BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 9 2005, pages 2049–2058
doi:10.1093/bioinformatics/bti268

Data and text mining

Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes

R. Jelier[1,*] G. Jenster[2], L. C. J. Dorssers[3], C. C. van der Eijk[1], E. M. van Mulligen[1], B. Mons[1] and J. A. Kors[1]

[1]Department of Medical Informatics, [2]Department of Urology and [3]Department of Pathology, Erasmus MC—University Medical Center, Rotterdam, The Netherlands

Received on July 15, 2004; revised on December 22, 2004; accepted on January 6, 2005
Advance Access publication January 18, 2005

ABSTRACT

Motivation: The advent of high-throughput experiments in molecular biology creates a need for methods to efficiently extract and use information for large numbers of genes. Recently, the associative concept space (ACS) has been developed for the representation of information extracted from biomedical literature. The ACS is a Euclidean space in which thesaurus concepts are positioned and the distances between concepts indicates their relatedness. The ACS uses co-occurrence of concepts as a source of information. In this paper we evaluate how well the system can retrieve functionally related genes and we compare its performance with a simple gene co-occurrence method.

Results: To assess the performance of the ACS we composed a test set of five groups of functionally related genes. With the ACS good scores were obtained for four of the five groups. When compared to the gene co-occurrence method, the ACS is capable of revealing more functional biological relations and can achieve results with less literature available per gene. Hierarchical clustering was performed on the ACS output, as a potential aid to users, and was found to provide useful clusters. Our results suggest that the algorithm can be of value for researchers studying large numbers of genes.

Availability: The ACS program is available upon request from the authors.

Contact: r.jelier@erasmusmc.nl

INTRODUCTION

The availability of whole genome sequences and the advent of high-throughput technology for molecular biology have dramatically changed the nature of biomedical research. Thousands of genes or proteins can now be studied in a single experiment. With this development arose the challenge to efficiently handle the huge amounts of data produced by these experiments. An important issue in the interpretation of data produced by DNA microarrays is the identification of the biological processes that underlie the observed differences in gene expression. Information needed for this task is for the larger part available in millions of free-text scientific publications, with thousands of new publications being added every day. When many genes are studied, the number of relevant publications will frequently be prohibitively large. This renders the traditional approach of manually searching bibliographic databases for every gene and reading scientific articles inadequate. It is therefore an important challenge at this time to make the available information both accessible and interpretable for molecular biologists.

An interesting current development is the use of annotations of genes with gene ontology (GO) terms (Ashburner *et al.*, 2000; Camon *et al.*, 2004) for the analysis of the results of microarray experiments (Zhang *et al.*, 2004; Al Shahrour *et al.*, 2004). The most reliable annotations are based on manually assigning GO codes to genes based on scientific literature. GO provides a structured description of biological information which is very amenable for use in bioinformatics. These methods are useful, though limited in flexibility by the focus of the ontology. GO annotations are for instance not very useful if one is interested in gene–disease relations. Additionally, the most reliable annotations are obtained by a difficult, slow and labor-intensive manual process. Clearly there is much more information stored in the whole body of literature than captured in current GO annotations. Therefore mining texts directly for relevant information on genes would be more flexible and could make an important addition to the molecular biologist's toolbox for microarray data analysis.

The recent years have seen new methods to efficiently use the large amount of literature for biomedical research. In an early effort, Masys *et al.* (2001) made keyword profiles for genes based on the manual annotations of articles with the controlled vocabulary Medical Subject Headings (MeSH) in the National Library of Medicine's MEDLINE database. For a group of selected genes, these profiles are combined and every keyword is given a value indicating its specificity for the group. An important developing field is the automatic extraction of relevant information from scientific texts [for a review, see Shatkay and Feldman (2003)]. The most important distinctions between current text-mining methods are the amount of linguistic information that is used and the number of documents that can be handled efficiently. One approach is to extract detailed information from documents by using natural language-processing techniques (Friedman *et al.*, 2001; Pustejovsky *et al.*, 2002). Many approaches though extract information about genes from scientific texts using only information about the co-occurrence of terms in a sentence or abstract (Chaussabel and Sher, 2002; Becker *et al.*, 2003; Tanabe *et al.*, 1999; Stapley and Benoit, 2000; Jenssen *et al.*, 2001;

*To whom correspondence should be addressed at Department of Medical Informatics, Erasmus MC—University Medical Center Rotterdam, PO Box 1738, 3000 DR Rotterdam, The Netherlands.

Wren *et al*., 2004). The use of simple co-occurrence is popular, because it allows for easy implementation and the efficient processing of huge amounts of texts. Also, the co-occurrence of gene names in an abstract frequently reflects an actual biological relationship between the two genes, as was shown by Jenssen *et al*. (2001) and Stapley and Benoit (2000).

Recently, we developed a new co-occurrence based text meta-analysis tool, the associative concept space (ACS) (Van Der Eijk *et al*., 2004). To construct the ACS, thesaurus concepts are automatically identified in texts. The use of a thesaurus allows synonyms to be mapped to the same concept, which reduces noise caused by natural language variation. Additional advantages are the possibilities of including multiword terms and using thesaurus hierarchies. The thesaurus we use contains genes but also many other biomedical concepts.

The ACS algorithm is a Hebbian-type of learning algorithm that in an iterative process positions the thesaurus concepts in a multi-dimensional Euclidean space. In this space the dimensions do not take a specific meaning, but just allow the positioning of the concepts relative to each other. The position of a concept follows from the mapping of co-occurrence relations (paths) between concepts to distances. A distance between two concepts will not only reflect the co-occurrence of the two concepts, a one-step relation, but also indirect, multi-step relations between the two concepts. The idea behind the algorithm is that concepts that are placed close to each other will be more likely to share an actual semantic relationship. An important feature of the ACS is that the multidimensional space can be visualized using standard dimension reduction techniques. The visualized ACS allows for easy and intuitively appealing browsing for relations between concepts that are derived from the underlying literature. The ACS can thus be used as a kind of portal to the literature, but it can also be used as a knowledge discovery tool. When in the ACS two concepts are placed close to each other while they do not have a co-occurrence, this would suggest that a relationship is not explicit in the literature set, but is likely to exist.

The ACS can be used in a similar way to how other authors have used co-occurrence as a basis for a knowledge discovery system. Swanson and Smalheiser (1997) discovered valuable knowledge hidden in medical literature. They searched for paths between two sets of related terms allowing for one intermediary term to connect terms from the two sets. Several other authors have built on their work using similar models (Srinivasan, 2004; Weeber *et al*., 2003b; Wren *et al*., 2004).

