EVALUATION OF THE VECTOR SPACE REPRESENTATION IN TEXT-BASED GENE CLUSTERING

P. GLENISSON, P. ANTAL, J. MATHYS, Y. MOREAU, B. DE MOOR
Department of Electrical Engineering, ESAT-SISTA,
Kasteelpark Arenberg 10,
B-3001 Leuven, Belgium

Thanks to its increasing availability, electronic literature can now be a major source of information when developing complex statistical models where data is scarce or contains much noise. This raises the question of how to deeply integrate information from domain literature with experimental data. Evaluating what kind of statistical text representations can integrate literature knowledge in clustering still remains an unsufficiently explored topic. In this work we discuss how the bag-of-words representation can be used successfully to represent genetic annotation and free-text information coming from different databases. We demonstrate the effect of various weighting schemes and information sources in a functional clustering setup. As a quantitative evaluation, we contrast for different parameter settings the functional groupings obtained from text with those obtained from expert assessments and link each of the results to a biological discussion.

1 Introduction

More and more, a successful understanding of complex genetic mechanisms (such as regulation, functional understanding,...) critically depends on the interaction between statistical analysis and various knowledge sources, such as annotations databases, specialized literature, and curated cross-links between them (Baxevanis[1]). Despite these efforts, the current interaction between the experimental (data) analysis and text-based information requires extensive user intervention. Gene expression experiments, which measure large-scale genetic activity under a variety of biological conditions, are excellent examples of environments that rely strongly on this interaction. Indeed as (1) the cost of data collection is high, (2) measurements are often noisy or unreliable, and (3) established relationships in the transcriptome are fragmentary at best, a deeper integration between data and text-based information will benefit the knowledge discovery process.

The present strategies for knowledge-based expression data analysis rely on the premise that statistical data analysis and biological knowledge can complement each other by *linking* two independently constructed sources that contain conceptually related records (Masys[2] and Vidal[3]).

In yeast for example, interpreting cluster patterns involves the consultation of curated functional databases such as the Saccharomyces Genome

Database[a] (SGD), which offers concise functional annotations and a variety of cross-references to other repositories. For more elaborate information, researchers can resort to MEDLINE, an online bibliographic source of citations and abstracts in biomedical research dating from 1966 till present. While the use of a controlled and curated index, like MeSH[b], is already common in automatically associating gene functions (see for example Jenssen[4], Masys[5], Kankar[6]), we tested additionally the use of free-text as a potentially more informative, and in the future possibly more dominant, information source (see also Stapley[7], Stephens[8], Renner[9], Iliopoulos[10], Raychaudhuri[11]).

In this work, we explore how representations borrowed from the field of information retrieval can be adopted for clustering genes based on their associated literature. We encode text-based information from various sources in a typical bag-of-words representation following the vector space model, a work horse in information retrieval research. We investigate the effect of *pooling* and *expanding* these sources, together with the question of which type of representation is more appropriate. To evaluate the biological usefulness of literature clustering, we formulate a clustering problem with gene sets from *Saccharomyces cerevisiae* for which the functional associations are well-established and biologically distinct. The reason not to start immediately from expression-based gene clusters is that these data-based clusters are often biologically complex and cannot provide a gold standard to interpret and quantify the correspondence between various data mining methods. Additionally, we seek to identify some inherent biases of the vector-space model by testing and quantifying its performance on a fairly simple biological problem. To compare different versions of the representation with respect to clustering performance, we use both external and internal scores for cluster validation (see Section 2). The aim of these evaluations is to establish a powerful statistical text representation as a foundation for knowledge-based gene expression clustering.

