Towards an integrated usage of expression data and domain literature in gene clustering: representations and methods

Peter Antal[*] **Patrick Glenisson**[*] **Tom Boonefaes**[+]
Geert Fannes[*] **Bart De Moor**[*] **Pieter Rottiers**[+]
Tamas Meszaros[#] **Yves Moreau**[*] **Johan Grooten**[+]

[ ]

Abstract

Due to the complexity of the underlying processes, the scarcity of the data and the high level of noise on the measurements, understanding large-scale gene interactions from expression data demands methodologies that incorporate domain literature with gene clustering. These methodologies should more and more comply with the circular nature of the overall clustering process, in which the human expertise iterates between the measured statistical data and the domain literature. In this paper we propose *Gene**, a genetic literature data structure representing the relevant literature for each gene based on its GeneCard and Medline abstracts. The *Gene** representation is appropriate for information retrieval and constructing textual profiles of a gene cluster. Based on this, we describe a system providing a textual profile for a gene cluster and supporting the corresponding information retrieval and revision steps according to the cyclic nature of the clustering process. We present results about clustering of genes based solely on their *Gene** literature representation and on the dual usage of this literature-based clustering with gene clustering based on expression data.

1 Introduction

The dissemination of microarray technology, providing the possibility of simultaneous measurement of the activity of large numbers of genes, gave new stimuli towards a broader understanding of their functional roles and their interactive networking. The complexity of these fundamental biological questions though, demand methodologies that allow a flexible organization of human expertise, the measured statistical data and the domain literature.

One of the main research directions in the post-genomic age is the clustering of genes to understand the regulatory patterns in their interactions, the pathways they belong to or the genetic network they constitute (for an overview, we refer to [8]). To date, the interpretation and validation of gene expression clusters remains a time-consuming manual task based on browsing the genetic literature and hypothesizing about the relationships within the clusters based on the data. While it is feasible for a small set of genes, this practice does not scale up to a medium or genome-wide set.

However, the availability of large amounts of electronic domain literature has changed the traditional way of data analysis. Nowadays a huge amount of domain knowledge is available in either unstructured or structured electronic format (e.g., natural language papers and structured documents, databases or knowledge bases respectively). Many efforts in the bioinformatics community are currently directed towards the development of machine-friendly representations, standardization of the state-of-the-art biological information resources and the establishment of cross-linked or combined information sources. Despite these efforts, much of the domain-specific

(*)Dept. of Electrical Eng., Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Heverlee (Leuven), Belgium, (+)Dept. of Molecular Biology, Ghent University and Flanders Interuniversity Institute for Biotechnology, K. L. Ledeganckstraat 35, B-9000 Ghent, Belgium, (#)Dept. of Measurement and Inf. Systems, Budapest Univ. of Tech. and Economics, Muegyetem rkp. 9, H-1521 Budapest, Hungary

information remains present in natural language text or scattered around different repositories, because of the rapid development of the domain.

The expert's environment can be characterized as split up in the *data world*, which contains the high-throughput data and statistical analysis methods, and the *knowledge world* which contains the domain knowledge dominantly present in free-text form. Within this terminology, the data analysis can be seen as deep interaction of human expertise with those two worlds. To increase the efficiency of this interaction, post-genomic intelligent systems should overcome the current artificial separation of the two worlds (i.e., separation between tools for data analysis and those for information retrieval) by following integrated approaches. Gradual steps in removing the barriers between the expression data analysis and literature oriented methods can be the following:

1. *Intelligent support.* More customized literature-based methods (e.g., information retrieval methods using contextual information from the user).
2. *Semi-automatic support.* Semi-automatic literature-based methods for interpreting the results of data analysis (e.g., automatic derivation of a textual profile for a gene cluster coming from the data analysis).
3. *Integrated approach.* Integration of methods based on the analysis of expression data and domain literature (e.g., comparison of clusterings of genes based on the measured data and the corresponding literature).
4. *Literature-data approach.* Methods using the domain literature in the same way as the expression data, possibly after complex transformation of the textual domain knowledge into a more numerical, statistical format (e.g., clustering of genes based on a dual vector representation for each gene, composed of relevance factors for domain terms and the expression levels in measurements).

This article addresses the first three issues from an integrated point of view on the data and domain knowledge, maintaining flexibility towards both the rapidly evolving field and expert interventions.