Compared to previously published algorithms, the ACS has the potential to stand out on several points. The ACS could improve on the performance of only direct co-occurrence of genes by improving recall. When only direct gene-gene co-occurrences are used, some relations will be missed, for example the relation between two genes that are involved in the same cellular process would be missed when their roles happen to be described in separate papers. The ACS can reveal relations between genes based on their contexts, i.e. the other concepts with which they are mentioned, and does not require the genes to be mentioned in the same article. The method introduced by Chaussabel and Sher (2002) also uses other co-occurring terms, and can pick up relations between concepts that do not necessarily co-occur in the same article. Our approach differs in that we use a thesaurus for identifying concepts in texts, which, as mentioned earlier, has several advantages. Additionally the ACS differs as it implicitly uses more information in that concept relations that involve more than two steps play a role. Raychaudhuri and Altman (2003), developed a method that assesses whether a group of genes is related by measuring the similarity of literature attributed to group members. Wren *et al*. (2004), use a thesaurus-based approach like we do. They use a probabilistic approach to identify whether a group of genes is functionally cohesive according to the literature and identify which terms connect the genes. Different from the previous two methods, the ACS does not assess functional coherence of groups. Instead, distances between concepts in the ACS reflect relatedness. Groups can be identified by clustering, as we shall illustrate, but relations between concepts can also be visualized. In this way a molecular biologist can quickly and intuitively inspect, based on a set of literature about a group of genes, relationships between these genes and other concepts associated with these genes.

In this paper we will assess whether the ACS is useful for molecular biologists. We will do this by evaluating how well the positioning of genes in the ACS reflects actual functional biological relationships between genes. This is the first systematic evaluation of the ACS on real data. A test set is constructed based on groups of genes that are known to be functionally related. We measure how well the method reproduces these groupings based on the literature about the genes. The performance of the ACS will be compared to a simple approach that only uses co-occurrences of genes. The results of the quantitative analysis are thoroughly reviewed in an attempt to understand the underlying phenomena. Additionally, we demonstrate how the ACS may assist molecular biologists in the interpretation of DNA microarray data.

MATERIALS AND METHODS

Selection of gene groups

We chose five groups of genes, each defined by a different aspect of gene biology, being function, organelle, biological process, metabolic pathway or association with a disease. Only human genes were taken into consideration. Three groups were derived from the functional annotation by the Gene Ontology (GO) annotation project (Ashburner *et al*., 2000; Camon *et al*., 2004) as stated in the Locuslink database of June 19, 2003. As evidence, the following annotation tags were accepted as being trustworthy: IDA, TAS, IGI, IMP, IPI, ISS (see http://www.geneontology.org/doc/GO.evidence.html). Note that the most prevalent annotation, inferred from electronic annotation (IEA), was not accepted as sufficient proof. The other two groups were acquired by alternative approaches. For genes associated with a disease, a review on breast carcinomas was used to identify eight genes regularly associated with this type of cancer (Keen and Davidson, 2003). For a metabolic pathway we used the KEGG database (Kanehisa and Goto, 2000) to identify the 10 genes involved in glycolysis in man. The selected groups are:

(1) Spermatogenesis; GO code 0007283, 41 genes, a biological process;

(2) Lysosome; GO code 0005764, 25 genes, an organelle;

(3) Chaperone activity; GO code 0003754, 23 genes, a biological function;

(4) Breast cancer; review, 8 genes, genes related to a disease;

(5) Glycolysis; KEGG database, 10 genes, a metabolic pathway.

None of the selected genes occurred in more than one group. From these genes, only those for which at least 10 abstracts could be retrieved by a PubMed query were added to the set used for the evaluation.

Selection of literature

Literature was selected by a PubMed query performed for each gene (Wheeler *et al*., 2003). The query was composed of gene symbols, including aliases, and full names that were derived from Locuslink (Wheeler *et al*., 2003). To

avoid the use of ambiguous terms in the query, we only used full gene names or gene symbols with a number in it. Gene symbols or full gene names that refer to more than one LocusLink gene were rejected as well. The accepted gene names and gene symbols were combined by 'OR' and were required to be found as text words. Additional requirements were the presence of the MeSH annotation 'Human' and an electronic publication date (EDAT) between 1-1-1965 and 31-12-2002.

Some genes within our set were found in many more abstracts than others. To assess how the number of abstracts per gene affects the outcome, three sets of literature were produced. For the first set, each gene contributed exactly 10 abstracts to the set randomly selected from the set of all abstracts available for that gene. In the second set, each gene contributed a maximum of 100 abstracts, though some contributed less. Similarly, we constructed a 1000 abstracts set. For each set, three versions were made to account for sampling effects. To assess the sensitivity to changes in the literature set we also experimented by adding 10 000 Medline abstracts randomly selected from those published in the years from 1997 to 2002.

Indexing

In this context, indexing means the identification of thesaurus concepts in text. The thesaurus we used for indexing was composed of three parts: the freely available thesauri/ontologies MeSH and GO, and a LocusLink derived human gene thesaurus. For each gene in the thesaurus, we considered all fields from LocusLink describing gene symbols, gene names, aliases and product names as synonymous. To match a common spelling variation, for every symbol that ends with a number, we also added to the thesaurus the same symbol with the number separated by a hyphen or a space and vice versa. For every word in the thesaurus we included the uninflected form produced by the normalizer of the lexical variant generator (McCray *et al*., 1994).

MEDLINE titles, MeSH headings and abstracts, if available, were indexed using Collexis software [http://www. collexis.com and Van Mulligen *et al*. (2002)]. Each concept found was assigned a concept weight to represent the importance of the concept for a particular citation. A document was thus represented by an M-dimensional vector $W = (w_1, w_2, \ldots, w_M)$, where M is the number of distinct concepts in the thesaurus, and $w_i = 0$ if t_i is not in the document. A concept's weight is defined as term frequency (TF) times inverse document frequency (IDF):

$$w_i = \mathrm{TF} \times \mathrm{IDF} = f_i \times \left({}^2\log \frac{N}{N_i} + 1 \right).$$

TF is the number of occurrences f_i of a concept t_i in a given document. IDF is a correction factor for the number of documents N_i containing t_i in a given set of N documents. Frequently occurring concepts, or general concepts, are thus given a lower weight. To calculate IDF we used 10 years of Medline. For each concept fingerprint the weights were normalized, i.e. divided by the largest value.

ACS and gene co-occurrence

For the gene co-occurrence method two genes co-occur if they are both found in the abstract, title or MeSH headings of one document. The gene co-occurrence method is based on a co-occurrence matrix. The matrix contains the number of times genes from the set co-occur.