2 Methods

2.1 Compilation of Information Sources

We collect and compile (as of September 2001) several sources for textual annotations of the genes. Firstly we retrieve the gene descriptions from the Saccharomyces Genome Database (SGD)[c]. Secondly, we use SWISS-PROT (SP)[d], a curated protein sequence database. We pool the SGD and SP information

[a] http://genome-www.stanford.edu/Saccharomyces/
[b] http://www.nlm.nih.gov/mesh/meshhome.html
[c] http://genome-www.stanford.edu/Saccharomyces/
[d] http://www.expasy.org/sprot/

into a local database we denote by YeastCard (YC). It serves as an extended textual resource for yeast genes. Finally, as a source for more detailed information, we use a collection of 493,923 yeast-related MEDLINE abstracts dated between January 1982 and November 2000. They were selected by retaining those abstracts coming from a list of 59 journals that was composed according to both impact factor[e] and relevance. The aim of this trimming is to retain a more domain-specific subset of abstracts, which is still diverse enough to hold essential genetic information. We evaluate how these sources influence text-based gene clustering and, more specifically, we investigate how the expansion of the SGD and YeastCard annotations with MEDLINE abstract information (see Section 2.3) affects clustering performance.

2.2 Text Representation

The representation called the vector space model encodes a document in a k-dimensional space where each component v_{ij} represents the weight of term t_j in document d_i. The grammatical structure of the text is neglected and therefore it is also referred to as a bag-of-words representation. As a basic index for each document in the collection, we construct a vocabulary consisting of 26,420 (possibly multi-word) terms extracted from the Gene Ontology[f] *Term* field. The Porter stemmer is used to canonize the words.Based on the *Term* field in GO and *Synonym* field in SWISS-PROT, we process candidate phrases and replace known synonyms. In this work we used the following common used indexing schemes (Baeza-Yates[12] and Korfhage[13]):

- $v_{ij}^{\text{bool}} = 1$ if $t_j \in d_i$, 0 otherwise
- $v_{ij}^{\text{freq}} = \frac{f_{ij}}{\max_{\forall j}(f_{ij})}$, where f_{ij} is the number of occurrences of t_j in d_i
- $v_{ij}^{\text{tf.idf}} = f_{ij} \log(\frac{N}{n_i})$, where N is the total number of documents and n_i is the number of documents containing term i in the collection

Additionally, we define another type of index called the reference representation (see Shatkay[14]). When a document contains references to other documents in the same or another repository, we can encode this as follows:

- $v_{ij}^{\text{ref}} = 1$ if annotation i contains a reference to document j, 0 otherwise

[e] http://jcrweb.com/
[f] http://www.geneontology.org

2.3 Relevance and Similarity

We express similarity between pairs of documents d_i and d_j, or between a text document d_i and a query document d_j, by the cosine of the angle between the corresponding normalized vector representations:

$$\text{sim}(d_i, d_j) = \cos(d_i, d_j).$$

The underlying hypothesis states that high similarity equals strong relevance. Further, the method termed *pseudo-relevance feedback* is geared towards *expanding* a query document with the n most similar documents in a collection and aims at refining the search or clustering process by a recalculation of the term weights (Yates[12]). We denote the annotations A expanded with n documents from collection C by A-C_n. A related application of pseudo-relevance feedback in combination with the reference representation can be found in Shatkay[14].

2.4 Cluster Algorithm

As divisive clustering algorithm we used the K-medoids algorithm (Rousseeuw[15]), which minimizes the objective function

$$\sum_{k=1}^{K} \sum_{j \in C_k} d(x_j, m_k)$$

over multiple partitionings $C = \{C_1, ..., C_K\}$ with $\{m_1, ..., m_K\}$ the corresponding representative points (called medoids) of each cluster. The parameter K denotes the number of clusters and is fixed in advance. One advantage of this algorithm over centroid-based methods, such as K-means, is that each medoid constitutes a robust representative data point for each cluster.

2.5 Cluster Quality

To measure the performance and quality of the clustering we define three scores: the silhouette coefficient, the performance of the clustering as a classifier, and the Rand index. The first two are termed *internal* scores since they rely on statistical properties of the clustered data, the last one is called *external* because it involves a comparison with a known, external labeling.