The paper is organized as follows: Section 2 summarizes the existing literature-based approaches for supporting the expression data analysis. In Section 3, an iterative, cyclic model of the overall clustering process is shown. Section 4 presents *Gene**, the so-called 'genetic literature data' representation of the domain literature and our semi-automatic system based on textual profiling of gene clusters to support this circular, iterative nature of the clustering process. Section 5 contains results about clustering the derived 'genetic literature data' and its relation to clustering expression data. Finally, Section 6 concludes and describes our ongoing work on this subject.

2 Literature-based microarray data analysis

The dominant strategies to support the expression data analysis by literature, rely on the fact that the statistical data analysis and the biological interpretation of the results are separated (i.e., the usage of statistical packages versus database or information retrieval tools). Consequently, most of the efforts are devoted to the compilation of the available genetic literature into cross-linked databases to support the structured and unstructured information retrieval. For the biomedical sciences in general, the US National Library of Medicine's Medline is a common bibliographic source of citations and abstracts ranging from 1966 to present. More recently created and focused on the human genome, the GeneCard system [6] offers a web-based, comprehensive and constantly updated knowledge repository for each human gene. These curated databases form basic tools for the manual interpretation of expression data.

A first step to automate the manual investigation of various domain knowledge repositories is serializing frequent subsequent steps as reported in [4]. It involves the extension of queries with terms coming from the GeneCards repository before submitting them to the Medline search engine. Following the four groupings above such methods can be classified as *intelligent support* for literature-based expression data analysis.

Another information retrieval-based approach is the use of domain-specific concepts and term hierarchies (or ontologies). Through their structure, ontologies can represent the semantics of the domain. In [2] for example, the GeneOntology (GO) Consortium demonstrates the potential of GO for interpreting high-throughput genomic data. Masys et al. [3] designed an interface that combines the Medline query engine with both the MeSH keyword list and the UMLS ontology to infer the function of a group of genes. The MeSH (Medical Subject Headings) is a controlled vocabulary used for indexing the abstracts included in Medline, while the UMLS (Unified

Medical Language Systems) is a biomedical concept ontology conceived to preserve semantic relations between the concepts described in its controlled vocabularies (more details available at http://www.nlm.nih.gov/databases/). The interface in [3] reports the quantitative significance of each result and provides links to different databases allowing further browsing. Such methods can be interpreted as *semi-automatic support* for literature-based expression data analysis.

The most ambitious methods, the *integrated approaches*, try to directly infer functional relationships between genes and their products from textual information. We can distinguish two important approaches. Rindflesch et al.[7] follow the deep-linguistic approach where Natural Language Processing forms the basis of the knowledge discovery process. An alternative way is to adopt the shallow statistical stance which basically discards the linguistic structural information and uses only statistics of individual terms in documents and the corpus. Our focus will be directed towards this approach. Shatkay et al. [9] for example derive a gene-to-gene similarity measure by selecting the N most relevant documents from a corpus of Medline abstracts for each gene. PubGene [5] on the other hand is the first on-line system that generates gene networks based on the *full* set of Medline abstracts, all indexed in a simplified, but scalable way. Finally, Stapley et al. [10] report on a similar, but prototypical application that is based on a binary gene-by-document matrix. It outputs the similarities between the genes that are related to a user-specified MeSH term.

3 An integrated approach for using data and literature in gene clustering

As the previous survey demonstrates, one of the main current trends is the increasingly close incorporation of the literature in the gene clustering process.

To design efficient *intelligent*, *semi-automated*, or *integrated* methods, requires the identification of a simple, but powerful model of the clustering process. Fig. 1 illustrates the overall cyclic process involving (1) clustering based on the data, (2) interpreting the clusters, (3) evaluating cluster members and potential non-members and finally (4) modifying the experimental setup, the parameterization of the clustering algorithm, and so on to return to the first step.