The ACS is a multidimensional Euclidean space, in which concepts are positioned. For the ACS per document only co-occurrences of concepts with a weight above a threshold are used, to diminish the impact of general terms. Concepts are positioned based on their co-occurrences, one-step relations and multi-step relations. For example, a two-step relation exists between concepts X and Z if they both co-occur with a concept Y. Concepts that are connected by many co-occurrence paths, either one step or multistep, are expected to have a small distance in the ACS, while concepts with few or no paths between them should be far apart. The algorithm starts by randomly positioning the concepts in the ACS. Subsequently, for each fingerprint, co-occurring concepts are moved towards each other. After all fingerprints are processed in this manner, the concept cloud is expanded and all concepts are moved away from each other. These attraction and relaxation steps are repeated until the relative position of the concepts is stable. In this way single and multi-step co-occurrence relations are mapped to a Euclidean space. The idea behind the algorithm is that in the ACS, distance between concepts takes the meaning of a semantic relatedness. The dimensions in this space do not have a meaning; they only accommodate the placing of concepts relative to each other. For a more detailed description of the algorithm see Van Der Eijk *et al*. (2004).

For the learning of the ACS, standard settings were used. The ACS algorithm iterated 150 times and every ACS had 10 dimensions. Because the ACS algorithm has a random initialization, the final positioning of concepts can be different each time a new ACS is built, even with the same literature set. To take this factor into account, we built and evaluated an ACS three times for each literature set. The results of the evaluation were averaged.

Evaluation

Both the ACS and the gene co-occurrence method were employed to produce a ranking of the set of genes relative to one so-called seed gene. All genes in turn served as a seed, producing a ranking for each of the 53 genes in our set. For the gene co-occurrence method, the genes from the set are rank-ordered according to their number of co-occurrences with the seed. Ties are ordered randomly. For the ACS, genes from the set are rank-ordered according to their Euclidean distances to the seed gene.

For each gene a receiver operating characteristics (ROC) curve was then constructed. ROC curves are commonly used to evaluate classifiers (Swets, 1988). They are two-dimensional (2D) graphs in which the true-positive (TP) rate is plotted against the false-positive (FP) rate. The TP rate is defined as correctly classified positives divided by all positives. The FP rate is defined as incorrectly classified negatives divided by all negatives. For each seed the set of genes was divided in two classes: members from the same functional group as the seed (positives) and non-group members (negatives). As input for the ROC curve served the set of genes ranked relative to the seed. The TP and FP rates were calculated for every rank. The area under the curve (AUC) was used as a performance measure (Hanley and McNeal, 1982). This value varies between 0 and 1. An AUC of 1 represents perfect ordering, i.e. all positives are at the top of the list with no negatives in between. The AUC has the useful property that a value of 0.5 represents random ordering (Hanley and McNeal, 1982). This property provides us, in a way, with built-in negative control.

To determine whether the AUC scores differed significantly between the two methods, we used the non-parametric Wilcoxon signed ranks test. The test requires the AUC scores of the genes to be independent. Because this is not true in this case, we had to apply bootstrapping (Efron and Gong, 1983) to estimate the distribution of the Wilcoxon test statistic. We generated 100 new sets of genes by sampling genes from the original set with replacement. The sampling was stratified over the five gene groups to obtain groups of equal size as in the original set. In the subsequent selection of literature every gene appearing more than once in the set was given the same set of literature, but with different IDs. This is important for the ACS as we have observed that the size of the literature set can have an influence. During indexing duplicate genes are treated as synonyms. AUCs are calculated for both simple gene co-occurrence and ACS, and the Wilcoxon signed ranks test is applied to measure the difference between the two methods per gene group. These 100 results are used to determine if the two methods differ in performance at the 0.05 level.

It is possible that relations exist between genes in different gene groups. In order to evaluate whether this is the case, we manually checked 108 of all possible 1081 inter-group gene pairs for functional biological relationships. Information sources used were GO annotations, KEGG, Online Mendelian Inheritance in Man (OMIM, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM), abstracts in which a co-occurrence was observed, and Swiss-Prot (http://www.expasy.org/sprot/). Relationships were acknowledged if they were of the following types: same or similar biological process, biological function, specific organelle, metabolic pathway, protein family, direct interaction or association with the same disease.

For visual inspection of the multi-dimensional ACS we utilized Sammon mapping (Sammon, 1969), which reduces the dimensionality to two, and hierarchical clustering as introduced for microarray analysis in Eisen *et al*. (1998). To apply the latter, the ACS coordinates of the set of genes were translated so that the center of the set was at the origin. The resultant coordinates were used as input for the clustering program. We used average linkage clustering with correlation (uncentered) as similarity metric.

RESULTS

We selected 5 groups of genes, with a total of 53 genes (Table 1). Genes in each group share a distinct functional biological characteristic: role in spermatogenesis, breast cancer, glycolysis, lysosome or chaperone activity. MEDLINE abstracts were selected by PubMed queries for every gene. For the PubMed query we used gene names and symbols extracted from Locuslink, excluding most of the ambiguous terms. We only included a gene in our study if at least 10 abstracts could be retrieved. The median number of retrieved articles for the breast cancer genes (median 2674) and glycolysis genes (median 787) is considerably higher than for the chaperones (median 61), lysosome genes (median 127) and spermatogenesis genes (median 21.5). The same tendency holds for the number of co-occurrences between a gene and other genes from the set (Table 1, medians in same order: 15, 12, 3, 6, 2). Twenty-nine genes co-occur with five or less genes and seven do not co-occur with any. To evaluate the quality of the grouping we estimated the amount of accidental intergroup relationships. From all 1081 possible pairs of genes from different groups, we manually assessed 108 (10%) randomly picked pairs for functional biological relationships. Seven gene pairs (6.5%) were found to have a relevant relationship (Table 2).

The ACS and simple gene co-occurrence were employed to produce for each gene, termed the seed, a ranking of the other 52 genes. A perfect ranking is when all genes that have a functional biological relationship with the seed, rank highest. To produce these rankings we used for the ACS the distances between genes and for simple co-occurrence the count of gene co-occurrences. To assess the quality of the rankings we determined ROC curves and used the area under the ROC curve as an outcome measure (Metz, 1978; Hanley and McNeal, 1982). The AUC has a value of 1 for perfect ordering, 0.5 for random ordering, and 0 for the worst possible ordering (genes related to the seed have the lowest ranks). We varied the maximum number of abstracts a gene could contribute to the set of literature used in the analysis, to take into account that some genes were mentioned in thousands of abstracts whereas others are only mentioned in ten.

