BIOINFORMATICS

Vol. 21 Suppl. 2 2005, pages ii259–ii267
doi:10.1093/bioinformatics/bti1143

Text Mining

Expert knowledge without the expert: integrated analysis of gene expression and literature to derive active functional contexts

Robert Küffner*, Katrin Fundel and Ralf Zimmer

Department of Informatics, Ludwig-Maximilians-Universität München,
Amalienstrasse 17 80333 München, Germany

ABSTRACT

Motivation: The interpretation of expression data without appropriate expert knowledge is difficult and usually limited to exploratory data analysis, such as clustering and detecting differentially regulated genes. However, comparing experimental results against manually compiled knowledge resources might limit or bias the perspective on the data. Thus, manual analysis by experts is required to obtain confident predictions about involved processes.

Results: We present an algorithm to simultaneously derive interpretations of expression measurements together with biological hypotheses from biomedical publications. It identifies active functional contexts ('concepts'), i.e. gene clusters that exhibit both a significant gene expression as well as a coherent literature profile. Manual intervention by an expert in specifying prior knowledge is not required. The approach scales to realistic applications and does not rely on controlled vocabularies or pathway resources.

We validated our algorithm by analyzing a current juvenile arthritis dataset. A number of gene clusters and accompanying literature topics are identified as an interpretation of the data that coincide well with the phenotype and biological processes known to be involved in the disease. We demonstrate that generated clusters are both more sensitive and more specific than Gene Ontology categories detected on the same data. The method allows for in-depth investigation of subsets of genes, the associated literature topics and publications.

Availability: Supplementary data on clusters is available upon request.

Contact: Robert.Kueffner@bio.ifi.lmu.de

1 INTRODUCTION

The interpretation of gene-expression data is a challenging task. It is generally accepted, though, that large-scale gene-expression measurements (see Xiang *et al*., 2003 for a recent review) allow to characterize specific states of biological systems, e.g. tissues or cell lines, and to identify relevant factors of the involved biological pathways. Frequently, clustering algorithms are used as the initial step in the analysis (see Jiang and Zhang, 2002; Speed, 2003 for surveys and textbooks) to generate groups of genes with similar expression profiles. The underlying assumption is that genes sharing similar expression patterns might be functionally related. As these methods usually work with sparse gene-expression data alone—typically much less than a 100 measurements for more than 10 000 genes—they are limited owing to statistical sampling errors, noise and spurious hits. Despite that, no consensus has been achieved on how to utilize the expert knowledge stored in the biomedical literature to make an analysis of expression data more robust and useful. As a partial workaround to knowledge utilization, intermediate representations have been manually derived from the literature, for instance, specialized controlled vocabularies, such as the Gene Ontology (GO, Harris *et al*., 2004) or pathway databases, such as BioCarta (http://www.biocarta.com/) or Transpath (Krull *et al*., 2003; Wingender, 2004). Such resources provide a structured view on biological information, but also have severe disadvantages. Keeping the resources up to date is tedious and time consuming. The incorporation of new biological concepts or functional annotations comes with a significant time lag so that some areas of biology are covered well, whereas others remain sketchy at best. Furthermore, concepts are forced to fit to a specific static perspective of such resources. Despite that, controlled vocabularies like GO have been very useful to inform algorithms of functional contexts, e.g. to correlate functional categories with gene-expression data (Doniger *et al*., 2003; Cheng *et al*., 2004) or annotate the results of a cluster analysis with GO terms (Gibbons and Roth, 2002). Another approach (Glenisson *et al*., 2003) reconstructs gene–GO term associations by jointly analyzing expression data and the literature.

Some methods focus on protein interactions and pathways for the interpretation of expression data while discarding much of the functional context of genes (Jenssen *et al*., 2001; Zien *et al*., 2000; Hanisch *et al*., 2002; Sohler *et al*., 2004). This way, biological hypotheses beyond the explorative data analysis can be derived but high-quality networks are required.

A way to cope with this situation might be the automated extraction of biological objects and their relations via text-mining. Although the recent BioCreative assessment on the performance of gene and protein name recognition (Fundel *et al*., 2005; Hanisch *et al*., 2005) is quite promising, the derivation of detailed interactions and, thus, of meaningful regulation contexts is still difficult. Several approaches employ fairly sophisticated literature analyses to avoid the use of manually compiled resources to derive gene annotations or relationships (Liu *et al*., 2004; Shatkay *et al*., 2000) or to annotate results from expression clustering (Hu, 2004; Blaschke *et al*., 2001; Masys *et al*., 2001). Other methods reverse the latter approach and annotate genes grouped by their literature profiles with expression data afterward (Chaussabel and Sher, 2002). Thus, gene-expression analysis

*To whom correspondence should be addressed.

is separated from literature analysis, i.e. publications cannot be fully exploited to improve the analysis of experimental measurements and vice versa.

In this paper, we present a new method that simultaneously integrates literature and gene-expression analysis in order to derive biological hypotheses. Our approach is based on two essential features: (1) a method to derive a literature topic or hypothesis from a set of genes that exhibit interesting patterns in gene-expression measurements and (2) a method to select genes that belong to the given literature topic and expression pattern. A topic or theme is defined as a consistent and coherent set of literature features (Shatkay *et al*., 2000). Our method iteratively refines literature topics based on an appropriate score indicating whether genes match the respective topic and exhibit the regulatory pattern. The final result of our method is a set of gene clusters that are significantly regulated and simultaneously represent a coherent literature topic. We call such clusters active functional contexts. They can serve as first biological hypotheses to interpret the given expression data. The topics further provide links into the literature so that topic-specific reviews can be generated automatically and navigated to investigate more refined subsets of genes.

Our approach has several important features:

(1) *Scalability*: Ability to process all protein-containing abstracts within the entire MEDLINE.

(2) *Usability*: Does not require expert intervention during the analysis procedure, does not require curated vocabularies, pathways or other manually curated resources (except for predefined lists of gene/protein names).

(3) *Precision and Recall*: Employs a state of the art engine for detecting gene names in publications, does not require prereduction to significantly regulated genes.

The paper is structured as follows: first, we review the methods to represent objects (document, terms, genes, topics) as a vector of term weights and to efficiently search a collection of objects (topic matrix) via latent semantic indexing. We then present our method to use these techniques for identifying biological hypotheses from given expression and topic matrices. We show that our method outperforms other explorative techniques without the need of manual expert intervention and predefined ontology categories by performing a comparative analysis of expression data from juvenile arthritis samples.

