Term-based literature mining across systems biology and biomedicine

Bert Coessens, Steven Van Vooren, Patrick Glenisson, Yves Moreau, and Bart De Moor

Address: Departement Elektrotechniek (ESAT), Faculteit Toegepaste Wetenschappen, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium

Tel.: +32 16 32 86 54; Fax: +32 16 32 19 07; E-mail: bcoessen@esat.kuleuven.ac.be

Correspondence: Bert Coessens, Kasteelpark Arenberg 10, 3001 Leuven

Abstract

In this paper we describe a term-based text-mining framework that allows a better integration of existing knowledge on the one hand and statistical models on the other. Three case studies are presented, ranging from systems biology over biomedicine to systems biomedicine, to show the power of text mining to integrate written information and natural language communications with wet-lab analysis results. The first case focuses on the term- and gene-centric textual analysis of groups of genes derived from expression experiments. In the second case we highlight how a text-mining approach can be adopted to link phenotype and genotype information in a clinical context. The third case describes how textual information can be used to screen chromosomal regions for candidate disease genes. The three cases clearly show the benefits of an integrated text-mining approach.

1 Introduction

The availability of the complete sequence of the human genome and that of several other organisms has sparked a novel research paradigm in the life sciences. The transition in biology from rather reductionist approaches to more integrative thinking draws mainly on the recent advances in high-throughput screening technologies. One of the most prominent examples of such breakthroughs are microarrays or gene chips. They measure the simultaneous activity of thousands of genes in a particular condition at a given time and enable researchers to identify potential genes involved in a great variety of biological processes or disease-related phenomena. In turn this has lead to the massive generation of novel hypotheses on the function of multiple genes, their interrelatedness, and the controlled circumstances in which the ensuing observations hold.

The shift in focus – from the characterization of individual molecular components to the study of their complex and orchestrated interplay within the cellular environment – requires methodologies that integrate different data and information sources at the level of genes, proteins, diseases, organisms, etc.

Scientific literature is more and more an essential resource in this integration process and ways to search and navigate it are indispensable. It is therefore not surprising that, over the past few years, the number of text mining applications that attempt overcoming some limitations of popular retrieval systems such as PUBMED, is on a sharp rise [1]. In this paper we demonstrate how a single literature mining framework can be applied to several biological and biomedical inquiries.

Through a variety of collaborations the Bioinformatics group of K.U. Leuven implements the recently suggested requirements by Hood *et al.* [13] for conducting effective systems biology research:

[1] Over a third of the current publications matching 'text mining' in PUBMED is dated between Jan 1st and Oct 31 2004.

with a team of cross-disciplinary scientists and through multiple projects we are exposed to various aspects of molecular biology and biomedicine, which enabled us to deploy our term-based text mining framework in a variety of research environments.

In this work we apply our framework in three fields that increasingly rely on 'systemic' computational approaches. Adopting a systemic viewpoint to complex problems implies the integration of various sources of data and information into an analysis pipeline. Citing Palsson [20]: 'Such systems analysis is often viewed as comprising four principal steps:

1. Enumerate as complete a list as possible of the biochemical components that participate in the process of interest.

2. Study the interactions between these components, a process referred to as network reconstruction

3. Describe the properties of the reconstructed network mathematically

4. *In silico* In silico' models can then be interrogated to analyze, interpret and predict the biological functions that can arise from reconstructed networks'

The contribution of semi-automated processing of literature information can be situated in the first step (assignment of function to components), the first phase of the second step (i.e., characterization or discovery of a variety relations from literature), and the final step (i.e., corroboration of model predictions according to literature). In our view, the knowledge discovery process should essentially be considered as cyclic, i.e. requiring one or more iterations between heterogeneous information sources in order to extract reliable hypotheses. In Figure 1 we translate this premise to the analysis cycle we deploy in three illustrative case studies. Each one adopts a different view on the 'systematization' of post-genome biology.