Figure 1 shows the performance of both the gene co-occurrence method and the ACS for the five gene groups. For the gene co-occurrence method performance for the chaperone, lysosome and spermatogenesis groups is not much better than random ordering, with AUC scores close to 0.5. These low scores are explained by a lack of co-occurrence between its group members (Table 1). For the breast cancer and glycolysis groups, performance is moderate for 10 abstracts per gene and improves when more abstracts are available. For the literature set of max 1000 abstracts per gene the score is good for the breast cancer genes (median 0.85) and excellent for the glycolysis genes (median 0.97). The addition of 10 000 randomly selected abstracts to the literature set does not affect the scores much. We found that only very few gene co-occurrences were extracted from these additional abstracts. Following from the AUC scores and also from a manual evaluation of co-occurrences of genes from different groups showed that almost all correctly found co-occurrences did represent actual biological relationships. Some wrong associations were found as a consequence of incorrect indexing due to ambiguity in the gene names. Pairs of genes with a general, though not a functional, biological relationship, were also found several times, such as the localization of two genes on the same chromosome, e.g. MPO and ERBB2.

The results for the ACS show that the ranking of genes scores better than random arrangement for all groups, except for the group of chaperones. The breast cancer genes have very high scores for the first three literature sets (median up to 0.93 for maximally 1000 abstracts per gene). The glycolysis group also has a very high score for the first two sets (0.92 for maximally 100 abstracts per gene) but decreases (median 0.8) for the set of 1000 abstracts per gene. The spermatogenesis group scores best for the set of 1000 abstracts per gene (median 0.88). The lysosome group scores best for the smallest literature set (median 0.75). The addition of 10 000 random abstracts results in a substantial decrease of the AUC scores for most groups. When the Wilcoxon signed ranks test is used in combination with bootstrapping, the ACS performs significantly better than the gene co-occurrence method for all literature sets when all gene groups are combined, but not when random literature is added. Results for the same test on a per group level are shown in the figure. For only 10 abstracts per gene ACS performance is better for the breast cancer group. ACS performs better in all literature sets for the spermatogenesis group. We observe that when no randomly selected abstracts are added and the chaperone group is not considered (see below), the ACS tends to score higher for all groups, with the sole exception being the glycolysis group in the literature set of 1000 abstracts per gene. We should note that due to the small size of the gene groups, statistical power is limited. The gene co-occurrence method only performs better when the randomly selected abstracts are added and only for the breast cancer group. The wrongly annotated genes and their effect on performance will be discussed below.

There are large differences in scores between different groups as well as between individual genes from the same group. The group of chaperones was not retrieved by both methods. In Table 1 it is shown that its group members have relatively few publications per gene and only a small number of gene co-occurrences. Upon closer inspection, it appeared that typical terms for chaperone activity were very scarce (except for TCP1). Chaperone activity was almost never the topic of the abstracts for these genes. Not surprisingly, in most cases, these genes were mentioned in the context of a disease or syndrome.

For the ACS, some of the genes have scores far below 0.5, which indicates that they were placed away from their group members. Especially the genes from the lysosome group have a large range of scores. Analysis of the function of some of these genes showed interesting results. To visually inspect the ACS, we made a 2D projection of a typical ACS for the literature set of a maximum of 100 abstracts per gene (Fig. 2). Six genes of the lysosome group had relatively low AUC scores ($\leq$0.6). It turned out that the gene products of TM4SF3, LRP2, RAMP2, RAMP3 and ADRB2 are not active in the lysosome. As can be seen in Figure 2, they are positioned dispersed and away from the majority of the lysosome genes. TM4SF3 is a membrane protein and was assigned the GO annotation via an apparently incorrect 'traceable author statement' (Gwynn *et al*., 1996). For the other genes their products are either a receptor or part of a receptor at the cell surface. LRP2 is a multiligand endocytic receptor, which binds molecules and facilitates their internalization by endocytosis

Table 1. Selected genes from five functional groups

LL-id	Gene symbol	Gene name	# abs	CC genes	Group
325	*APCS*	Serum amyloid P component	325	7	Chaperone activity
3998	*LMAN1*	Mannose-binding lectin 1	61	6	Chaperone activity
6102	*RP2*	Retinitis pigmentosa 2	96	4	Chaperone activity
6687	*SPG7*	Spastic paraplegia 7	17	0	Chaperone activity
6950	*TCP1*	t-complex 1	35	3	Chaperone activity
7249	*TSC2*	Tuberous sclerosis 2	279	3	Chaperone activity
11140	*CDC37*	Cell division cycle 37 homolog (*S.cerevisiae*)	17	0	Chaperone activity
154	*ADRB2*	Beta-2-adrenergic receptor	1309	9	Lysosome
410	*ARSA*	Arylsulfatase A	434	10	Lysosome
411	*ARSB*	Arylsulfatase B	101	5	Lysosome
412	*STS*	Steroid sulfatase	286	12	Lysosome
1200	*CLN2*	Neuronal ceroid-lipofuscinosis 2	112	3	Lysosome
2548	*GAA*	Acid alpha-glucosidase	312	34	Lysosome
2581	*GALC*	Galactosylceramidase	201	4	Lysosome
3916	*LAMP1*	Lysosomal-associated membrane protein 1	36	6	Lysosome
4036	*LRP2*	Low density lipoprotein-related protein 2	127	2	Lysosome
4353	*MPO*	Myeloperoxidase	3837	17	Lysosome
4758	*NEU1*	Sialidase 1	753	14	Lysosome
7103	*TM4SF3*	Transmembrane 4 superfamily member 3	16	0	Lysosome
8692	*HYAL2*	Hyaluronoglucosaminidase 2	27	7	Lysosome
10266	*RAMP2*	Receptor (calcitonin) activity modifying protein 2	47	3	Lysosome
10268	*RAMP3*	Receptor (calcitonin) activity modifying protein 3	22	2	Lysosome
226	*ALDOA*	Fructose-bisphosphate aldolase A	1853	10	Glycolysis
2023	*ENO1*	Enolase 1	2550	14	Glycolysis
2597	*GAPD*	Glyceraldehyde-3-phosphate dehydrogenase	26	18	Glycolysis
2821	*GPI*	Glucose phosphate isomerase	1015	15	Glycolysis
5230	*PGK1*	Phosphoglycerate kinase 1	110	7	Glycolysis
5236	*PGM1*	Phosphoglucomutase 1	558	8	Glycolysis
2302	*FOXJ1*	Forkhead box J1	13	4	Spermatogenesis
2492	*FSHR*	Follicle stimulating hormone receptor	310	6	Spermatogenesis
2649	*NR6A1*	Nuclear receptor subfamily 6, group A, member 1	13	1	Spermatogenesis
3010	*HIST1H1T*	Histone 1, H1t	26	0	Spermatogenesis
3206	*HOXA10*	Homeo box A10	33	3	Spermatogenesis
3640	*INSL3*	Insulin-like 3	36	2	Spermatogenesis
5619	*PRM1*	Protamine 1	74	3	Spermatogenesis
5620	*PRM2*	Protamine 2	65	3	Spermatogenesis
6046	*BRD2*	Bromodomain containing 2	26	2	Spermatogenesis
6847	*SYCP1*	Synaptonemal complex protein 1	10	0	Spermatogenesis
8287	*USP9Y*	Ubiquitin specific protease 9, Y chromosome	11	2	Spermatogenesis
8607	*RUVBL1*	RuvB-like 1 (*E.coli*)	11	0	Spermatogenesis
8900	*CCNA1*	Cyclin A1	17	2	Spermatogenesis
9191	*DEDD*	Death effector domain containing	14	0	Spermatogenesis
9240	*PNMA1*	Paraneoplastic antigen MA1	30	4	Spermatogenesis
23626	*SPO11*	Sporulation protein, meiosis-specific, SPO11 homolog (*S.cerevisiae*)	11	1	Spermatogenesis
672	*BRCA1*	Breast cancer 1, early onset	2674	12	Breast Cancer
675	*BRCA2*	Breast cancer 2, early onset	1530	8	Breast cancer
1956	*EGFR*	Epidermal growth factor receptor	7502	22	Breast cancer
2064	*ERBB2*	Erythroblastic leukemia viral oncogene homolog 2	2791	16	Breast cancer
2066	*ERBB4*	Erythroblastic leukemia viral oncogene homolog 4	227	7	Breast cancer
2099	*ESR1*	Estrogen receptor 1	36	15	Breast cancer
5241	*PGR*	Progesterone receptor	3656	22	Breast cancer
5915	*RARB*	Retinoic acid receptor, beta	15	5	Breast cancer
7157	*TP53*	Tumor protein p53	19919	29	Breast cancer