2 METHODS

2.1 Detection of protein and gene names in the biomedical literature

Our literature analysis is based on NCBI's PubMed database that is the most comprehensive resource for publication abstracts in the biomedical domain. To integrate information from abstracts with other data sources we need to detect the gene or protein names within documents and map them to unique symbols or database identifiers. A simple but effective method to tackle this problem has recently been developed and evaluated (Hanisch *et al*., 2005). This method is based on synonym lists, i.e. comprehensive lists of gene or protein names derived from public databases. The lists have been extensively processed by a largely automated procedure to improve precision and recall, e.g. by removing ambiguous synonyms, expanding/collapsing abbreviations, etc. (Fundel *et al*., 2005). We identify gene and protein names contained in PubMed abstracts by relaxed string matching using ProMiner. This named entity recognition produces a table of PubMed-IDs and matching protein/gene object identifiers.

Table 1. Ranks of the bigrams as returned by the mutual information score

1	Sprague dawley, de novo, epidermolysis bullosa, situ hybridization, retinitis pigmentosa, epstein barr, vice versa, charcot marie
3000	Uncoupling protein, antisocial personality, surgical resection, colon cancer, conformational change, lod theta, blood flow
10 000	carbonyl chlorophenylhydrazone, lumbar cord, protein phosphatase, neutral endopeptidase, immunotoxin saporin, rho gdi
40 000	Electronic absorption, association risk, glycosylated fructosamine, retinal breakdown, celiac mesenteric, sporadic mutation

2.2 Term selection and document representation

Documents are represented via the occurrence of certain terms. We generate a list of these terms from four sources: (1) We construct a vocabulary of terms directly from the PubMed texts. Each word in the title, abstract and MeSH terms (Nelson *et al*., 2001) of each document is a possible term. Specific English stop-words are excluded from the list of terms (ftp://ftp.cs.cornell.edu/pub/smart/english.stop). (2) We used all combinations of two words (bigrams) within a window of three consecutive words to generate additional terms and keep only those with a large enough mutual information score (Table 1). (3) Additional terms are derived from complete MeSH phrases, once with and once without a subheading qualifier and (4) identifiers for proteins as detected by our named entity approach (Section 2.1). We denote the set of used terms as $t = \{t_i | i = 1..|t|\}$. It should be noted that sources 1–3 do not require controlled vocabularies; only the fourth source builds upon precompiled lists of protein names. A frequency cutoff is applied to terms from sources 1–3 (Section 3.3). In order to reduce the number of documents, we discarded documents without proteins or with very frequent proteins only (i.e. that are mentioned in more than t_{abs} abstracts) because we expect them to contain very little specific information.

2.3 Vector space models representing weighted term occurrences in documents

Vector space models represent objects, e.g. documents from the set $d = \{d_j | j = 1 \ldots |d|\}$, via appropriate $|t|$-vectors with weights for terms from t. The set of documents d is represented for the algorithms by an $|d| \times |t|$ topic matrix A containing a row vector for each document. The term weights in a document vector are designed to balance the local importance of a term within a given document against the global frequency of the term, i.e. rare terms are considered more important than common terms. A well-known term weighting scheme is the *tfidf* score (see Sebastiani, 2002, for its application to text-mining). We use a variant of the *tfidf* score based on a Poisson model (Kim *et al*., 2001) to adjust the score to the length of the documents.

2.4 Latent semantic indexing

We use latent semantic indexing (LSI) as an additional preprocessing step for document representation (Deerwester *et al*., 1990). LSI extracts the underlying semantic structure of a set of documents instead of relying on individual words. The topic matrix A for the documents d and terms t as defined in the last section is the starting point for computing the LSI space by singular value decomposition (SVD) of A into a product of three matrices:

$$A = D\Sigma T^{\mathrm{T}}$$

where T and D are matrices of singular vectors that correspond to the terms t and documents d, respectively, and Σ is a diagonal matrix with singular

values on the main diagonal. The singular vectors describe the principal components of covariance of the terms and the documents. The importance of the singular vectors for the total covariance observed is proportional to the value of the singular values in Σ and the singular vectors and singular values are sorted according to their importance. By discarding the singular vectors with the smallest (least important) singular values the noise in term usage in the corpus can be reduced. All further analysis takes place on such an adequately truncated representation D_k or T_k, where the number of remaining dimensions k is significantly smaller than the original dimensions, i.e. $k \ll \min(|d|, |t|)$.

We will briefly describe the method of query projection (Deerwester *et al*., 1990) as a means to query and compare genes, documents and terms in the above setting. A query represents a set of objects, in our case documents, terms or genes. A term query (a set of terms q or a cluster of genes C) is projected into the LSI space via

$$\mathrm{tgp}(A_k, C) = \mathrm{tgp}(A_k, q) = q^{\mathrm{T}} T_k \sum\nolimits_k^{-1}$$

The resulting $\mathrm{tgp}(A_k, C)$ can be interpreted as a profile representing a gene cluster in SVD space. Overall, query projection is a way to represent arbitrary sets of documents or arbitrary sets of genes as a unique k-vector, called a document (dgp) or term/gene group profile (tgp). In SVD space objects or sets of objects, such as (sets of) genes, can be compared via an appropriate similarity measure on k-vectors normalized to unity length. Here, we compute the cosine of the angle between vectors, i.e. $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$.

3 THE CONCEPTMAKER ALGORITHM

3.1 Gene/protein name detection

We will briefly summarize the preprocessing of the document corpus before describing the ConceptMaker Algorithm. As of December 2004, the PubMed database contained >13 million abstracts starting from as early as 1965. We restricted our analysis to the 6.8 million abstracts published in 1990 or later. The goal of the first processing step was to detect the occurrences of protein names within these abstracts using a synonym list for human proteins and genes (Fundel *et al*., 2005). This list consists of some 29 000 different proteins and a total of 350 000 synonyms. In order to detect protein names we applied the ProMiner software (Hanisch *et al*., 2005). Within the examined abstracts, we identified 1.8 million abstracts containing protein or gene names. Although we used a human synonym list, ∼10% of the detected abstracts actually did not focus on humans but were found, owing to overlaps, in the protein name spaces. It might be desirable to include abstracts on other mammals or vertebrates in our analysis but we decided to remove abstracts on non-vertebrates based on their MeSH annotation. With the procedure outlined, so far we identified 1.7 million abstracts containing 16 100 different protein objects and 4.3 million abstract–protein links.