Silhouette Coefficient

As a first internal score for cluster quality we use the silhouette coefficient per cluster $S_k = \frac{1}{n_k} \sum_{i=1}^{n_k} s_{ik}$ and the overall silhouette coefficient $S = \frac{1}{n} \sum_k \sum_{i=1}^{n_k} s_{ik}$

with n_k is the size of cluster k, n the number of objects, and

$$s_{ik} = \frac{b(i) - a(i)}{\max(a(i), b(i))},$$

where $a(i)$ is the average dissimilarity of member i to all other members of its cluster and $b(i)$ the dissimilarity of member i to the nearest member of the nearest cluster. It is a metric-independent measure designed to describe the ratio between cluster coherence and separation and to assist in choosing which clustering is preferable according to the data (Rousseeuw[15]).

k-NN Learnability

For the second measure of internal cluster quality, we look upon the problem as being semi-supervised. Using the clustering result as a labeling for all the points, we assess the performance of a given classifier on a class (or cluster) in a cross-validated leave-one-out setup. Following Pavlidis[16], we use a k-NN classifier jointly with the $(1 - \cos(\cdot, \cdot))$ distance measure to compute a misclassification score for each class. The statistical significance of this score m, is expressed by a p-value derived from a binomial $\mathcal{B}(m, p_{\text{misclass}})$ with p_{misclass} the prior chance of misclassification, which can be computed analytically in case of a k-NN classifier (details can be found in Pavlidis[16]).

Rand Index

As an external measure for cluster validity we use the adjusted Rand index[17]. Given a set of n points, an external partition $P = \{P_1, ..., P_k\}$, and a clustering $C = \{C_1, ..., C_l\}$, define a as the number of pairs of points that co-occur in a group in the partitioning P as well as in the clustering C, d the number of pairs of points that are in different groups in P as well as in C, and b and c as the number of pairs of points that co-occur in a group in P, but not in C or vice-versa. The Rand index is then defined by

$$R = \frac{a + d}{a + b + c + d}.$$

The correction for random partitioning is $R_{\text{adj}} = \frac{R - E(R)}{\max(R) - E(R)}$, where a hypergeometric baseline distribution is used to compute the expected values. In a comparative study[17], the adjusted Rand index is recommended as the external measure of choice.

3 Results

3.1 Construction of Test Set

We construct a set of genes for which the functional associations are well-established. From the MIPS catalogue [1], we select three biologically distinct functional groups consisting of 116 genes in total. For all genes we select their corresponding SGD and YC annotations (see Section 2.1) and proceed with the 105 genes that have entries in both databases[9]. The first group holds 63 genes that encode lysosomal proteins. The second group consists of 30 genes involved in translational control and the third contains 23 genes related to amino acid transport.

3.2 Cluster Performance

Following the strategies outlined in Section 2, all gene annotations are represented by various indices and subsequently *expanded* with the 20 best matching MEDLINE abstracts. More specifically, we perform the expansion by re-indexing the enriched annotations, again following various indexing schemes. Table 1 summarizes the impact of these settings on cluster performance, expressed by means of the Rand index R_{adj}. Firstly we discuss the effect of information source, afterwards follows the results on the indexing schemes.

Table 1: R_{adj} scores for clustering the three groups using various representations. Note that some results are duplicated along the blocks to facilitate discussion.

	Representation	Weight	R_{adj}	$\sharp(t_i)$
Source	SP Keywords	$bool$	0.1767	3
	SGD	tf	0.4050	
	YeastCard	$tf \cdot idf$	0.4617	
Index	SGD	$bool$	0.3386	8
	SGD	tf	0.4050	
	YeastCard	$bool$	0.3323	
	YeastCard	tf	0.4028	
	YeastCard	$tf \cdot idf$	0.4617	26
	$YC\text{-}ML_{20}$	$bool$	0.3726	
	$YC\text{-}ML_{20}$	tf	0.2953	
	$YC\text{-}ML_{20}$	$tf \cdot idf$	0.7344	396
	$YC\text{-}ML_{20}$	ref	0.2354	20
Expansion	$SGD\text{-}ML_{20}$	$tf \cdot idf$	0.5920	
	$YC\text{-}ML_{20}$	$tf \cdot idf$	0.7344	

Effect of Indexing Scheme In the second block of Table 1 we write the performance of the boolean ($bool$), frequency (tf) and $tf \cdot idf$ index on typical

[9] ftp.esat.kuleuven.ac.be/sista/glenisson/reports/webSuppl_TR02_121/yeastcardTable.htm

free-text entries in annotation databases, and on a set of our top 20 retrieved MEDLINE abstracts. For very brief keyword-based descriptions (less than 8 words) the boolean representation is found to be the best one. If all fields from SGD are used, tf (0.4050) improves on *bool* (0.338). For the YC database, typically containing 30 to 50 terms per entry, $tf \cdot idf$ (0.4617) outperforms *bool* (0.3323) and tf (0.4028) slightly.