Given are the LocusLink identification number (LL-id), preferred symbol, gene name, number of abstracts retrieved by the PubMed query (#abs)number of genes from the set with which the gene co-occurs, taking all abstracts into account (CC genes), and the functional group to which the gene belongs.

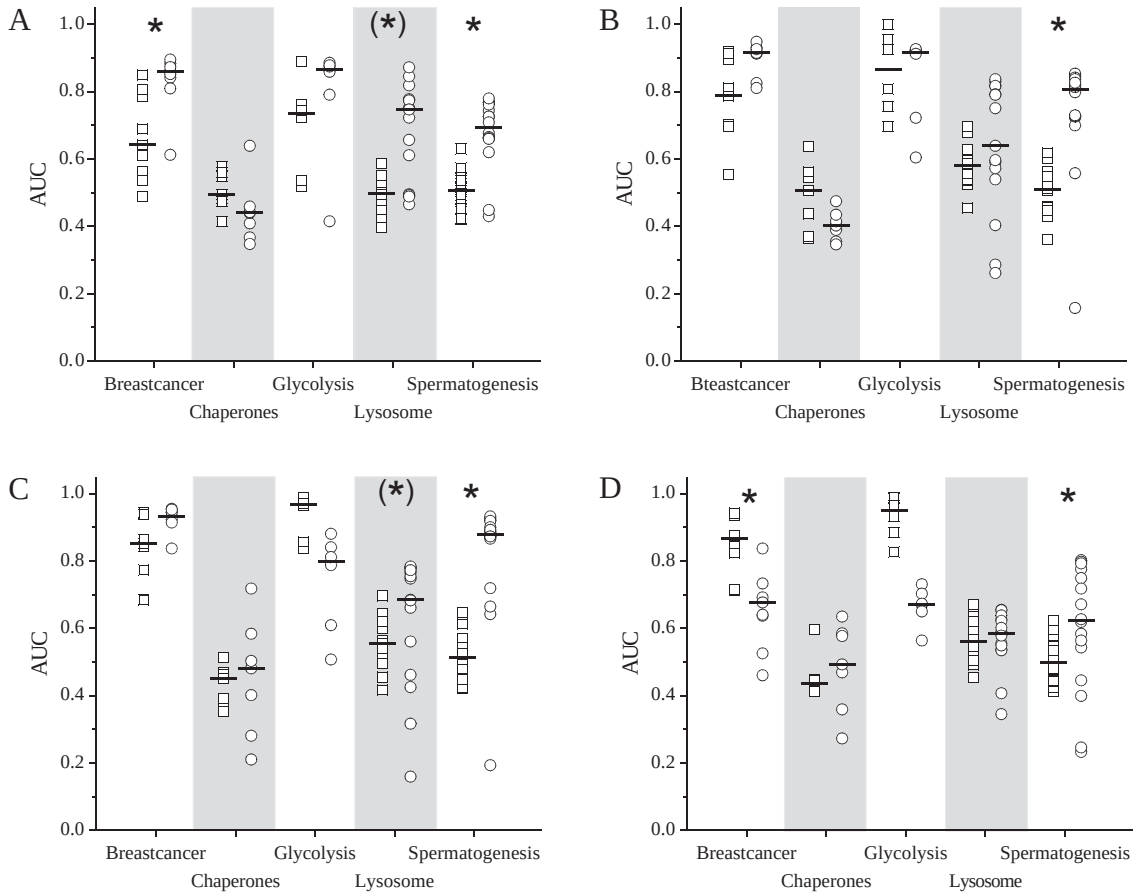

Fig. 1. AUC scores for individual genes per group for the gene co-occurrence method (open boxes) and the ACS (open circles). The different graphs represent results for the different literature sets: (**A**) 10 abstracts per gene, (**B**) maximum 100 abstracts per gene, (**C**) maximum 1000 abstracts per gene, (**D**) maximum 1000 abstracts per gene + 10 000 randomly selected abstracts. An asterisk, above a group indicates that the difference in performance of the two methods is statistically significant (at the 0.05 level). An asterisk in parentheses indicates that the difference would be statistically significant if wrongly annotated genes are removed (see Results section).

Table 2. Found inter-group functional biological relations

Description of relation	Gene pair	PMID
Both involved in		
Cancer	RARB–PNMA1	10050892
Cancer	PGR–PNMA1	10050892
Cancer	RAMP2–TP53	11420706
Alzheimer	MPO–APCS	12052532, 12015594
Cryptorchidism	STS–INSL3	6135610, 10319319
Female reproductive cycle	FSHR–ESR1	11089565, 10342864
Epilepsy	BRD2–TSC2	12830434

The last column gives examples of PubMed identification numbers of articles that support the relationship identified.