3.2 Mapping genes to gene-expression arrays

In this paper, we focus on gene expression measurements derived from the Human Affymetrix HG-U95Av2 microarray that detects mRNA targets on 12 600 different spots. We constructed a mapping between our protein synonym list and the array specific annotation. The major class of targets that could not be matched to our synonym list consisted of predicted ORFs and hypothetical proteins without a known gene product. Out of the remaining 11 800 spots we were able to match 11 400 to 9900 distinct entries in the synonym list, i.e. some targets were represented several times. The set of identifiers that match spot-Ids and have been detected in publication abstracts contained 8500 proteins thereby covering 10 300 spots. Further processing takes place on the 960 000 abstracts that reference proteins from this set.

3.3 Computing the singular value decomposition

We computed the singular value decomposition (SVD) using SVDLIBC (http://tedlab.mit.edu/∼dr/SVDLIBC/). According to our experiments, SVDLIBC is able to process up to about 6×10^5 abstracts on a standard workstation with 2 GB of main memory (a single precision variant is able to process 1.2×10^6 abstracts). We discarded all terms that appear within <30 abstracts and abstracts that contain <30 such terms. We disregarded documents containing only very frequent proteins ($t_{\mathrm{abs}} = 900$) resulting in 516 000 documents with 128 000 terms including 43 000 bigrams and 38 000 MeSH/qualifier terms. The bigrams returned by the mutual information score (Table 1) captured phrasal terms as well as many phrases, where the meaning would not have been obvious from the constituent words alone, e.g. blood flow. The number of SVD dimensions k is determined empirically, (in the literature 100–200 are used); we therefore set $k = 150$.

3.4 ConceptMaker: identifying contexts that are active in gene-expression measurements

The main goal of this work is to utilize gene-expression measurements as an aid to formulate questions or define queries to find significant patterns in the literature. Vice versa, a literature pattern should be valuable for refining a corresponding gene-expression pattern. Thus, a pattern needs to be shared between measurements of a limited and focused number of states of a biological system and available publications that describe many of the different possible states a given biological system can adopt. We aim to find such active functional contexts, i.e. a set of genes that are significantly regulated on the gene-expression level and also represent a coherent topic in the space of the biomedical literature. We wish to simultaneously identify such topics and the corresponding sets of genes in an exploratory, unsupervised fashion. We also have to take into account that there might be a number of different, partially overlapping coherent topics that are represented in the measurements and that one gene might participate in several topics. We will refer to the set of genes under consideration as a cluster. We will refer to the literature features that are shared by the members of the cluster as a common topic and associated clusters and topics as concepts. To measure the coherence of topics together with its significance with respect to the measured gene-expression data (i.e. the quality of a concept) we use a combined score (Section 3.5). As described above, we represent a cluster of genes or a literature topic by a query projected profile cp. To evaluate if a cluster represents a coherent topic we also need to compare all genes to a given cp and measure an overall similarity. As all the genes under consideration have been represented as terms in the singular value decomposition of the topic matrix, we can transform a set of genes into a query q and use query projection to represent it in SVD space. The resulting $\mathrm{tgp}(A_k, q)$ now summarizes the contributions from the constituent genes and, therefore, is a suitable representation of the required cluster profile cp. The similarity between a gene g and a given profile cp for cluster C is given by $\langle \mathrm{cp}, g \rangle$.

We now describe the algorithm ConceptMaker that consults the gene-expression data to check the hypotheses on the regulation of the

genes in the actual measurements. Certain genes need to be classified as belonging to the topic, others have to be discarded based on the experimental data. This modified cluster of genes will then be used to derive a new topic. The algorithm requires gene-expression measurements (e.g. by P-values quantifying the significance of expression level differences between two states) for a set of genes and a reduced topic matrix A_k (containing the measured genes in the set of terms), identifies a predefined number m of gene clusters and corresponding topics, which are both coherent according to published knowledge in the literature and significantly regulated between the two measured states. The algorithm constructs a sequence of topic matrices A_k^i and finds the best topic relative to such a matrix and the expression data. Once a cluster is found, a new matrix is constructed via orthogonalization with respect to the found cluster. For a given matrix, the algorithm (**Reduce**)—starting from a cluster of all genes—iteratively reduces the current cluster to a coherent topic (analogous to Hastie *et al*., 2000).

ConceptMaker(A)
  $A^0 = \mathbf{SVD}_k(A)$
  for ($j = 0$; j < |d|; j++)
    $C^j = \mathbf{Reduce}(A^j)$
    $A^{j+1} = \mathbf{Orthogonalize}(A^j, C^j)$
    output C^j;

Reduce(A^j)
G^1 = [all genes] ; L = [G^1]
for (i = 1; (G^i! = empty); i++)
  G = G^i;
  While (G changed)
    L = **sort** [**score**(**tgp**(A^j, G), g)| g gene]
    G = $\mathbf{best}_g$(L)
  G^{i+1} = G; L = L + +[G]
return **select**(L)

Orthogonalize(D,C)
$D_{.j} = D_{.j} - < D_{.j}, C > * \mathbf{tgp}(D_j, C)$

The ConceptMaker algorithm starts from a topic matrix and computes a k-truncated topic matrix $A^0 = A_k = (D_k, \Sigma_k, T_k)$ via singular value decomposition ($\mathbf{SVD}_k$). It then produces a sequence of m (m a user defined input parameter) modified k-truncated topic matrices A^j via orthogonalization of the current matrix with respect to found topics/clusters. For each topic matrix the procedure **Reduce** iteratively reduces the set of possible genes to a smaller set of genes with the best compatibility to a topic according to the current topic matrix. For this purpose, $\mathbf{best}_g$ selects the top $|G^i|$ * *fraction* (*fraction* being an input parameter) from a list of genes sorted according to a **score** (Section 3.5) measuring the simultaneous compatibility of genes to the gene cluster, its literature profile and regulatory pattern. In the inner While-loop the size of a gene cluster is fixed for a given I, the algorithm repeats the steps of cluster profile generation and selection of compatible genes until the gene set remains constant. In this iteration, the algorithm determines a cluster of a particular size that is most compatible with a topic, where the topic is allowed to shift, if appropriate, to a completely new one. The new cluster/topic is stored in the list of generated clusters. Finally, the set of nested gene sets is inspected to **select** (Section 3.6) the best cluster produced during the procedure that is returned as the best topic for the current topic matrix.