The first case aims at supporting data mining in a systems biology context. In microarray studies that focus on unraveling the functions of genes [27, 14] and their interrelations, linking up experimentally derived gene groups (from cluster analysis) to the existing databases and published literature still requires numerous queries and extensive user intervention. This process of drilling down into database entries of hundreds of genes is notably inefficient and requires higher-level views that can more easily be captured by a (non-)expert's mind. Figure 1 shows how a gene group derived from expression data ('Data World') is used to scan for existing information ('Text World') to yield an improved characterization and understanding. We will show how to involve textual annotation and MEDLINE abstracts in the knowledge discovery process.

Whereas microarrays query expression levels of genes, genomic hybridization arrays (CGH arrays), which are an extension of the microarray hybridization technology, are able to scrutinize variations within the DNA such as amplifications and deletions in full genome information [15]. Hence, they are of direct relevance to a clinical research environment. In our second case, the text mining framework we present is deployed to put human (full-)genome information, experimental results, and patient records into a single loop to find novel markers that can be used in prenatal screening of congenital deficiencies.

This type of biomedical inquiries are likely to undergo the systematization as patient records and disease information will be increasingly connected to molecular biology and genetic information. Our third case is in this respect an example of how the systems approach meets biomedicine as we show how to prioritize disease genes in a give chromosomal region based on the literature. The aim in this final example is to use the literature semi-automatically to produce a list of candidate genes that can act as an enhanced priority list for *in vitro* validation studies.

2 Term-based literature mining framework

2.1 Document processing

The construction of a literature database starts with the collection of a set of documents in ASCII format. These documents might be MEDLINE abstracts, patient reports, database annotations, etc. The document corpus is typically processed by removing punctuation, case, and document structure. For this reason the vector space representation is often referred to as *bag-of-words*.

Standard English *stemming* (Porter's method [22]) canonizes the words according to morphological and inflectional endings (for instance plurals, tenses, ...) and helps to reduce to a certain extent the dimensionality as well as the dependency between terms. Removing words including articles, prepositions, and conjunctions is desirable to reduce noise and is done using term statistics (see below) and/or with a predefined *stop word* list. *Phrases* are terms consisting of several tokens (for instance 'amino acid transport') and although little is known on whether and how they affect the performance of learning text (for instance document classification/retrieval), they are important when extracting comprehensive information (for instance document summarization). *Synonyms* are different terms conveying the same meaning or referring to the same object (for instance 'tumour' and 'tumor', 'P53' and 'TP53'). *Polysemy* refers to terms conveying different meanings according to the context it appears in (for instance 'CD' as compact disk, Crohn's disease, cytosine deaminase, etc.).

The idiosyncrasies of biomedical text pose particular challenges to the implementation of the concepts introduced above. Biological nomenclatures for example can contain comma's, dots, hyphens, primes, parentheses, brackets, ... and can be case-sensitive (e.g., '1-chloro-2,2-bis(4'-chlorophenyl)ethane dehydrogenase'). Moreover, inspired gene names such as 'CELL', 'AND' or

'ALPHA' make accurate extraction of crucial biological entities such as genes and proteins a daunting task. Conversely, applying procedures such as stemming can modify harmless words such as 'may' (if not appearing in the list with stop words) into gene names ('MAI') or vice versa. Such issues can seriously trouble the outcome when looking for co-cited genes in literature. We deal with these issues in a pragmatic way by manually processing the domain vocabularies. The detection of gene names [9] and relationships between terms [2] remain important research topics.

2.2 Construction of text profiles

The representation called the vector space model encodes a document in a k-dimensional term space where each component w_{ij} represents the weight of term t_j in document d_i (Figure 2). Hence, each point i in this vector space is a row vector containing weights for each term j. We call this a *text-* or *term-profile*.