(Nykjaer *et al*., 1999). After internalization these endosomes can become lysosomes (Lisi *et al*., 2003). RAMP2, RAMP3 and ADRB2 are involved with the activity of receptors whose activity is regulated, upon agonist activation, via internalization and degradation in a lysosome (Kuwasako *et al*., 2000; Gagnon *et al*., 1998). If the lysosome group would have been better defined without these genes, the score for the lysosome group would have been improved, up to a median of 0.87 for the set of 100 abstracts per gene. Using the statistical test mentioned earlier, the ACS would perform significantly better than using simple gene co-occurrence for the set of 10 abstracts per gene and a maximum of 1000 abstracts per gene. Interestingly, RAMP2 and RAMP3 are placed near breast cancer genes (Fig. 2). In the abstracts retrieved for these genes, they were not directly implicated in cancer nor directly linked to the breast cancer genes. The only exceptions are two co-occurrences between TP53 and RAMP2 in non-cancer related abstracts. These proteins are involved with the adrenomedullin receptor. Adrenomedullin is an angiogenic factor, and has been linked to a response in (breast) cancer cells in solid tumors that protects against hypoxic cell death (Martinez *et al*., 2002; Oehler *et al*., 2001; Zudaire *et al*., 2003). Because of the role of adrenomedullin in cancer, the genes that are associated with its receptor are also relevant for the study of cancer (Fernandez-Sauze *et al*., 2004).

The position of some genes in the ACS is not easily explained in terms of functional relations. The spermatogenesis gene PNMA1,

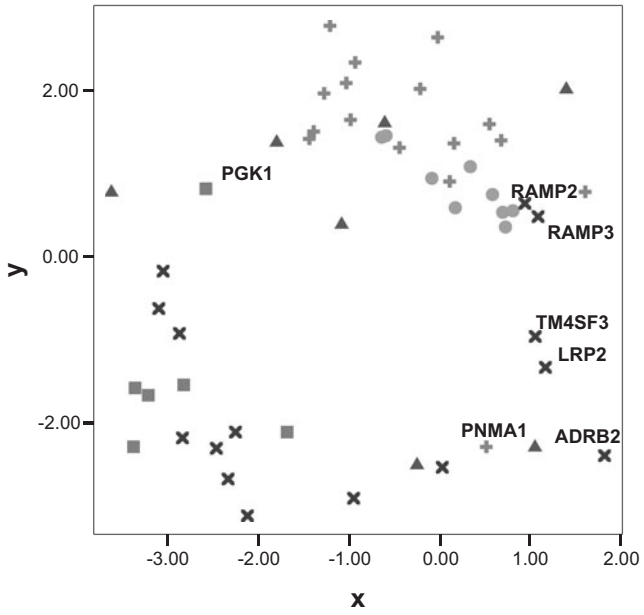

Fig. 2. Two-dimensional projection of the ACS with Sammon mapping. The different groups are marked; triangles, Chaperone activity; circles, Breast cancer; squares, Glycolysis; cross symbols, Lysosome; plus symbols, Spermatogenesis. Some genes with aberrant behavior are labeled.

for instance, is placed apart from the spermatogenesis group. Its position was found to be caused by an ambiguity problem. The term MA1 is a synonym for the gene and was used for the PubMed query, but is unfortunately also used for numerous other concepts, such as monoclonal antibody 1. The glycolysis gene PGK1 had very different AUC scores for different builds of the ACS for the same literature set. Apparently it did not find a stable position in the ACS. A study of the abstracts in which PGK1 was mentioned showed that it was referred to in a large number of very different contexts and only occasionally in the context of its role in glycolysis. Given the different contexts in which the gene is mentioned, it is hard to imagine how it could be placed in the ACS so that its surroundings correctly reflect all contexts.

In practical applications of the ACS where the labeling of genes is unknown, clustering algorithms can be used to provide a grouping of the genes of interest. Figure 3 shows an example of how a standard clustering technique can be applied. The result for three clusters gives a group of genes with roles in cell-cycle control, regulation of gene expression and other forms of DNA–protein interactions (cluster 1), a group mostly containing genes with ambiguity problems or deviating annotations (cluster 2), and a group of enzymes (cluster 3, except for the chaperones). If we allow 14 clusters, 4 clusters contain only one gene and half of the remaining 10 clusters contain only genes which share a functional biological relationship.

DISCUSSION

Our experiments show that the positioning of genes in the ACS reflects functional biological relationships. Four of the five functional groups that were tested were clustered very well (median AUC > 0.85), if we exclude the aberrant annotations from the lysosome group. Genes with aberrant annotations were correctly placed away from their supposed group members. Interestingly, the ACS placed two of these genes, RAMP2 and RAMP3, in the breast cancer cluster, while there are hardly any co-occurrences between these genes and the breast cancer genes. A study of the literature revealed that a relation to breast cancer is supported by, among others, the role of these genes in angiogenesis. Although not the focus of this study, this is an example of how the ACS could be useful as a knowledge discovery tool.

Both the gene co-occurrence method and the ACS show large differences in performance for different groups. The genes from the breast cancer group have very good scores for both methods. The genes from the glycolysis group score close to perfect for the gene co-occurrence method and good for the ACS. For both groups, the number of abstracts retrieved per gene was high when compared to the other groups. The group of chaperones could not be reconstructed by both methods. The poor performance for this group is explained by the scarce reference to their chaperone activity. Clearly, the relationships that the text meta-analysis tools can extract are limited to those described in the literature set, and biomedical abstracts are better represented in Medline than those about basic biology. For the spermatogenesis and lysosome groups it appears that the genes are frequently referred to in the expected context. The ACS method can reproduce both the lysosome and spermatogenesis groups quite well. The gene co-occurrence method, on the other hand, scored very poorly for these groups. Most gene co-occurrences between genes do reflect actual biological relations; see also Jenssen *et al*. (2001). The low scores for these groups were caused by a lack of actual gene co-occurrences. As these groups could be retrieved with the ACS, this is a clear indication that additional information from the abstract should be used.

The ACS can produce good results with a limited amount of literature, such as only 10 abstracts per gene. This is an important feature as for a large number of genes limited literature is available. In contrast to the gene co-occurrence method though, the performance of the ACS was affected by the addition of large amounts of randomly selected abstracts. We hypothesize that the effect of the addition of these abstracts is caused by the appearance of new relationships between concepts. In order to reflect these changes, these concepts will move away from their original positions in the ACS, which apparently disrupts a meaningful clustering. With more relations added the ACS has more problems with accurately representing them in its Euclidean space. This finding makes it necessary to focus the selected literature set on the studied genes, e.g. similar to what we did for the automated selection of literature in this paper. Though this is not a large limitation, it is important to take it into account. We are currently working on improving the robustness of the ACS algorithm.