Orthogonalize replaces the singular vectors in the SVD by vectors which are orthogonal to the group profile of the selected cluster. Thereby, the contribution of the identified cluster and its literature profile is removed from the topic matrix.

3.5 Scoring

We need to take special care that weak literature topics with a strong gene-expression pattern are not dominated by strong topics with genes that are not consistently regulated. Correspondingly, we need to avoid forcing regulated members of a class of proteins into a cluster without considering that other members of that class might not be regulated. We aim to suppress such inconsistent regulatory patterns to detect clusters that are more relevant to the condition measured in the experiment. Therefore, we need to distinguish between the cluster of genes and the literature topic that is represented in the cluster. Given a cluster C of genes, the minimum similarity t_{cs} for the genes in the cluster to cp can be determined. We can derive additional genes that belong to the topic by comparing all available genes with the profile. The topic yields a different set of genes which is a superset of the cluster (the hull $H(C)$ of the cluster C). A gene g belongs to $H(C)$ if its similarity to the cluster profile is larger or equal to t_{cs}, i.e. $\langle g, C\rangle \geq t_{cs}$. If a gene is removed from the current cluster, the cluster itself changes and so does the score. Taking this into account would require recomputing the topic with/without the gene and reevalute cluster coherence (for every gene), resulting in an infeasible runtime of ConceptMaker for realistic problem instances.

We avoid this unfavorable complexity by relying on a score based on the current cluster C. This combined score balances three aspects of cluster/topic coherence, the fit to (1) the cluster C, (2) the hull H of the cluster and (3) a vector to shift the cluster away from the hull. To be precise, for a current topic matrix A, we construct three profiles,

$$\begin{aligned}
&\mathbf{cp1} = \mathrm{tgp}(A, C), && \\
&\mathbf{cp2} = \mathrm{tgp}(A, H), && \text{where } H = \{g \in G \backslash C \,|\, \langle g, \mathbf{cp1}\rangle \geq ts\} \\
& && \text{and ts} = \min(g \in C, \langle g, \mathbf{cp1}\rangle), \\
&\mathbf{cp3} = \mathrm{tgp}(A, S(C)), && \text{where } S(C) = \{\Delta(g) * \mathrm{g|g \text{ in } } C\} \\
& && \text{and } \Delta(g) = \langle g, \mathbf{cp1}\rangle - \langle g, \mathbf{cp2}\rangle.
\end{aligned}$$

For the overall score, we compute a weighted average of the topic (literature) score with the cluster (expression) score:

$$\mathrm{score}(M, C, g) = (1 - \alpha) * \mathrm{rpv} + \alpha * (\Delta(g) + \langle g, \mathrm{cp}_3\rangle)$$

The influence of the gene-expression measurements is solely captured by $\mathrm{rpv} = 1 - (r-1)/(n-1)$, where r is the rank in the sorted list of P-values derived from a t-test for the expression data from all n genes. The parameter $0 \leq \alpha \leq 1$ is used to balance the influence of gene-expression measurements against the importance of the literature features. The term $\Delta(g) + \langle g, \mathrm{cp3}\rangle$ exhibits different behaviors depending on the following two cluster scenarios:

(1) The cluster represents a coherent pattern in the regulation and the literature profile cp_1 of the gene cluster (Fig. 1: size $\neq$ 21)
  (a) The genes $g \in C$ are coherent with respect to cp_1.
  (b) Most of the genes that match cp_1 are already contained in the cluster, i.e. $|H|$ is small.

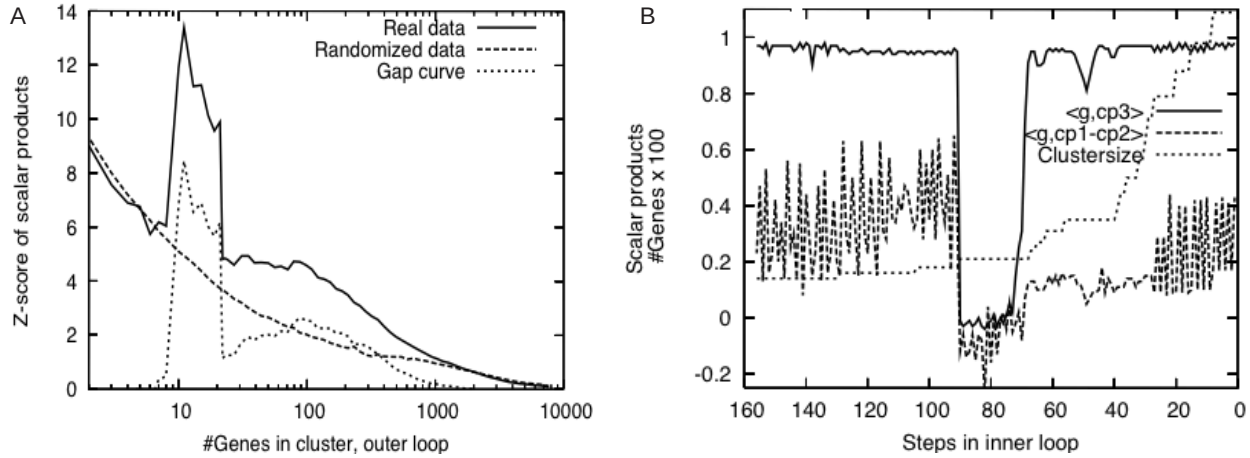


Fig. 1. Concept building and scoring. (**A**) depicts a run of the ConceptBuilder algorithm via an averaged cluster Z-scores $z(g, \text{cp1})$ and a gap estimate (Section 3.6) that compares Z-scores from real and randomized data. The cluster decreases in size as the algorithm proceeds, so the panels are read from right to left. Using real data, the literature topic captured by the profile cp1 changes only a little until a significant shift occurs as the cluster size drops to 21 genes also indicated by the terms shown below. This shift led to a steep increase in coherency of the literature topic and the corresponding Z-score (A). Such a shift becomes necessary if the current concept cannot be improved by removing genes from the cluster. The shift itself is triggered by a changed balance between the two components ($\langle g, \text{cp1–cp2}\rangle$, $\langle g, \text{cp3}\rangle$) of the score in (**B**) in the inner loop. The stepped curve (B) corresponds to the x-axis in (A). Cluster 24 genes: DNA binding, radiation effect, skin, RNA, DNA repair; 21 genes: antigens cd, T-lymphocyte, B-lymphocyte, lymphoma.