As explained before, any structure in the text is thus obliterated. When all terms are assigned equal importance, binary weights are adopted:

$$w_{ij}^{\text{BOOL}} = \begin{cases} 1 & \text{if } t_j \in d_i \\ 0 & \text{otherwise} \end{cases}$$

which is referred to as Boolean weighting. Slightly more elaborated is the TF-IDF weighing scheme, which allows for a partial matching of corresponding terms and is defined as follows:

$$w_{ij}^{\text{TF-IDF}} = f_{ij} \log(\frac{N}{n_i}),$$

where f_{ij} is the number of occurrences of t_j in d_i and is called term frequency (TF). It accounts for the assumption that terms occurring often within a document are more important. The logarithmic term on the other hand is called inverse document frequency (IDF), where N represents the total

number of documents and n_i is the number of documents containing term i in the collection. It proportionally downweights terms that often recur throughout the whole document set.

2.3 Document similarity

Similarity between pairs of documents d_1 and d_2 is expressed by the cosine of the angle between the corresponding vector representations:

$$\mathrm{Sim}(d_1, d_2) = \cos(d_1, d_2) = \frac{\sum_j w_{1j} w_{2j}}{\sqrt{\sum_j w_{1j}^2}\sqrt{\sum_j w_{2j}^2}}.$$

Since $w_{ij} \geq 0$, for all i, the degree of similarity $\mathrm{Sim}(\cdot, \cdot) \in [0, 1]$. The underlying hypothesis states that high similarity between documents corroborates a strong semantic connection between them.

2.4 Construction of an entity index

Depending on the research issue at hand the basic elements of reasoning will differ. In our first and third study for example, genes constitute the elementary concepts as we analyze a cluster of co-expressed genes and a genomic disease region. Conversely, in the second case we aim at linking genomic and patient information, so our information in the text world now is patient- as well as gene-centered.

Hence, prior mappings between objects of reasoning and textual sources need to be constructed. Gene-literature associations for example, are already available in several forms:

- Hits from PUBMED or other search engines [26] when entering terms related to a gene;
- Gene-to-literature mappings in curated repositories such as SGD, GO, LocusLink;
- Gene-to-literature links constructed by computational classification techniques [23].

After computing a vector-based literature index for each document in a given collection, we can combine for each gene/patient the text indices of all documents associated to it.

The text index of a gene i is then a vector of terms j obtained by taking a (possibly weighted) average over the N_i indexed documents to which it is linked:

$$g_i = \{g_i\}_j = \{\frac{1}{N_i}\sum_{k=1}^{N_i} a_k w_{kj}\}_j, \tag{1}$$

where a_k represents a weight factor that can translate the relevance or quality of the document annotation, if available. In this work we ignore this parameter.

Equation 1 pools the keyword information contained in all documents related to a gene into a single term vector. As a result, documents describing the same gene and containing different but related terms are joined.

3 Case 1: Systems biology and microarrays

Microarray experiments allow researchers to measure gene expression of thousands of genes at once, resulting in vast amounts of raw data. After analysing the data, the experiment results need to be validated and interpreted in the context of the existing knowledge. Since most knowledge resides in textual formats, a text mining approach is evident.

Our methodology can support gene expression analysis by summarizing curated free-text information about genes in the form of textual profiles. The methodology allows recovery of general functional terms shared by a given group of genes and can assist in the interpretation of gene groupings stemming from advanced clustering methods. The retrieved terms can then be used to construct subsequent queries to the PUBMED interface. This approach can considerably speed up the analysis by facilitating retrieval of highly relevant information as we will demonstrate in the

example below.

As gene expression only makes sense if the corresponding cellular conditions are well-characterized, we tested whether our method is capable of discriminating between different states of a given pathway. To this end, we use the analysis performed by Ihmels *et al.* [14] where the authors use their signature method to discover expression modules[2]. They identify two subparts of the yeast TCA (tricarboxylic acid) cycle that are autonomously regulated in different cellular contexts (see Figure 3).