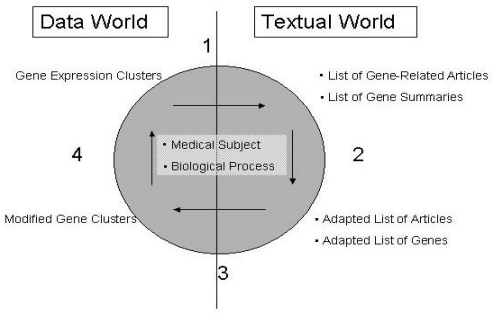

Figure 1: The interaction loop between the statistical data world and the literature world

The research into clustering strategies for high-throughput data, most notably microarray data, is a top priority in the current post-genomic stage. The algorithms and their evaluation are increasingly taking into account concepts as statistical cluster quality, a priori confidence levels and predefined causal relations (cfr. [8]). However, attaching biological context to the computed clusters remains a major bottleneck like ruling out biological discrepancies, peculiarities or experimental errors. Although current efforts to collect related genetic information (or links to them) into centralized repositories has proved extremely valuable, they (being a closed system) lack the interactive component needed to fine-tune the retrieval or discovery process. To our knowledge, the existing *intelligent* and *semi-automated* systems operate only unidirectionally, lacking the possibility to support the circular, iterative nature of the clustering process. As a continuation, the rationale behind our approach to use data and literature in clustering, is to cover and support the shown cyclic process.

The basis of our system is a representation of the literature for each known human gene, which allows an automatic derivation of a textual profile for a set of genes and clustering of genes based on this literature representation. The 'textual profile' of a gene cluster consists of relevant terms with relevance and a stability factors corresponding to this cluster (see later).

The potential of these functionalities in the iterative process can be demonstrated on the following paradigmatic examples:

1. Investigation of a gene cluster of interest to the researcher
2. Investigation of a gene cluster identified by an expression-based clustering algorithm
3. Simultaneous Investigation of multiple gene clusters identified by an expression-based clustering algorithm

Firstly, analyzing a cluster of genes for *exploration* aims at retrieving highly relevant keywords for the set of genes of interest to the researcher. The returned profile identifies either the most important or most stable terms for the gene cluster from a selected vocabulary. Since these can be an excellent starting point for further literature exploration, the gene cluster profiling is directly integrated with an information retrieval system capable of accepting queries based on the resulting gene cluster profile. The subsequent browsing of the returned documents of the information retrieval system drives the user to a better understanding of the clusters.

Secondly, analyzing a cluster of genes directly obtained from expression analysis could aim at providing feedback for the clustering algorithm. While in the previous part genes were not necessarily connected to the data world (i.e., no expression information was considered), here the set of genes is defined by a clustering algorithm. As the number of members per cluster typically lie around the one-hundred range, the scientist will rather skim the reported profiles, than thoroughly check the literature for each gene. This global inspection of textual profiles enables to detect outliers – genes assigned to a cluster in the data world, but excluded from it by the literature. Such outlying gene can suggest a flaw in the parameterization of the cluster algorithm or can represent a new discovery. Depending on the outcome, the user can iterate back to the data world to repeat the analysis or check certain findings and, if necessary, the loop can be entered again.

Thirdly, analyzing several clusters in general aims at performing a 'dual' clustering: to compare a clustering based on the expression data versus domain knowledge. Till now, the domain knowledge-based clustering was implicit and based on human expertise. Our goal at this point is to provide an explicit, literature-based clustering of genes ; based on the comparison of the clusterings derived from data and literature, clusters can be adapted and alternative cluster strategies or parameterizations can be considered (see also Fig. 1).

4 Automatic textual analysis of gene clusters

In this section we will briefly summarize the main features of our experimental system based on the integrated approach for the usage of expression data and genetic literature.

4.1 A 'genetic literature data' representation of genetic literature: *Gene**:

Our long-term purpose is to exploit the genetic literature inside various statistical methods, hereby giving it an equal status as experimental data. Therefore we call representations suitable to this end genetic literature data.

The most difficult question in deriving genetic literature data is how to collect, for a given gene, the most relevant documents from a large collection of related documents to construct its literature representation. We used a kernel-based methodology, a simplified version of [9], where for each gene its corresponding GeneCard card was chosen as the literature kernel. This gene kernel functions as a query document for selecting the most relevant Medline abstracts.

In each phase, the documents (either Medline abstracts or GeneCards) are represented by the relevance factors of 70.000 terms stemming from controlled vocabularies including the GO [2] and the MeSH ontologies. The relevance factors are computed by the TF-IDF weighting scheme [1] and the cosine similarity measure was used to select the most relevant Medline abstracts.

The derived genetic literature data – *Gene** – for each gene is thus based on the corresponding GeneCard and the N most related Medline abstracts. Based on this representation, similarity matrices were generated which were used as input for two well-known clustering algorithms: hierarchical and k-medoids.