(2) The cluster represents an incoherent pattern (Fig. 1: size = 21)
   (a) $|H|$ is large.
   (b) The genes $g \in C$ are regulated but $g \in H$ are not.

In the first case, the cluster and the topic are sufficiently different with regard to the literature features expressed in the respective sets of genes. It follows that $\Delta(g) \neq 0$ and, thus, $\Delta(g)$ will have more weight to select genes than $\langle g, \text{cp3}\rangle$ which is approximately equal to $\langle g, \text{cp1}\rangle \approx 1$ (Fig. 1B). Genes will be selected that fit well to the cluster profile, i.e. the coherence of the literature features will be improved. In the second case, the literature features are about evenly distributed among the genes in the cluster and in the topic. The decisive term will now be $\langle \text{cp}_3, g\rangle$ while $\Delta(g) \approx 0$. Incoherent patterns will be eliminated as cp_3 becomes orthogonal to cp_1, cp_2.

According to our experiments with ConceptMaker, the runtime of the inner loop of the algorithm is independent of the number of genes n. The runtime complexity of the outer loop is bounded by $O(n \log_{1/\text{fraction}} n)$ as compared with $O(n^2)$ for hierarchical clustering. For $n = 8500$ it takes about 60 s for constructing one topic.

3.6 Cluster boundary detection and selection

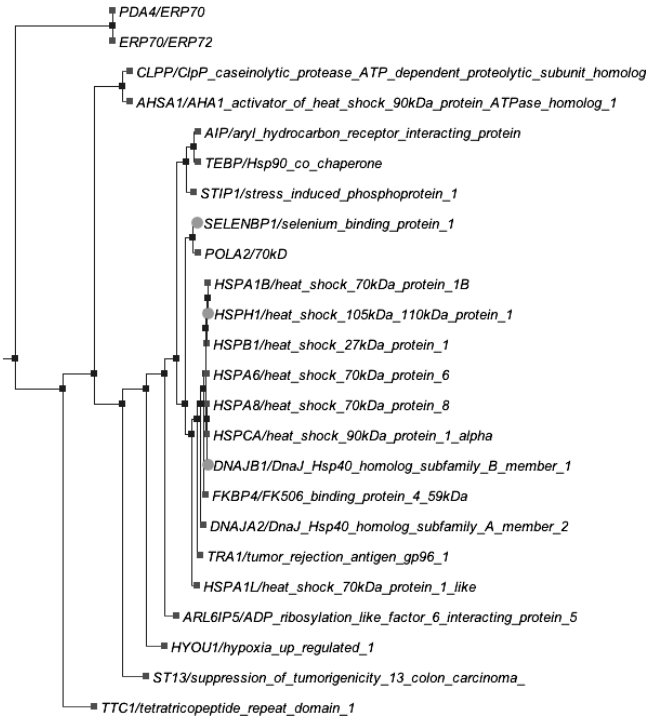
So far, we described an algorithm that constructs a series of clusters of decreasing size until the final cluster contains a single gene. We first extract distinct clusters by evaluating how much the group profile has changed at each step of the algorithm i, i.e. we compute the similarity between successive steps via $\langle \text{cp}_{1,i-1}, \text{cp}_{1,i}\rangle$. During most steps, the profile vector cp_1 typically changes only a little, i.e. this similarity will be close to (1). This corresponds to scenario (1) that has been described above. In scenario (2) we observe that the profile vector gradually shifts during an iteration, usually repeating the inner loop multiple times and exhibiting a similarity <1. If the similarity falls below a given threshold t_{sp}, we take this as evidence that there has been a significant shift in the topic and in the set of genes contained in the cluster. This threshold will typically be set to a value between 0.5 and 0.9. The switch depicted for cluster = 21 genes in Figure 1 occurred with a similarity of 0.04. We need to decide which of the clusters represent the appropriate topic to be selected. To find clusters that represent real patterns and discard clusters with spurious patterns, we compare the similarity of the clusters and group profiles derived from real as well as random data. We create a number of matrices T_k^b with $b = 1, ..., \text{B}$ from the input matrix T_k by random permutation, i.e. we shuffle the weights of the individual singular vectors in T_k (adapted from (Hastie *et al.*, 2000). At each step i we evaluate how well genes fit to the group profile cp_1 of the cluster on average via their respective Z-scores, i.e. $z(g, \text{cp}) = (\langle g, \text{cp}\rangle - \mu)/\sigma$ with the mean μ and the standard deviation σ. We compare this coherency criterion between the value from the real matrix and the averaged value from the randomized matrices via a gap function for each cluster c_i where $|c_i| = |c_i^b|$:

$$\text{Gap}(i) = \frac{1}{|c_i|}\sum_{g \in c_i} z(g, \text{cp}_1) - \frac{1}{B}\sum_b \frac{1}{|c_i^b|}\sum_{g \in c_i^b} z(g, \text{cp}_1^b).$$

We start from the cluster i that produced the largest gap. As clusters with $\text{size}_j > \text{size}_i$ will represent only slight variations of a topic, we examine if the clusters of larger size still satisfy the threshold t_{sp}. We select the largest such cluster and its corresponding profile for orthogonalizing T_k.