We profiled these two subsets and examined how our textual profiles differentiate the two subnetworks by examining both term- and gene-centric summaries (see Section 2.4). The first module constitutes four genes that Ihmels and colleagues reported to be 'upstream of alpha-ketoglutarate, a primary precursor of glutamate and coregulated under experimental conditions of deletion of RTG1 and genes that affect mitochondrial function'. In Table 1 we print the term-centric profile of the first transcriptional module. 'Glutamate' is a module-specific term whereas 'isocitrate dehydrogenase' and 'TCA cycle' are also ranked highly when profiling the entire pathway. Using these three terms to query PUBMED we are able to retrieve three highly specific abstracts (see Table 1) related to reduction in respiration capacity (or hypoxia). We note that the reference PMID:10490611 is exactly the one used by Ihmels *et al.* to support their results.

Under a different set of conditions the authors found a module containing 'the genes upstream of alpha-ketoglutarate coexpressed with genes whose expression is dependent on the transscriptional activator CAT8, which suggests they are involved in gluconeogenesis'. Profiles for the second transcriptional module are also displayed in Table 1. Again, although terms as 'carbon' and 'succinate dehydrogenase' are shared with profiles for the more general TCA pathway, they rank particularly

[2]A module is defined here as a set of genes that are jointly active under very specific conditions.

high in this cluster and using them as building blocks for a PUBMED query in combination with 'glyoxylate cycle' or 'gluconeogenesis' we are again able to retrieve very specific MEDLINE abstracts (Table 1). Interestingly, the abstract PMID:10328823 contains references to gene ACN9, which also scores high in our gene-centric profile (cfr. infra) and is new with respect to the analysis by Ihmels. The abstract PMID:11514507 refers to the role of CAT8 in this module and confirms the manual analysis of Ihmels and colleagues. This implies our framework is a powerful alternative to extensive manual analyses in PUBMED.

As genes take a central position in molecular biology they can be regarded as one of the building blocks in biological reasoning. Contrary to regular terms, which, in our context, confer either a general or specific functional connotation, genes (or, more specifically, gene names) convey information in a more 'compact' way. Indeed, they often function as primary keys between many public genome databases. We start from the hypothesis that genes are often cited by their official symbol and look for simultaneous presence in the *pool* of abstracts that are linked to the group of genes under consideration. As gene interaction networks are known to be scale-free[3] [11] and gene citation is sensitive to scientific *trends* [10], we can expect a handful of genes to dominantly occur in this kind of analyses. We interpret such hub-genes as less informative and therefore correct partly for their invasiveness by applying, as before, the IDF weighting scheme to the gene-centric index.

In Figure 4 we show the co-linkage network for module 2 from Ihmels *et al.*. The query is composed of the module's 15 genes and results in 212 nodes and 457 edges. The idiosyncrasy of the gene network (i.e. lots of genes with few connections and few genes with lots of connections) troubles a transparent analysis. Our method proposes candidate genes of interest by assigning a high score

[3]A scale-free network is a specific kind of network in which the distribution of connectivity is extremely uneven. Instead, in scale-free networks some nodes act as intensively connected hubs using a power-law distribution, $\text{Prob}(\text{nr_connections} > x) = x^{-\gamma}$, with $\gamma \geq 0$.

to genes that are co-linked multiple times with the genes from module 2. In Table 2 we list the top 30 genes ranked by our system. Three genes from the original list are not recovered because of their small weights.