Table 1 summarizes possible alternatives and parameterizations for the representation, actually present in *Gene**. Almost each of the shown parameterizations can be combined with each other.

Representation	Top N	Similarity	Clustering
binary	0	matching coefficient	hierarchical
DF-IDF	5	cosine	k-medoids
	20		

Table 1: Parameterizations for the generation of *Gene**, a genetic literature data. Listed are the specific vector representation, the number of incorporated Medline documents, the similarity measure used to select the most relevant Medline documents and to compute the gene-to-gene similarity matrix and the clustering method applied on the similarity matrix.

Since it supports both the generation of textual profiles for gene clusters and the generation of clusters purely on the literature, the *Gene** representation can be used optimally for an integrated usage of data and literature in the gene clustering process.

4.2 An iterative scheme based on textual profiles of gene clusters

The first applicability of *Gene** in the overall gene clustering process is to generate profiles for gene clusters as described in Section 3. A more elaborated iterative scheme is presented in Fig. 2, which follows the circular nature of this process shown previously in Fig. 1.

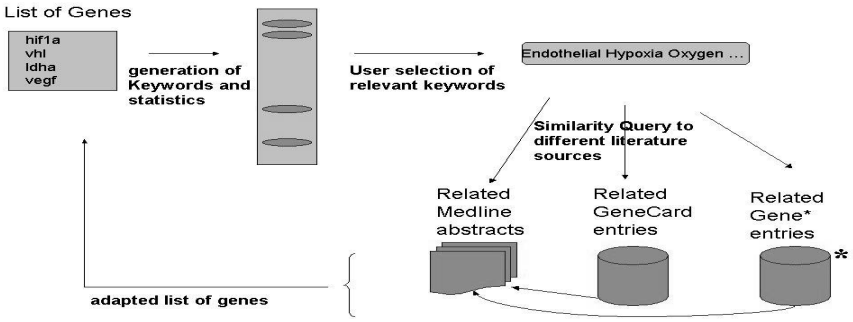

Figure 2: *Gene**-based interaction between statistical data world and literature world

Based on *Gene**, the system computes the mean and variance of the relevance factors of the terms corresponding to the representations of the genes in the given cluster. Taking into account either the average or the variability of the relevance factors of the terms, the user can select a subset of keywords to query the Medline collection, the GeneCards or - again - the *Gene**. Two advantages of using a non-binary representation here, are its ability to generate more detailed term statistics and its flexibility to encapsulate the binary approach. However, the system remains open in the sense that the user can select the underlying genetic literature data representation from the alternatives presented in Table 1.

Preliminary, qualitative results about the usefulness of this system are encouraging. For example, in an experiment, four highly relevant genes for hypoxia (oxygen deprivation) were selected (being HIF1A, VHL, VEGF and LDHA). Then, from the derived textual profile for this cluster, relevant keywords were selected and used to

successfully retrieve related genes (such as Tie-1 and Tie-2) based on *Gene**. However, this application still needs a thorough quantitative evaluation, so it is not covered further in this paper. We continue with the evaluation of the applicability of *Gene** for generating gene clusters based on the literature.

5 Towards full integration: clustering the 'genetic literature data'

Beside the full GeneCard database, we used a collection of 30.000 Medline abstracts dated from 01/1992 to 11/2000. The abstracts come from ten journals selected according to their impact factor and their relevance as assessed by a biologist in the field of Molecular Oncology. The rationale behind using a restricted subset of journals, but widening the temporal scope, was both to keep our prototype system in the medium-scale range and to eliminate noise expected from unrelated information sources by keeping the literature more domain-specific.

We looked into the performance of clustering methods based on various *Gene** parameterizations, as a first attempt to answer questions such as "which *Gene** representation provides an optimal similarity matrix for gene clustering?". Though this question needs further investigation, we found the *Gene** representation incorporating the 20 most relevant Medline abstracts, to have the best performance. Therefore the subsequent reported results are computed based on this representation, denoted as $Gene^*_{20}$. The similarity matrices derived from $Gene^*$ representations were used in a hierarchical and k-medoids clustering algorithm. In Fig. 3 we depict a typical result after application of hierarchical clustering.