In the expansion step, we collect and re-index the 20 best matching MEDLINE abstracts for each gene. This operation provides a profile for each gene with the number of terms ranging typically between 200 and 400. Among the indexing options for this set of abstracts, $tf \cdot idf$ (0.7344) scores considerably higher than *bool* (0.3726) and tf (0.2953), even after stopword removal.

Basing ourselves on the same 20 top-scoring abstracts we also evaluate the performance of the reference representation ref, which characterizes a gene in document space instead of term space. It has an R_{adj} value of 0.2354, indicating that it is a less descriptive representation. This can be explained by the fact that ref is probably more dependent on the retrieval of highly relevant abstracts (see also Shatkay[14]).

Effect of Information Source In the first block of Table 1, we see that for the gene groups considered, the *keywords* field in SWISS-PROT does not provide sufficient information for an acceptable clustering result (0.1767). For instance, the SWISS-PROT keyword list only provides an average of 2 to 3 meaningful keywords for 86 out of 105 genes. The remaining genes are described with no or irrelevant keywords such as *hypothetical protein*, which will not allow for correct classification. Using the GO entries and especially the description line of SGD improves the results and raises the Rand score to 0.4050. Only two genes have no meaningful representation, YKL002w and YLR309c, whereas the others are now described by 7 to 8 biologically relevant terms. When resorting to our pooled information source YC (see Section 2.1), we obtain a score of 0.4617, misclassifying 21 out of 105 genes. Although the clustering itself is not dramatically influenced by the expansion with YC, for most of the genes, the textual representation is greatly improved (e.g., the weights of specific terms are increased and additional specific terms are incorporated). For instance, Table 2 shows the text profiles of the medoids of the vacuolar cluster for various representations.

In the clustering based on SWISS-PROT keywords, the vacuolar cluster itself is not found. Instead, the algorithm identifies a cluster of ATP-binding proteins that contains the vacuolar ATPases but also a number of ATP-binding proteins involved in translational control. The SGD representation ensures the grouping of vacuolar proteins solely based on one relevant term, *vacuolar*. Both

Table 2: Text profiles of the medoids for group1 (only 25 top-scoring terms are shown).

SP keywords	SGD	YC	YC-ML_{20}
ATP (0.45)	vacuolar (0.38)	vacuolar (0.54)	vacuolar (0.54)
ATP bind (0.45)	vps41 (0.38)	ATPas (0.4)	vacuol (0.45)
bind (0.45)		vacuolar membran (0.32)	snare (0.36)
		vma13 (0.21)	vacuolar membran (0.18)
		subunit (0.2)	T snare (0.17)
		associ (0.17)	syntaxin (0.16)
		organel (0.16)	vacuolar assembli (0.12)
		vacuolar acidif (0.16)	Golgi (0.1)
		acidif (0.15)	carboxypeptidas (0.1)
		sector (0.14)	vam3 (0.09)
		hydrogen (0.13)	pep12 (0.09)
		membran (0.1)	V snare (0.08)

Table 3: Text profiles of gene YPL029w based on the SGD and YeastCard representations.

SGD	YC
ATP (0.27)	helicas (0.57)
ATP depend helicas (0.27)	mitochondri(0.36)
depend (0.27)	ATP depend helicas(0.29)
helicas (0.53)	suv3 (0.29)
RNA (0.27)	ATP (0.23)
RNA helicas (0.27)	depend (0.2)
suv3 (0.27)	RNA (0.2)
	RNA helicas (0.19)
	post (0.19)
	ATP depend RNA helicas (0.18)
	elem (0.16)
	translat (0.13)
	control (0.13)
	interact (0.11)
	transcript (0.09)

Table 4: Text profiles of gene YLL048c and YPL149w based on the YeastCard representation and the corresponding expansion to MEDLINE (only the top-scoring terms are shown).