Contrary to a 'whole network' analysis, our approach comprehensibly summarizes, in a first approximation, the most prevalent gene links from literature. CAT8 for example is a transcriptional regulator suggesting the implication of the member genes of module 2 in gluconeogenesis. This is also indicated by Ihmels and colleagues as well as our term-based MEDLINE analysis. ACN9 is a weakly understood gene (3 curated abstracts) involved in gluconeogenesis that also appeared in the course of our term-based analysis and constitutes a good candidate gene for further analysis. Indeed, when submitting ACN9 jointly with the genes from module 2 to SGD's *Expression Connection*[4], we find expression correlation (Pearson correlation ≥ 0.8) of this gene with KGD1 and PCK1 in an experiment describing the evolution of expression during glucose limitation [6] (see Figure 5). We note that the accompanying paper was not present as a literature reference for ACN9 nor was this gene addressed in this paper according to SGD curators.

Finally, we see some genes from module 1 (marked with an asterisk) present in profiles in Figure 2, indicating a difficulty to retrieve context-specific genes from a module. There are two possible reasons for this: either our term-inspired method to detect subnetworks is simple and leaves room for improvement [24, 16] or the analyzed conditioned gene relations are weakly, or not, reported in literature.

As a conclusion we state that our ranking of co-linked genes via a gene-centric vocabulary can offer useful clues to interpret and further unravel sub-networks from expression clusters. Moreover, the gene-centric approach is found to be complementary to the term-based analysis as terms refer more

[4]Expression Connection allows for searching gene expression data from multiple yeast microarray studies.

to candidate functional characteristics, while co-linkage analysis infers hypothetical relations to the investigated gene group. The putative relation of ACN9 to module 2 is also retrieved in a particular expression experiment. This supports our cyclic views on the analysis process (see also Figure 1), where modules or clusters from the data world are analyzed in the text world and subsequently yield candidates that can be confirmed again in the data world.

4 Case 2: Biomedicine and CGH arrays

Recently, DNA microarrays have been introduced into the clinical practice to detect genomic abnormalities in the human genome that are not visible by classic karyotype analysis (chromosome banding) or classic Comparative Genomic Hybridization (CGH). This novel technique, ArrayCGH [15], allows screening of the patient genome for deletions and duplications well below the resolution of classical karyotyping as it is able to pick up micro-amplifications and deletions that span only a few hundred kilobases. As micro-deletions and duplications are an important cause for dysmorphology and developmental delay in humans, ArrayCGH can boost diagnostics and functional genome annotation. Genome annotation implies linking phenotype and genotype information, which involves the identification of candidate genes that are in some way involved in a disease phenotype such as dysmorphology or a deviation in the developmental process.

Text mining proves to be an interesting approach to link phenotype and genotype information in this context. In this section, we propose a text mining technique based on three sources of information. A first source is raw or statistically analyzed data from ArrayCGH-microarray experiments. Data from these experiments is used to provide a high resolution delineation of the chromosomal aberration in a patient. A second information source is a set of informative and complete case descriptions on patients that display genomic micro-deletions and micro-duplications. Specifically,

a list of multi-word terms describing the patient phenotype or a piece of natural language text containing a patient description is used to construct a phenotypical profile. A last source of information is the existing knowledge on genes, phenotypical traits, and diseases or syndromes embedded in the biomedical literature.

Aggregation of patient reports that have detailed descriptions of aberrant regions as well as informative and correct phenotypical descriptions allows the construction of a cytogenetic database that is amenable to search, data- and literature mining. Here, we will apply our text mining framework to support the functional annotation of the human genome.

Using the vector space model on both case reports and PUBMED abstracts, as described in the methods section, we can mine for patients showing similar phenotypes, similar genomic aberrations, and related literature. We indexed the available literature and patient descriptions again by means of controlled vocabularies, as we did in the section above. In this case, the tailor-made word lists were disease- and gene centric.

In a first application, we applied a dysmorphology domain vocabulary. This word list was derived from the London Dysmorphology Database (LDDB). The LDDB is a database of over 3000 non-chromosomal, multiple congenital anomaly syndromes that is used both as an aid to diagnosis for the clinician and as a reference source. It is a hierarchically organized set of dysmorphology related concepts which we have extended manually with synonyms and related concepts.