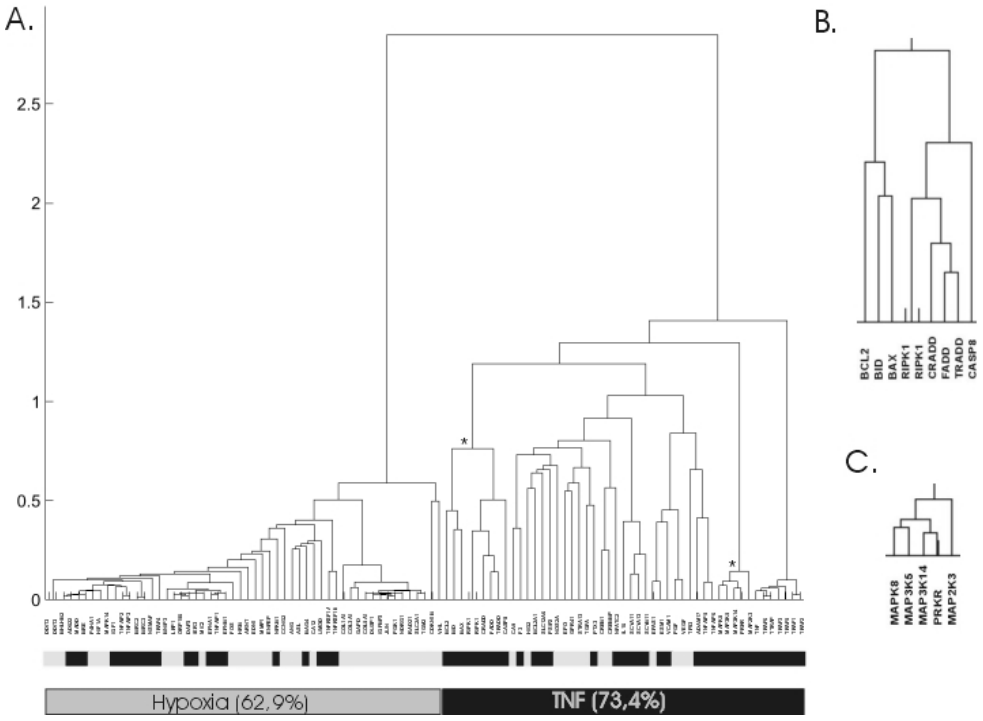

Figure 3: Hierarchical clustering of genes based on $Gene^*_{20}$.

To evaluate the biological usefulness of a *Gene**-based literature clustering, we compared the resulting clusters to the sets of genes for which the functional associations are well-established. The reason not to start immediately from expression-based gene clusters, is that real data-based clusters themselves can be unreliable, so the obtained results are hard to interpret. Consequently, in a first approximation we simulated the behaviour of an 'ideal' expression data cluster algorithm by having a molecular biologist construct two disjunct groups of genes; The first group ('hypoxia group') consisted of 54 genes known to be regulated by hypoxia (oxygen shortage) or to be involved in the hypoxic signal transduction; a second group ('TNF group') included 58 genes mediating tumor necrosis factor signaling. From a biological point of view, the two processes under consideration are not entirely independent, as both hypoxia and TNF are able to induce apoptosis (programmed cell death), even utilizing some common mechanisms and signaling molecules. However, this overlap can be taken into account in the

evaluation.

We merged these two groups and performed a Ward's linkage hierarchical clustering, as shown in Fig. 3. The distribution of the 'knowledge-based' clusters is plotted below the dendogram: Light boxes indicate genes belonging to the 'Hypoxia Group'; Dark boxes indicate genes belonging to the 'TNF Group'. This clustering divided the genes into two large clusters: the one to the left contained predominantly hypoxia related genes and the one to the right mainly TNF related genes. Within both clusters, highly relevant sub-clusters of functionally related genes could be identified. Two sub-clusters of the TNF group are shown in detail in Fig. 3B and C (marked with an asterisk in Fig. 3A). Fig. 3B represents genes associated with TNF-induced cell death, whereas Fig. 3C stands for MAP Kinases involved in alternative TNF signaling. Still, both clusters show a substantial degree of contamination which could be due to the above-mentioned overlap between the hypoxic and the TNF signaling pathways. Subsequent literature searches revealed that some of the contaminating genes were indeed linked to both hypoxia and TNF (e.g. NF-kappaB and p38 MAPK). Thus, this clustering method might provide new insights about associations between genes, as well as about their function or regulation.