YLL048c YC	YC-ML_{20}	YPL149w YC	YC-ML_{20}
bile (0.68)	bile (0.92)	autophagi (0.89)	autophagi (0.87)
transport (0.46)	bile acid transport (0.28)	apg5 (0.43)	apg5 (0.17)
bile acid transport (0.25)	bile acid (0.22)		conjug (0.15)
ybt1 (0.25)	hepatocyt (0.06)		apg1 (0.13)
ATP (0.20)	transport (0.06)		cAMP (0.13)
abc (0.15)	abc (0.05)		starvat (0.11)
ATP bind (0.14)	ATP (0.05)		kinas (0.11)
integr membran (0.14)	ATPas (0.03)		phosphati-dylinositol(0.08)
integr (0.13)	apic (0.03)		vacuol (0.08)
membran (0.11)	vesicl (0.03)		apoptosis (0.08)
acid (0.1)	cotransport (0.03)		hepatocyt (0.07)
similar (0.1)	sister (0.03)		antagonist (0.06)
depend (0.09)	voltag (0.03)		ubiquitin (0.06)
bind (0.07)	glycoprotein (0.02)		apg12 (0.06)
famili (0.03)	triphosph (0.02)		amino-peptidas (0.04)

the YC representation and the MEDLINE expansion of the YC annotation result in a large cluster containing most of the vacuolar proteins. The text profiles of the corresponding medoids confirm the success of the MEDLINE expansion and the feasibility of our approach to identify relevant terms that characterize individual genes or groups of genes. For the other two groups a similar improvement is observed.

Table 3 shows two examples of text profiles of individual genes that were misclassified when the SGD representation was used whereas the YC representation assigned the genes to the correct cluster. For the RNA helicase, YPL029w, terms like *mitochondri* and *translat* are added to the text profile. The clustering of YBR024w in the group of translation-related proteins is based on terms such as *mitochondri*, *inner*, and *membrane*.

Expansion to MEDLINE improves the text profiles of almost all of the genes and even the clustering of a few genes such as YLL048c, a lysosomal bile transporter, and the genes that encode autophagy-related proteins. In the clustering based on the YC representation, YLL048c was wrongly assigned to the group of amino acid transporters. However, the expansion strongly decreased the weight of the term *transport* and introduced the term *ATPase* in the text profile, resulting in a correct classification of the gene. For the autophagy-related genes, retrieval of the term *vacuol* ensures correct grouping after MEDLINE expansion as shown in Table 4. However, some of the genes are incorrectly clustered no matter what representation or weighting scheme was used. For instance, Group 1 and Group 3 include several proteins that regulate transcription, a process that is closely related to translation and shares many of its keywords. The proteins YLR025w (Group 1), YLR375w (Group 3) and YDL048c (Group 3) are therefore persistently misclassified into Group 2. One gene, YLR309c (Group 1), is consistently assigned to the wrong cluster because it lacks proper annotation. The only terms that characterize YLR309c are vague, aspecific words such as *gene product* and the name of the gene *imh1*. This information is insufficient for successful expansion with MEDLINE. A manual search via the PUBMED engine did not reveal much information on *imh1* (YLR309c) either.

3.3 Cluster Quality

Because of the absence of a gold standard or prior knowledge in regular clustering problems, internal measures of quality are used to evaluate a cluster result (Jain[17]). They are based on various statistical properties of the grouped data and provide clues to choose between different parameterizations of a single algorithm (such as the number of clusters) or even between various clus-

ter algorithms. Here we use two measures, the silhouette coefficient and a k Nearest Neighbour (k-NN) learnability index to study the influence of the text representation on a standard clustering procedure. Our concern is that, although high Rand scores may be encouraging, they do not provide any level of confidence in the result: it might be that clusters or groups lie very close to each other or that clusters exhibit a high spread. Therefore, we compute for each clustering their score over the major text representations. From Table 5 we see that the silhouette score does not contain any indications towards the optimal representation (i.e., the one with the highest R_{adj}). The more local 10-NN misclassification rate performs better, indicating that individual cluster structures should be examined more carefully. We expect that groups that are easy (and therefore do not need an elaborate representation to be learned) will end up in clusters having low misclassification rates over all the representations. Harder groups will behave inversely.