3.7 Visualizing clusters: genes, terms and documents

The clusters that have been selected by the above procedure represent groups of genes that share a specific literature profile. From the viewpoint of the available gene-expression data, further partitioning of such sets of genes does not make sense. There might still be interesting substructures within the cluster of genes based on their individual literature context that can be visualized using hierarchical



1.00 heat
0.35 heat-shock protein
0.29 stress
0.26 heat-shock proteins 70
0.18 molecular chaperone
0.17 crystallin
0.11 heat-shock proteins 90
0.10 heat-shock response
0.10 endoplasmic reticulum
0.07 ischemia
0.06 ubiquitin
0.06 chaperonin
0.05 arsenite
0.04 protein denaturation
0.04 adenosine triphosphate

Fig. 2. ConceptMaker identified a heat shock cluster with 24 genes (boxes: upregulated in juvenile arthritis). A fixed number of terms that match the literature profile of the cluster is reported together with a relevancy score (Section 3.7).

clustering. We used complete linkage clustering with an uncentered correlation to compare the genes within the cluster based on their representation in T_k. To provide a summary of the cluster and its group profile we compare all terms with the profile cp_1 in LSI space. We weight the similarity by the inverse document frequency (IDF) of the corresponding term t to penalize frequent terms:

$$\text{term\_score}(t, \text{cp}) = \langle t, \text{cp}\rangle * \text{IDF}(t).$$

We report a fixed number of MeSH terms and/or abstract terms and/or genes that achieved the largest term_score(t, cp_1) (Fig. 2).

This approach is not limited to rank and visualize genes or terms. Analogously, a publication d $\in$ D_k can be weighted and ranked according to a given cluster profile by a similar measure:

$$\text{doc\_score}(d, \text{cp}) = \langle d, \text{cp}\rangle.$$

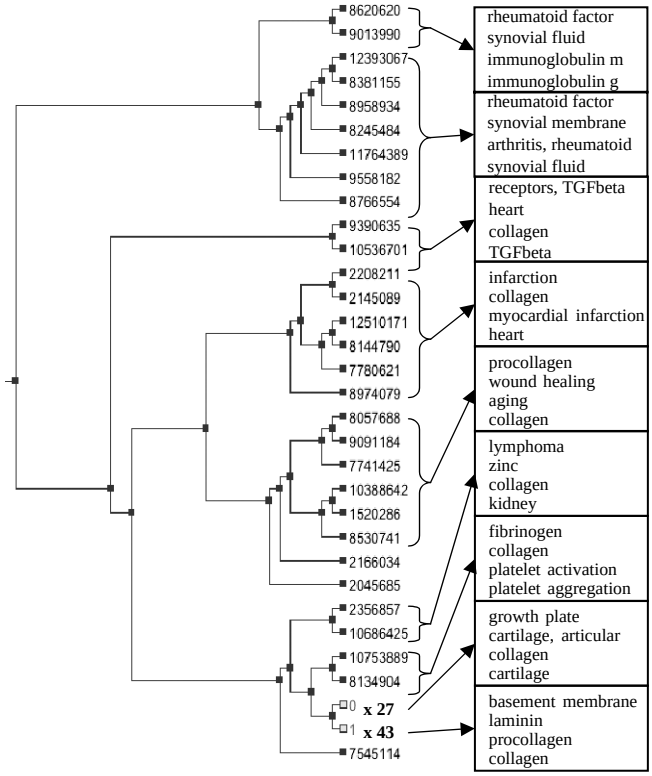


Fig. 3. Literature review of the collagen turnover concept (Section 3.7). We chose 100 publications associated with the profile of the collagen ConceptMaker concept (Table 2). The dendrogram depicts publications, referenced by NCBI PubMedIds and their mutual similarity. The dendrogram has been cut and the resulting clusters (representing subtopics) have been automatically annotated by English words. The clusters 0 and 1 have been collapsed to increase legibility.

If D_k instead of T_k is used we can also construct a dendrogram for documents (Fig. 3) after choosing a fixed number of documents that score well with respect to a given cluster profile. To enable biologists interested in a particular field to quickly screen the literature available for a given topic or cluster, we also compute sub-profiles sp $=$ dgp(A_k, S) for subtrees S in the dendrogram of publications. In turn, these subtopics can be annotated by english terms using term_score(t, sp) (Fig. 3).

4 RESULTS

We evaluate the effectiveness of the method outlined above based on a gene-expression dataset published by Barnes *et al*. (2004). The authors aimed to identify disease-specific gene-expression patterns in cells of juvenile arthritis patients, an autoimmune disease that affects joints in children. They generated microarray data (Affymetrix U95Av2) for 11 healthy controls and 15 patients with active disease from peripheral blood mononuclear cells (PBMC). These data are available via the gene-expression database GEO (http://www.ncbi.nlm.nih.gov/geo) as dataset GDS711.

Presently, two main hypotheses on the cause of juvenile arthritis are discussed in the literature, a viral/bacterial infection or an overactive immune system. So far, the molecular biology of the

Table 2. Cluster comparison/assignment to disease processes

Process	ConceptMaker	#b	Barnes *et al*. (2004): GO category		#a	#b
Immune system	T&B lymphocyte activation/antigens	21	**JAK-STAT cascade**	**GO:0007259**	**23**	**6**
	Chemokine/chemotaxis/eosinophil	25	**Chemoattractant activity**	**GO:0042056**	**23**	**6**
	Neutrophil activation	34	Regulation of cell adhesion	GO:0030155	11	4
	Macrophage activation/lipopolysaccharide	14				
	Bacterial infection/lipopolysaccharide	14				
	Viral infection/hepatitis/influenza	11				
Apoptosis	Apoptosis/caspases/death	12				
Stress response	Oxidative stress/superoxide/glutathione	56	Ubiquitin conjugating enzyme	GO:0004840	33	6
	Heat shock proteins/protein degradation	24	Response to heat	GO:0009408	39	6
	Fatty acid modification/degradation	27	Proteasome endopeptidase activity	GO:0004299	32	5
Extracellular matrix	Matrix formation/collagen turnover	22				
	Bone growth/density	8				
Other	Purinergic receptors/Adenosine	22	Embryonic development	GO:0009790	22	4
	Leukemia	16	Prenylated protein tyrosine phosphatase	GO:0004727	32	5
	Receptor protein tyrosine kinase/Src-family kinase	20	Receptor protein tyrosin phosphatase	GO:0005001	10	3
			Protein tyrosin phosphatase activity	GO:0004725	50	7
			Single stranded RNA binding	GO:0003727	13	4
			Amino acid metabolism	GO:0006520	99	13
			Glutamine metabolism	GO:0006541	9	3
			Amine biosynthesis	GO:0009309	39	7
			Aromatic compound metabolism	GO:0006725	55	8
			Hearing	GO:0007605	28	5
			Amino acid biosynthesis	GO:0008652	28	5
			Perception of sound	GO:0009592	29	5
			Mitochondrial inner membrane	GO:0005743	92	11
			Protein transporting 2-sector ATPase	GO:0016469	14	3
			Aromatic amino acid metabolism	GO:0009072	22	4