Since the dendrogram in Fig. 3 suggests a division in two clusters, we executed a k-medoids clustering with $k = 2$ to check this partitioning. The left of Fig. 4 shows the comparison of literature clustering versus the test clusters.

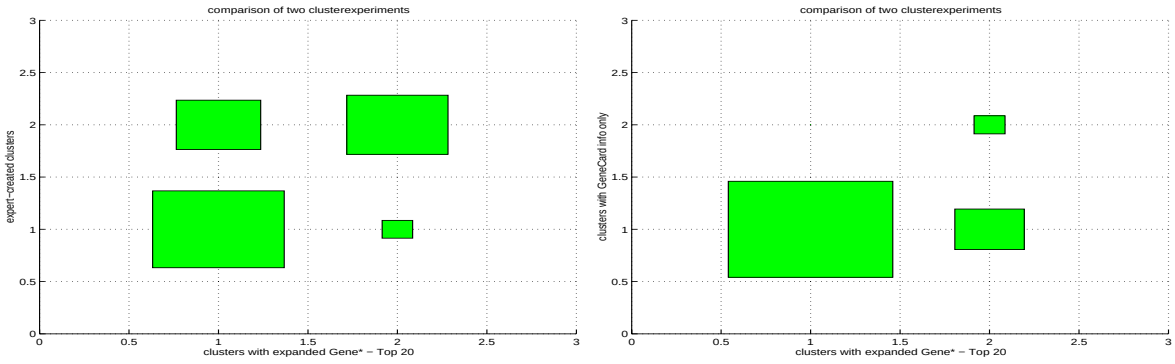

Figure 4: Comparison of clusterings. Left: manually created test clusters (Y-axis) versus clusters computed on $Gene^*_{20}$ representation (X-axis). Right: literature clusters computed on Genecard (Y-axis) versus $Gene^*_{20}$ with 20 Medline abstracts (X-axis).

In general, this type of graph matches two different divisions of gene sets by looking, for each cluster, where its members are in the dual set. In case of a perfect match, the surface of squares in the diagonal sums to 1. On the other hand, if some elements belonging to one cluster are distributed along many clusters in the dual division, their joint share in each of these is depicted by squares with normalized surface. In particular, we observe that only a small proportion of the genes known to be hypoxia-related (cluster 1 on Y axis) were assigned to 'another' cluster. Genes labeled to be mediated by TNF, however, were found to be more spread out over the two clusters. As explained before, this contamination is partly caused by the relatedness of the two processes.

Finally, we investigated whether the $Gene^*_5$ and $Gene^*_{20}$, being representations expanded with the most relevant Medline abstracts, are more informative than the gene kernel they are derived from (the GeneCard or $Gene^*_0$). The right part of Fig. 4 shows the difference between k-medoids applied to similarity matrices derived from $Gene^*_0$ and $Gene^*_{20}$, the $Gene^*$ representations using 0 and 20 Medline abstracts for expansion. We observe that the $Gene^*_{20}$-based method readily isolates genes from the large cluster generated by the strategy based solely on the GeneCard($Gene^*_0$) information. This finding confirms that, in our current approach, enriching the GeneCard kernel with other relevant information, increases the performance of the literature clustering.

In this section, we demonstrated the clustering of genes based on their literature representation using different parameterizations of $Gene^*$. We simulated on ideal test-case to rule out various uncertainties typically present in experimental data. The outcome suggests that literature-based clustering can help to interpret or re-parameterize clusterings from the gene expression domain.

6 Discussion and future work

In this article we summarized an integrated approach to use expression data and domain literature for gene clustering. The basis of our approach is *Gene**, a representation of relevant literature for each gene. This representation consists of numerical relevance factors for a set of domain-specific terms, which are computed on both a kernel document (currently the GeneCard) and the most corresponding Medline abstracts. The document collection on which the experiments were performed is a subset of the complete Medline abstract collection, obtained by selecting on journals according their relevance and impact factor. The *Gene** representation is appropriate for information retrieval and for constructing textual profiles for a gene cluster based on the representation of the cluster members. Furthermore it can be easily integrated within various statistical methods, hereby assessing it an equal status as the expression data. Therefore, this representation can be termed as 'genetic literature data'.

As a first step towards an integrated application of data and literature, we described a system capable to provide a textual profile for a gene cluster and to support the information retrieval in an integrated fashion. We demonstrated how its iterative and cyclic nature can benefit the analysis process. In a second step, we clustered the genes based solely on their *Gene** literature representation. We investigated this method for two manually constructed test sets and found promising results.