In Figure 1 we plot the misclassification rate against the silhouette coefficient to look for possible discrepancies between the two scores in this problem. We show the results for 10-NN. From Figure 1 (left) we estimate the group of translational control as the hardest to learn from text, since it has the highest misclassification rate, even with the MEDLINE expanded representation. Additionally, there exists a discrepancy between the silhouette and learnability score. For the amino acid group there exists great variation in the silhouette value, while the misclassification rate stays below 0.1. This indicates that the shape and constitution of the cluster changes over the representation without changing its relative position with respect to the other clusters.

The quality of the cluster is highly affected by the presence of distant genes: genes that have a poor or biased description (and hence representation) will end up far away from the cluster center (i.e., the medoid). We illustrate this in Figure 1 (right), where we plot the growing of the silhouette score (from right to left on the x-axis) while increasingly dropping members beyond a given distance. Flat regions indicate the absence of members in that distance region and sudden changes in silhouette scores show the detrimental effect of those, more distant, genes on the scores. Since a biologist is not always interested in clustering *all* the genes *per se*, this information can be utilized to prune genes from the clustering process or to check the information given by that gene.

Table 5: Various cluster quality scores for the three major text representations

Representation	Silhouette	10-NN p-value (miscl. rate)	R_{adj}
SGD	0.220	4.634^{-4}(0.2286)	0.4050
YC	0.1576	1.2^{-3}(0.2095)	0.4617
YC-ML_{20}	0.2192	10^{-9}(0.1143)	0.7344

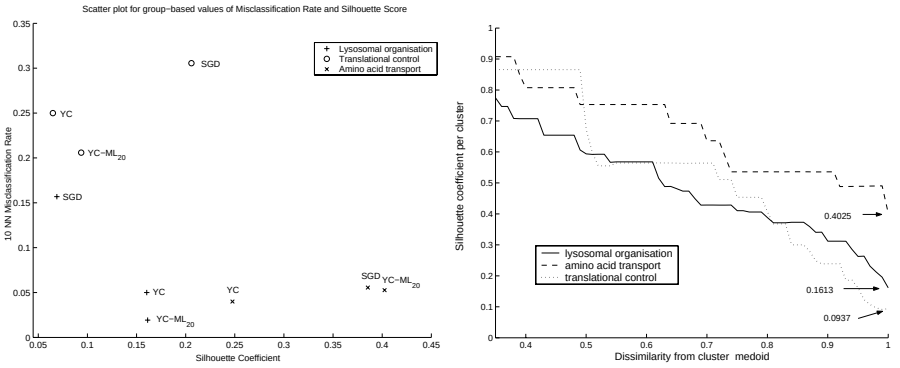

Figure 1: Correspondence between spatial cluster information as captured by the silhouette coefficient and learnability in Nearest Neighbour sense (left) and effect of distant members in each cluster on its silhouette score for (right).

4 Conclusion

Our aim was to investigate the potential of the vector-based representation for functional and text-based gene clustering.We looked into which bag-of-words representation was optimal for what type of information source. We expanded various gene annotations with abstracts that were closest for the cosine measure. Since similarity ranking scores are often hard to threshold and provide a poor quantification for relevance, we retained the top 20 matching entries. This approach considerably improved clustering results because of the inclusion of important terms not present in the annotation databases or because of a relative weight change of already included terms.

Next to a biological evaluation, we computed two complementary internal cluster quality measures to examine some statistical properties of the text representations. The k-NN learnability score gave useful clues on how difficult a class or cluster was to learn. The outcome matched our biological expectations, indicating that our recommended representation is usable in an unsupervised learning task. The silhouette profiles gave more insight into the nature of the clustered annotations and were used to prune or check the information of genes distant from a cluster's medoid.