We retrieved case reports that are related to a patient of interest from a phenotype point of view by constructing a patient ranking based on cosine scores between the dysmorphology profiles of all patients. These dysmorphology profiles are nothing more than the case reports' text vectors corresponding to the LDDB domain vocabulary.

In Table 3, we illustrate how the dysmorphology profile of a patient is constructed by extracting

phenotypic traits from a chunk of text that a clinical researcher or geneticist has entered into the database. This text is usually the patient description that is sent to the MD requesting the array experiment, or a part of a manuscript that is being prepared on the case, or a list of keywords deemed relevant by the medical geneticist who entered the patient into our cytogenetic database. We extracted dysmorphology terms from this information in an intelligent way, not just by term matching. For instance, the dysmorphology term 'speech defect' is triggered when any of these terms occur: articulation defect, donald duck speech, dysarthria, no active speech, slurring speech, speech defect, staccato speech, stutter, speech delay.

To further illustrate the benefits of this approach, we also constructed text vectors for biomedical literature. Again, we used tailored domain vocabularies. For this, we started from a selection of about 8 million biomedical abstracts that are provided through PUBMED. Using a dysmorphology-centric vocabulary, we constructed dysmorphology-profiles for papers and case reports available through PUBMED and linked them to related patients, again using a cosine score ranking.

To illustrate this feature, we added a case report to the system describing a patient that has Rosai Dorfman disease. We then extracted and stored a dysmorphology profile for the patient. Using this profile, we scored the available biomedical literature. The top ranking papers indeed described 'sinus histiocytosis with massive lymphadenopathy (SHML), also called Rosai Dorfman disease'.

Next, we wanted to identify candidate genes for the phenotypic features linked to a group of cases of interest by identifying genes they share in aberrant genomic regions. For this, a second domain specific vocabulary was composed, like in Section 3, of human gene names. Again, the available biomedical literature was indexed using this tailored word list, thus generating gene vectors for the available biomedical literature, allowing retrieval from a gene perspective.

Here, gene vectors for patient case reports were constructed by considering the associated genomic

region that is duplicated or deleted in the patient and scanning it for (putative) genes. The standardized GO names of these genes were then used to construct a gene vector linked to the patient, again allowing the calculation of cosine similarities. Using this approach, we found patients that share affected gene copy numbers, and retrieved relevant literature discussing these genes.

We have shown how text mining can contribute to the annotation of the human genome: combining micro-deletion and duplication case reports on patients that show congenital defects and dysmorphology with text mining features supports the construction of phenotypic genome maps and thus enables the identification of genes involved in developmental processes and constitutional anomalies and will delineate novel clinically recognizable entities.

5 Case 3: Towards systems biomedicine – literature-based gene prioritization

In the post-genomic era researchers are often confronted with selecting the most promising genes from a large list of candidates for further analysis in the lab. In the case of complex multigenic diseases linkage analysis results in the identification of several large genomic loci comprising hundreds of candidate disease genes, and manual validation of all of them is often not feasible.

Several methods have been described to prioritize a list of disease-related genes. However, only Perez-Iratxeta *et al.* [21] adopt a text mining approach. They use literature abstracts from MEDLINE and annotations from LocusLink to calculate fuzzy relationships between Gene Ontology (GO) terms and pathological conditions (diseases). Genes are prioritized according to the strength of the relations between their GO annotations and the disease terms present in the abstracts annotated to them. The relationships are fuzzy because they are inferred indirectly via co-occurrence

of disease terms with chemical terms.

We present a method that uses complete textual profiles of MEDLINE abstracts instead of term co-occurrence information. It can be expected that two genes involved in the same pathology have a comparable textual profile. Therefore, the cosine-based distance between the weight vectors of their profiles is a valid measure for the similarity between them.