The use of homonymous gene names is widespread and has a large impact on text mining applications. The amount of different meanings for one symbol can be quite startling, especially for gene symbols that are two or three letter acronyms, such as ER or GAA (Weeber *et al*., 2003a). We therefore adapted our literature selection to exclude ambiguous gene symbols. For some genes this will have reduced sensitivity, as their preferred terms (according to LocusLink) are ambiguous acronyms, e.g. GAPD, MPO and PGR. While the selection step did reduce the amount of ambiguity in our literature set, the manual analysis still revealed some problems with indexing, such as GAA which is also a DNA sequence or also, as in the case

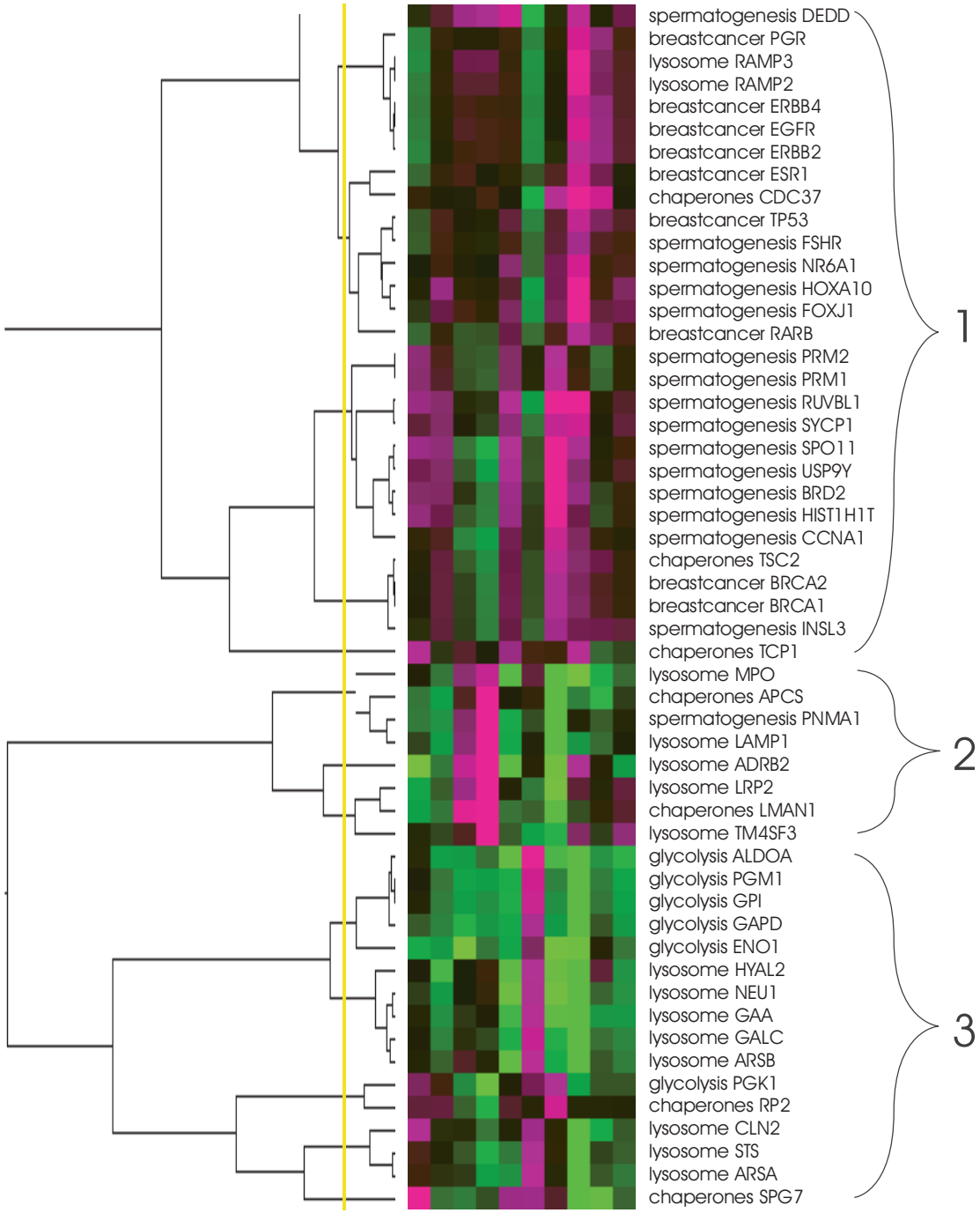

Fig. 3. Analysis of the structure of the ACS by hierarchical clustering. The rows represent the different genes and the columns represent the ten axes of the 10D ACS. The marked clusters indicate by approximation: 1, genes with roles in cell-cycle control, regulation of gene expression and other forms of DNA–protein interactions; 2, genes with ambiguity problems or deviating annotation and 3, enzymes. The yellow line in the clustering tree indicates 14 clusters.

of PGK1, when the promoter is intended instead of the gene. Clearly methods for disambiguation are needed. Word sense disambiguation has been studied for years (Liu *et al*., 2001; Resnik and Yarowsky, 2000), but only recently has a disambiguation tool been developed specifically for the disambiguation of gene names (Podowski *et al*., 2004). A tool for the disambiguation of gene names will be built into our indexing engine as soon as possible. Another common case of ambiguity is that a gene symbol can refer to the gene itself, its associated mRNA or relevant proteins. In this paper we chose not to distinguish between genes, mRNA or proteins. Such a distinction will sometimes be artificial, difficult to achieve (Hatzivassiloglou *et al*., 2001) and not relevant for our purposes.

Currently, a problem with biomedical literature mining tools is the lack of gold standards and established evaluation procedures (Shatkay and Feldman, 2003). The evaluation method we used handles this problem by depending on external and high-quality functional annotation of genes. The use of a genuine list of differentially expressed genes derived from microarray experiments for a quantitative analysis of performance is difficult and requires a substantial investment, as the annotation process would require extensive reading of scientific literature and extensive expert knowledge. While such annotated datasets are available for several organisms, to our knowledge no exhaustive annotation has been performed on a microarray dataset for human genes. The evaluation set we used was limited in size with its 5 functional groups and 53 genes, and this allows for the detailed and useful analysis we performed. The five gene groups were drawn from the broad category of functional biological relationships between genes to reflect the broad types of relations in biology. The generalizations concerning ACS performance that we can make based on our results apply only to this category. The manual annotation of genes with GO codes gives a gold standard and has been used by several authors (Raychaudhuri and Altman, 2003; Wren *et al*., 2004). It is far from perfect though, as our analysis showed that almost half of the genes from the lysosome group had only a remote connection to the lysosome. These cases did have an impact on performance, as the ACS correctly positioned them away from the lysosome group.