Finally, the ultimate goal of our approach is to use key elements of the shallow-statistical approach as extra background information in the clustering of expression data.

Acknowledgments

Patrick Glenisson and Peter Antal are research assistants with the KUL. Janick Mathys is a post-doctoral researcher with the KUL. Yves Moreau is a post-doctoral researcher with the FWO Vlaanderen. Dr. Bart De Moor is a full professor at the KUL, Belgium. Research supported by Research Council KUL: GOA-Mefisto 666, IDO (IOTA Oncology, Genetic networks), several PhD/postdoc/fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects G.0115.01 (microarrays/oncology), G.0413.03 (inference in bioi), G.0388.03 (microarrays for clinical use), G.0229.03 (ontologies in bioi), research communities (ICCoS, ANMMM); AWI: Bil. Int. Collaboration Hungary/ Poland; IWT: PhD Grants, STWW-Genprom (gene promotor prediction), GBOU-McKnow (Knowledge management algorithms), GBOU-SQUAD (quorum sensing), GBOU-ANA (biosensors); Belgian Federal Government: DWTC (IUAP IV-02 (1996-2001) and IUAP V-22 (2002-2006); EU: CAGE; ERNSI;

References

1. A.D. Baxevanis. The molecular biology database collection: 2002 update. *Nucleic Acids Research*, 30:1–12, 2002.
2. D.R. Masys. Linking microarray data to the literature. *Nature Genetics*, 28:9–10, 2001.
3. M. Vidal. A biological atlas of functional maps. *Cell*, 104:333–339, 2001.
4. T.K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28:21–28, 2001.
5. D.R. Masys, J.B. Welsh, J.L. Fink, M. Gribskov, I. Klacansky, and J. Corbeil. Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, 17:319–326, 2001.
6. P. Kankar, S. Adak, A. Sarkar, K. Murari, and G. Sharma. Medmesh summarizer: text mining for gene clusters. In *Proceedings of the Second SIAM International Conference on Data Mining*, 2002.
7. B.J. Stapley and G. Benoit. Biobibliometrics: Information retrieval and visualization from co-occurences of gene names in medline abstracts. In *Proceedings of the Fifth Annual Pacific Symposium on Biocomputing (PSB 2000)*, 2000.
8. M. Stephens, M. Palakal, S. Mukhopadhyay, R. Raje, and J. Mostafa. Detecting gene relations from medline abstracts. In *Proceedings of the Sixth Annual Pacific Symposium on Biocomputing (PSB 2001)*, 2001.
9. A. Renner and A. Aszodi. High-throughput functional annotation of novel gene products using document clustering. In *Proceedings of the Sixth Annual Pacific Symposium on Biocomputing (PSB 2000)*, 2000.
10. I. Iliopoulos, A.J. Enright, and C.A. Ouzounis. Textquest: document clustering of medline abstracts for concept discovery in molecular biology. In *Proceedings of the Sixth Annual Pacific Symposium on Biocomputing (PSB 2001)*, 2001.
11. S. Raychaudhuri, J.T. Chang, P.D. Sutphin, and R.B. Altman. Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Research*, 12:203–214, 2002.
12. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, 1999.
13. R. Korfhage. *Information Storage and Retrieval*. New York: Wiley Computer Pub., 1997.
14. H. Shatkay, S. Edwards, W.J. Wilbur, and M. Boguski. Genes, themes and microarrays: Using information retrieval for large-scale gene analysis. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, Menlo Park, CA, USA*, pages 317–328. AAAI, 2000.
15. L. Kaufman and P. Rousseeuw. *Finding groups in data*. Wiley-Interscience, 1990.
16. P. Pavlidis, D.P. Lewis, and W.S. Noble. Exploring gene expression data with class scores. In *Proceedings of the Seventh Annual Pacific Symposium on Biocomputing (PSB 2002)*, 2002.
17. A. Jain and R. Dubes. *Algorithms for clustering data*. Prentice Hall, 1988.