The outcome opens the way to a variety of research paths related to genetic literature data. Together with a team of biologists, we are working to further improve the profiling-based system and the dual clustering - both from a biological and a statistical viewpoint. A crucial part of this research is to evaluate the various *Gene** representations and to incorporate outcomes of supportive linguistic methods such as those handling the synonyms (aliases), domain specific phrases and stemming. Beside these, we are investigating how the *Gene** genetic literature data can be incorporated fully inside statistical methods, such as gene clustering or genetic network learning by probabilistic networks based on expression data.

Acknowledgements

Bart De Moor is Full Professor at the K.U.Leuven. Peter Antal and Patrick Glenisson are a Research Assistant with the K.U.Leuven. Tamas Meszaros is a Research Assistant with the B.U.T.E. Geert Fannes is a Research Assistant with the F.W.O. Vlaanderen. Yves Moreau is a Post-Doctoral Researcher with the F.W.O. Vlaanderen. Tom Boonefaes is a Biotech Fellow with the Ghent University, Pieter Rottiers is a Senior Scientist with the V.I.B. and Ghent University, Johan Grooten is Full Professor at the Ghent University and Head of the Molecular Immunology Unit. This work was carried out at the ESAT laboratory and supported by grants and projects from the Flemish Government: Concerted Research Action GOA-MEFISTO-666 (Mathematical Engineering for Information and Communication Systems Technology) and F.W.O. project G.0262.97: Learning and Optimization: an Interdisciplinary Approach and the F.W.O. Research Communities: ICCoS (Identification and Control of Complex Systems) and Advanced Numerical Methods for Mathematical Modelling and Bilaterale Wetenschappelijke en Technologische Samenwerking Flanders-Hungary, BIL2000/19; from National Fund for Scientific Research (OTKA) under contract number T030586; from National Fund for Scientific Research (OTKA) under contract number F-030763; from the IDO/99/03 project (K.U.Leuven) "Predictive computer models for medical classification problems using patient data and expert knowledge", of the FWO grants G.0326.98 and G.0360.98, the FWO project G.0200.00. and Cult. Affairs-Interuniversity Poles of Attraction Programme (IUAP P4-02 (1997-2001): Modeling, Identification and Control of Complex Systems. We thank Isabel Van Dooren, Janick Mathys and Kathleen Marchal for their help.

References

[1] R. Baeza-Yates and B. Ribiero-Neto, *Modern information retrieval*, Addison-Wesley, 1999.

[2] The Gene Ontology Consortium, *Gene ontology: tool for the unification of biology*, Nature Genetics (2000), 25–29.

[3] D.R. Masys et al., *Use of keyword hierarchies to interpret gene expression patterns*, Bioinformatics **17** (2001), 319–326.

[4] L. Tanabe et al., *Medminer: an internet text-mining tool for biomedical information, with application to gene expression profiling*, Biotechniques **27** (1999), 1210–1214, 1216–1217.

[5] T.K. Jenssen, L.M.J. Oberg, M.L. Andersson, and J. Komorowski, *Methods for large-scale mining of networks of human genes*, Proceedings SIAM Conference on Data Mining (SDM 2001), Chicago, IL, USA, SIAM, 2001.

[6] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet, *Genecards: A novel functional genomics compendium with automated data mining and query reformulation support*, Bioinformatics **14** (1998), 656–664.

[7] T.C. Rindflesch, J.V. Rajan, and L.Hunter, *Extracting molecular binding terms from biomedical text*, Proceedings of the 6th Applied Natural Language Processing Conference, 2000, pp. 188–195.

[8] R. Shamir and R. Sharan, *Algorithmic approaches to clustering gene expression data*, Current Topics in Computational Biology (T.Jiang, T. Smith, Y. Xu, and M. Q. Zhang, eds.), MIT Press, 2001.

[9] H. Shatkay, S. Edwards, W.J. Wilbur, and M. Boguski, *Genes, themes and microarrays. using information retrieval for large-scale gene analysis*, Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, Menlo Park, CA, USA, AAAI, 2000, pp. 317–328.

[10] B.J. Stapley and G. Benoit, *Biobibliometrics: Information retrieval and visualization from co-occurences of gene names in medline abstracts*, Pacific Symposium on Biocomputing (2000), 529–540.