The outcome of a DNA microarray experiment can be a sizable set of genes (>100) that are differentially expressed. A tool to quickly identify the genes that according to literature have a functional biological relationship, would facilitate the identification of biological processes underlying the gene expression profile and assist in selecting genes for further analysis. Since distances between genes in the ACS reflect functional biological relatedness, the ACS offers an intuitively appealing presentation that can be of value for molecular biologists. We are currently developing a user-friendly and interactive interface to allow for better browsing of the ACS for genes, related concepts and their relations and to give easy access to descriptions of concepts, database entries for genes and the underlying literature.

In conclusion, the positioning of genes in the ACS reflects functional biological relationships. When the literature set is focused on the studied genes, performance of the ACS is good to excellent A focused literature set is important, as it was shown that when large amounts of randomly selected abstracts are added, performance decreases. When compared to a simple gene co-occurrence method, the ACS is capable of revealing more functional biological relations and can achieve results with less literature available per gene. The ACS can be of value for researchers studying large numbers of genes, for example in DNA microarray analyses.

ACKNOWLEDGEMENT

We would like to thank Theo Stijnen for his advice on the statistical test.

REFERENCES

Al Shahrour,F. *et al*. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Becker,K.G. *et al*. (2003) PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics*, **4**, 61.

Camon,E. *et al*. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with gene ontology. *Nucleic Acids Res.*, **32** (Database issue), D262–D266.

Chaussabel,D. and Sher,A. (2002) Mining microarray expression data by literature profiling. *Genome Biol.*, **3**, RESEARCH0055.

Efron,B. and Gong,G. (1983) A leisurely look at the bootstrap, the jackknife, and cross-validation. *Am. Stat.*, **37**, 36–48.

Eisen,M.B. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Fernandez-Sauze,S. *et al*. (2004) Effects of adrenomedullin on endothelial cells in the multistep process of angiogenesis: involvement of CRLR/RAMP2 and CRLR/RAMP3 receptors. *Int. J. Cancer*, **108**, 797–804.

Friedman,C. *et al*. (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, **17** (Suppl. 1), S74–S82.

Gagnon,A.W. *et al*. (1998) Role of clathrin-mediated endocytosis in agonist-induced down-regulation of the beta2-adrenergic receptor. *J. Biol. Chem.*, **273**, 6976–6981.

Gwynn,B. *et al*. (1996) Genetic localization of Cd63, a member of the transmembrane 4 superfamily, reveals two distinct loci in the mouse genome. *Genomics*, **35**, 389–391.

Hanley,J.A. and McNeal,B.J. (1982) A simple generalization of the area under the ROC curve to multiple class classication problems. *Radiology*, **143**, 29–36.

Hatzivassiloglou,V. *et al*. (2001) Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics*, **17** (Suppl. 1), S97–S106.

Jenssen,T.K. *et al*. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.*, **28**, 21–28.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Keen,J.C. and Davidson,N.E. (2003) The biology of breast carcinoma. *Cancer*, **97**, 825–833.

Kuwasako,K. *et al*. (2000) Visualization of the calcitonin receptor-like receptor and its receptor activity-modifying proteins during internalization and recycling. *J. Biol. Chem.*, **275**, 29602–29609.

Lisi,S. *et al*. (2003) Preferential megalin-mediated transcytosis of low-hormonogenic thyroglobulin: a control mechanism for thyroid hormone release. *Proc. Natl Acad. Sci. USA*, **100**, 14858–14863.

Liu,H. *et al*. (2001) Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. *J. Biomed. Inform.*, **34**, 249–261.

Martinez,A. *et al*. (2002) The effects of adrenomedullin overexpression in breast tumor cells. *J. Natl Cancer Inst.*, **94**, 1226–1237.

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

McCray,A.T. *et al*. (1994) Lexical methods for managing variation in biomedical terminologies. *Proc. Annu. Symp. Comput. Appl. Med. Care.*, 235–239.

Metz,C.E. (1978) Basic principles of ROC analysis. *Semin. Nucl. Med.*, **8**, 283–298.

Nykjaer,A. *et al*. (1999) An endocytic pathway essential for renal uptake and activation of the steroid 25-(OH) vitamin D3. *Cell*, **96**, 507–515.

Oehler,M.K. *et al*. (2001) Adrenomedullin inhibits hypoxic cell death by upregulation of Bcl-2 in endometrial cancer cells: a possible promotion mechanism for tumour growth. *Oncogene*, **20**, 2937–2945.

Podowski,R.M., Gleary,J.G., Goncharoff,N. and Hayes,W.S. (2004) AzuRE, a scalable system for automated term disambiguation of gene and protein names. *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference*, Stanford, CA, 415–424.

Pustejovsky,J., Castano,J. *et al*. (2002) Robust relational parsing over biomedical literature: extracting inhibit relations. *Pac. Symp. Biocomput.*, 362–373.

Raychaudhuri,S. and Altman,R.B. (2003) A literature-based method for assessing the functional coherence of a gene group. *Bioinformatics*, **19**, 396–401.

Resnik,P. and Yarowsky,D. (2000) Distinguishing systems and distinguishing senses: new evaluation tools for words sense disambiguation. *Natural Language Engineering*, **5**, 113–133.

Sammon,J. (1969) A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.*, **C-18**, 401–409.

Shatkay,H. and Feldman,R. (2003) Mining the biomedical literature in the genomic era: an overview. *J. Comput. Biol.*, **10**, 821–855.

Srinivasan,P. (2004) Text mining: generating hypotheses from MEDLINE. *JASIST*, **55**, 396–413.

Stapley,B.J. and Benoit,G. (2000) Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pac. Symp. Biocomput.*, 529–540.

Swanson,D.R. and Smalheiser,N.R. (1997) An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artif. Intell.*, **91**, 183–203.

Swets,J.A. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

Tanabe,L. *et al*. (1999) MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques*, **27**, 1210–1217.

Van Der Eijk,C.C. *et al*. (2004) Constructing an associative concept space for literature-based discovery. *JASIST*, **55**, 436–444.

Van Mulligen,E.M. *et al*. (2002) Research for research: tools for knowledge discovery and visualization. *Proc. AMIA Symp.*, 835–839.

Weeber,M. *et al*. (2003a) Ambiguity of human gene symbols in LocusLink and MEDLINE: creating an inventory and a disambiguation test collection. *Proc. AMIA Symp.*, 704–708.

Weeber,M. *et al*. (2003b) Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. *J. Am. Med. Inform. Assoc.*, **10**, 252–259.

Wheeler,D.L. *et al*. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.

Wren,J.D. *et al*. (2004) Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics*, **20**, 389–398.

Wren,J.D. and Garner,H.R. (2004) Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. *Bioinformatics*, **20**, 191–198.

Zhang,B. *et al*. (2004) GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics*, **5**, 16.

Zudaire,E. *et al*. (2003) Adrenomedullin and cancer. *Regul. Pept.*, **112**, 175–183.