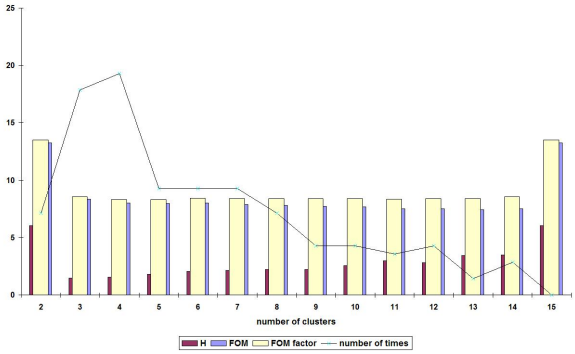
Financial support from BOF (Special Research Fund) is gratefully acknowledged.

References

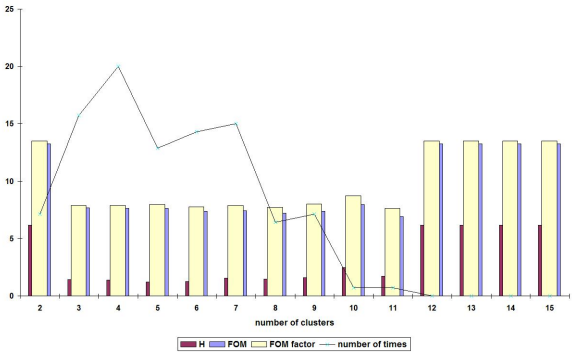
[1] J. Supper, H. Fröhlich, C. Spieth, A. Dräger and A. Zell, *Inferring gene regulatory networks by machine learning methods*, Proceedings of the 5th Asia-Pacific Bioinformatics Conference, pp. 247-256, 2007.
[2] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*, Prentice Hall, 1988.
[3] K. Y Yeung, D. R. Haynor and W. L. Ruzzo, *Validating clustering for gene expression data*, Bioinformatics, vol. 17, pp. 309-318, 2001.



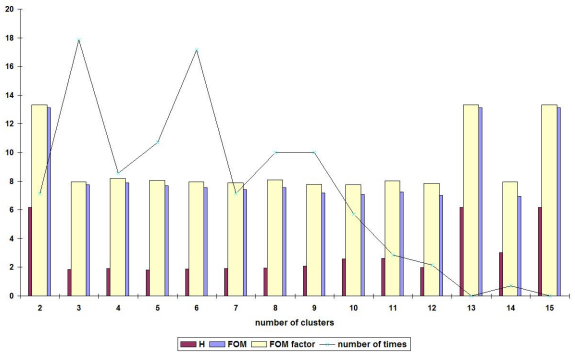
(a) $\sigma_1 = 0.025, \sigma_2 = 0.0251$



(b) $\sigma_1 = 0.025, \sigma_2 = 0.075$



(c) $\sigma_1 = 0.075, \sigma_2 = 0.025$



(d) $\sigma_1 = 0.075, \sigma_2 = 0.075$

Fig. 1. Validation measures on artificial data

[4] M. Halkidi, Y. Batistakis and M. Vazirgiannis, *On clustering validation techniques*, Journal of Intelligent Information Systems, vol. 17, pp. 107-145, 2001.

[5] G. Rustici et al., *Periodic gene expression program of the fission yeast cell cycle*, Nature Genetics, vol. 36, pp. 809-817, 2004.

[6] L. J. Heyer, S. Kruglyak and S. Yooseph, *Exploring expression data: identification and analysis of coexpressed genes*, Genome Research, vol. 9, pp. 1106-1115, 1999.

[7] E. W. Forgy, *cluster analysis of multivariate data: efficiency versus interpretability of classification*, Biometrics, vol. 21, pp. 768-769, 1965.

Wim De Mulder received a M.Sc. degree in Computer Science (2004) from Ghent University, and is currently working on his PhD with as topics clustering and the modeling of gene regulatory networks.

Martin Kuiper received a M.Sc. degree in Molecular Biology and Biochemistry (1982), and a PhD degree (1987) in Biology from the University of Groningen, Netherlands. He was a doctoral fellow of the Ohio State University, US (1987) and Utrecht University, Netherlands (1989). After 9 years in industry he joined the VIB institute in Gent, Belgium, as PI in Computational Biology. He transferred to the Norwegian University of Science and Technology in 2008, where he now works as Professor in Systems Biology. His research interests include the development of approaches and tools for analysis of biological data, and the use of semantic web technologies for integration of biological knowledge.

René Boel received a degree in electromechanical engineering (1969) and in nuculear engineering (1970) from Ghent University, an M.Sc. (1972) and a Ph.D. degree (1974) in Electrical Engineering and Computer Sciences from the University of California, Berkeley. He has held temporary appointments at the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, at the Mathematics department of the Katholieke Universiteit Leuven, at the Mathematics and Statistics department of Bell laboratories Murray Hill (N.J.), at the Department of Systems Engineering, Australian National University, Canberra, at the department of Electrical and Computer Engineering, University of Newcastle, Australia, and at the Electrical and Electronic Engineering department, Melbourne University. He is currently a professor at the School of Engineering, Ghent University (Belgium). His research interests are in the fields of distributed and supervisory control and in estimation and fault detection for stochastic, discrete event and hybrid systems, with applications to traffic networks and power systems, as well as in bio-informatics.