#a is the total number genes in the GO category, #b the number of significantly regulated genes. It should be noted that, in contrast to GO categories, in our approach the number of genes associated with a particular concept is not a fixed parameter. The size of the clusters that result from our procedure is determined with respect to the defined optimality criterion. It is possible to increase/decrease the number of genes contained in a cluster with a slight loss/gain on the significance of the regulatory/literature patterns. Barnes *et al*. (2004) claimed that the two categories printed in bold face were particularly disease relevant.

disease is only poorly understood. The main disease relevant biological processes known include immune responses (inflammation, cell mobility, apoptosis), stress responses and extracellular matrix processes (collagen turnover, proteolysis, tissue repair).

We prepared a corpus of 516 000 publication abstracts, the mapping between abstracts, proteins and the spots on the microarray as well as the corresponding singular value decomposition as described in Section 3. This representation is independent of the given gene-expression dataset. The algorithm aims to generate clusters with coherent literature profiles that are significantly regulated. In the following, we will compare the 15 distinct clusters that have been generated by our procedure (out of 20 clusters, redundant clusters have been removed) to the results of Barnes *et al*. (2004) that also aimed to identify functional groups of genes. First, they restricted the genes represented on the microarray to 342 differently expressed genes according to a P-value <0.0001 (t-test). In a second step, they analyzed this set of genes regarding the possible overrepresentation of their respective annotation in the GO hierarchy (Stevens, 2002). They identified 21 GO categories that yielded a P-value better than 0.01 according to a hypergeometric distribution. We compared our findings with the list of GO categories compiled by Barnes *et al*. (Table 2). Here, we used the mappings between Affymetrix HG-U95Av2 and GO categories as provided by Pavlidis *et al*., one of the co-authors of the above paper (http://microarray.genomecenter.columbia.edu/annots). Among the 21 categories, the authors discuss just two overrepresented GO categories as particularly disease relevant, i.e. activation of chemokines and the cytokine receptor pathway (Jak–Stat pathway).

For comparison in terms of overlaps and differences to related GO categories with overrepresentation analysis (ORA) we discuss the cluster of heat shock proteins in more detail (Table 2). In the literature, associations between juvenile arthritis and heat shock proteins have already been discussed (Prakken *et al*., 2002). Although there is no dedicated category on heat shock proteins in GO, it could be described by a combination of closely related categories: 'response to heat', 'response to stress' and 'unfolded protein binding', i.e. detecting significantly regulated genes in all three categories would strongly hint at heat shock proteins as an important category. Response to stress (1075 proteins) and protein folding (163) contain a higher number of proteins and therefore are likely to receive higher P-values compared with response to heat (39 proteins). Consequently, only response to heat has been identified by Barnes *et al*. (2004), with a moderately significant P-value. Different categories cannot be joined, so ORA cannot identify heat shock proteins as highly significant.

We will now focus on the overlaps/differences between our cluster (24 genes) and the response to heat category (39). The intersection between both groups contains nine genes. Out of the 15 genes our method identified additionally, 13 can be explained by examining related GO categories: protein folding (8 genes), response to stress (3) and protein disulfide isomerase activity (2). Out of the two remaining genes, ARL6IP5 (ADP-ribosylation-like factor 6 interacting protein 5), is not annotated within GO and is possibly relevant for the ATPase activity of certain heat shock proteins. The inclusion of the remaining gene, POLA2 (polymerase DNA-directed, alpha), is probably a text mining artefact because this gene is also referred to as 70 kDa (synonym from the LocusLink database), a term likely to overlap with 70 kDa heat shock proteins.

Further differences between the approaches can be illustrated by the inspection of the expression profiles of the identified groupings. Barnes *et al*. (2004) focused their analysis on 342 genes, i.e. the range of about 4% most significantly regulated genes. This restriction led to the detection of 6 of 39 genes related to response to heat (Table 2). Our approach does not use such fixed thresholds and will, therefore, generate clusters that also include moderately regulated genes if they match the overall literature profile of the cluster.

The profile from our heat shock proteins cluster (24 proteins) contained seven proteins from the (0–4%) range of most significantly regulated genes, 8 (4–10%) and 9 (10–25%).

Genes from the reference GO category, response to heat, not included in our cluster can be explained as follows. First, many of these genes are not significantly regulated. Second, genes from the Pavlidis GO category are not reported by ConceptMaker if they do not match the literature profile of the cluster. There are six additional genes in the GO category, which are regulated [within (0–25%)] but not reported in our cluster that deserve a closer examination. Three of them are not annotated within response to heat by the official GO-annotation group and may represent possible mapping errors, i.e. KRN1 (keratin, cuticle, ultrahigh sulphur 1), STK3 (serine/threonine kinase 3) and TGFB1I1 (TGFB1 induced transcript 1). The remaining proteins (TRAL1, TRAP1, GroEL) exhibited only weak regulatory patterns [(20–25)% range]. Other functions besides heat shock have been attributed to these proteins i.e. their literature profile was rather weak with respect to the category heat shock proteins.

To facilitate a more in-depth analysis of the generated clusters we compiled a literature review, i.e. a dendrogram of publications that fit the profile of individual clusters. Figure 3 depicts a literature compilation for the collagen turnover concept. The approach is analogous to the ranking and visualization of terms and genes (Section 3.7). By cutting the dendrogram we can identify subtopics that, in turn, are annotated by descriptive English terms. Thereby, researchers can quickly zoom in to find publications according to their interest. As juvenile arthritis is an autoimmune disease involving extracellular matrix degradation, collagen-related processes are relevant. Guided by the annotation, we focus on the top cluster in Figure 3 to learn about the autoimmune aspect. The associated papers discuss the binding of autoantibodies to collagens and fibronectin and discuss their involvement in collagen degradation.

5 DISCUSSION

We presented an approach to generate concepts, e.g disease hypotheses and clusters of genes relevant to these hypotheses, by an integrated analysis of gene-expression measurements together with biomedical publication abstracts. Thereby, the ConceptMaker algorithm overcomes limitations of methods based on gene-expression information alone as well as approaches that employ manually compiled knowledge resources. The algorithm additionally annotates clusters using English terms derived from the abstract text and/or the corresponding MeSH annotation (Fig. 2). A graphical tree representation visualizes the resulting gene sets (Fig. 2). There is no need to manually compile controlled vocabularies (except for predefined synonym lists of gene/protein names) or pathway databases prior to applying our algorithm. The hypotheses and associated gene sets assembled by our algorithm are tailor-made for the given gene-expression measurements to be interpreted. To derive annotations or literature reviews, topics can be converted into and scored against terms, genes and documents (Fig. 3).

We applied our algorithm to a current gene-expression dataset on juvenile arthritis (Barnes *et al*., 2004) and compared our results with an ORA against GO categories performed by the authors of the study. Barnes *et al*. (2004) were able to match 2 of the 21 identified categories to biological processes relevant to the disease. Many of the categories identified by ORA were too unspecific, such as amino acid metabolism or mitochondrial inner membrane. Some categories were not obviously related to juvenile arthritis, such as embryonic development (Table 2). Moreover, processes located in the extracellular matrix or apoptosis could not be detected by ORA. In contrast to this, the ConceptMaker approach introduced in this paper detected 12 (out of 15) categories that are quite specific disease relevant processes related to the immune system, apoptosis, stress response and extracellular matrix turnover, and to identify the corresponding relevant genes in the data (Table 2). As our approach utilizes the entire set of genes for which expression data are available (in contrast, ORA utilizes only the small fraction of highly regulated genes), the identified topics are also more sensitive with respect to the associated gene clusters as genes can be included into these clusters owing to the topic score even if they exhibit only moderate significance in the expression measurement. Thus, the resulting topics and clusters can be expected to be a good starting point for further hypotheses refinement and experimental research.

ACKNOWLEDGEMENTS

This work has been partially funded by the German Ministry of Research (Project BOA, Leitprojekt Osteoarthritis, grant 01GG9824).

Conflict of Interest: none declared.

REFERENCES

Adryan,B. and Schuh,R. (2004) Gene-Ontology-based clustering of gene expression data. *Bioinformatics*, **20**, 2851–2852.

Barnes,M.G. *et al*. (2004) Gene expression in juvenile arthritis and spondyloarthropathy: pro-angiogenic ELR+ chemokine genes relate to course of arthritis. *Rheumatology*, **43**, 973–979.

Blaschke,C. *et al*. (2001) Mining functional information associated with expression arrays. *Funct. Integr. Genomics*, **1**, 256–268.

Chaussabel,D. and Sher,A. (2002) Mining microarray expression data by literature profiling. *Genome Biol*., **3**, R55.

Cheng,J. *et al*. (2004) NetAffx Gene Ontology mining tool: a visual approach for microarray data analysis. *Bioinformatics*, **20**, 1462–1463.

Deerwester,S. *et al*. (1990) Indexing by Latent Semantic Analysis. *JASIS* **41**, 391–407.

Doniger,S.W. *et al*. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol.*, **4**, R7.

Fundel,K. *et al*. (2005) Exact versus approximate string matching for protein name identification. *BMC Bioinf.*, **6**, S15.

Gibbons,F. and Roth,F. (2002) Judging the quality of gene-expression based clustering methods using gene annotation. *Genome Res.*, **12**, 1574–1581.

Glenisson,P. *et al*. (2003) Meta-clustering of gene expression data and literature-based information. *SIGKDD Explor. Newslett.*, **5**, 101–112.

Hanisch,D. *et al*. (2005) ProMiner: Organism-specific protein name detection using approximate string matching. *BMC Bioinformatics*, **6**, S14.

Hanisch,D. *et al*. (2002) Co-clustering of biological networks and gene expression data. *Bioinformatics*, **18**(Suppl 1), S145–S154.

Harris,M.A. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.

Hastie,T. *et al*. (2000) Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.*, **1**, R3.

Hu,X. (2004) Integration of Cluster Ensemble and Text Summarization for Gene Expression Analysis. BIBE 2004.

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

Jiang,D. and Zhang,A. (2002) A cluster analysis for gene expression data: a survey. *IEEE Transactions on Knowledge and Data Engineering*.

Kim,W. *et al*. (2001) Automatic MeSH term assignment and quality assessment. *AMIA S*, 319–323.

Krull,M. *et al*. (2003) TRANSPATH: an integrated database on signal transduction and a tool for array analysis. *Nucleic Acids Res.*, **31**, 97–100.

Liu,Y. *et al*. (2004) Text mining functional keywords associated with genes. *Medinfo*, 292–296.

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

Nelson,S.J., Johnston,D., Humphreys,B.L. (2001) Relationships in Medical Subject Headings. *Relationships in the organization of knowledge*. C. A. G. Bean, Rebecca. New York, Kluwer Academic Publishers, 171–184.

Prakken,B. *et al*. (2002) Heat shock proteins in juvenile idiopathic arthritis: keys for understanding remitting arthritis and candidate antigens for immune therapy. *Curr. Rheumatol. Rep.*, **4**, 466–473.

Sebastiani,F. (2002) Machine Learning in Atomated Text Categorization. *ACM Comp Surveys*, **34**, 1–47.

Shatkay,H. *et al*. (2000) Genes, themes, microarrays using information retrieval for large-scale gene analysis. *ISMB*, **8**, 317–328.

Sohler,F. *et al*. (2004) New methods for joint analysis of biological networks and expression data. *Bioinformatics*, **20**, 1517–1521.

Speed,T.P. (2003) Statistical Analysis of Gene Expression Microarray Data. Boca Raton, FL, Chapman & Hall/CRC.

Stevens,R. (2002) Ontology based document enrichment in bioinformatics. *Comput. and Funct. Genomics*, **3**, 42–46.

Wingender,E. (2004) TRANSFAC, TRANSPATH and CYTOMER as starting points for an ontology of regulatory networks. *In Silico. Biol.*, **4**, 55–61.

Xiang,Z. *et al*. (2003) Microarray expression profiling: analysis and applications. *Curr. Opin. Drug. Discov. Devel.*, **6**, 384–395.

Zien,A. *et al*. (2000) Analysis of gene expression data with pathway scores. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 407–417.