We created several textual profiles for every genetic locus in LocusLink and repeated this for different domain vocabularies. Again, the vocabularies define which aspect of the literature describing these loci is emphasized. They allow us to filter irrelevant information and focus on the information germane to the query at hand. In the example below we compare the performance of two vocabularies: one consisting of terms and phrases extracted from GO, and one comprising pathology terms and phrases derived from OMIM's Morbid Map.

Another advantage of working with textual profiles is the straightforward way we can construct disease models. We assume that the literature about a set of genes known to be involved in a certain pathological condition contains all the relevant information and that the average textual profile of all profiles of these genes is a sound model for this pathology. Once a model is built this way, we can prioritize a list of genes by measuring the cosine similarity between the weight vectors of all gene profiles and the weight vector of the model.

The Charcot-Marie-Tooth (CMT) disease will be used to exemplify our approach. We retrieved all genes known to be involved in the disease from OMIM's Gene Map, resulting in a training set of 16 genes. Our model consists of the average textual profile of all gene profiles.

We performed a leave-one-out cross-validation to establish the performance of our method. 16 different models were constructed by removing every gene once from the training set. For every model we recorded how high the removed gene was ranked when it was prioritized together with 49

randomly selected genes. This information can be visualized in a Rank ROC (Receiver Operating Characteristic) curve [1], plotting the proportion of true positives or sensitivity (i.e. the number of times the gene left out is ranked above the cut-off) against one minus the proportion of true negatives or (1-specificity) (i.e. all genes below the cut-off except for the gene left out):

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}.$$

The Rank ROC curve of the CMT case is shown in Figure 6. The area under the curve (AUC) is 0.84 when prioritizing on the textual profiles created with the 'GO' domain vocabulary and 0.83 when using the 'OMIM' profiles (a perfect classification has an AUC of 1). The Rank ROC curve for a model comprising 20 randomly selected genes is clearly worse than that of our Charcot-Marie-Tooth model.

Although statistically sound, we wanted to know how our approach would rank previously unidentified genes. To this end, we queried the Entrez Gene database with 'Charcot-Marie-Tooth' and found two genes (HOXD10 and MFN2) that are only recently identified to be involved in the disease [25, 28]. We note that the abstract texts of the papers describing this link were not yet present in the indices and that no occurences of 'Charcot-Marie-Tooth' were found in those that were used to create the textual profiles of HOXD10 and MFN2. We then retrieved all known genes located in the same chromosomal band as HOXD10 and MFN2 (2q31.1 and 1p36.22, respectively) using the EnsMart system [17], resulting in a group of 62 genes for the HOXD10 band and a group of 48 genes for the MFN2 band. After prioritizing these groups with the CMT model and using the 'GO' or 'OMIM' textual profiles, HOXD10 is ranked 14th and 7th of 62 genes, MFN2 13th and

19th of 48 genes, in both cases higher than randomly expected.

We thus conclude that a straightforward approach of using textual profiles to prioritize lists of genes can already yield promising and biologically meaningful results.

6 Discussion

As contemporary biology is partly evolving towards an information science, integrative views on biological problems will be of increasing importance. Also, published literature has become one of the key factors in such systemic approaches to biology. Through three representative case studies we have shown how the term-based text mining framework we present, provides added value when integrated with post-genome inquiries.

Integration, however, is a broad term and is understood differently by biologists, database people and statisticians. The case studies illustrate two viewpoints:

- **Connection of information sources**: Originally, the creation of text indices was driven by a cyclic view on the knowledge discovery process, where researchers iterate between their data and the literature. Through the creation of various indices on text-oriented databases (the annotation database LocusLink and OMIM, the literature repository MEDLINE or patient records from a university hospital database) we enabled a text-based analysis of groups of DNA sequences that are reported from genome-based pedigree analysis, CGH assays or microarray results. We leveraged the idea of mere term analysis by basing ourselves on annotation standards, nomenclature conventions and taxonomies, and hereby included several *views* on the literature. The underlying rationale is that relevant keywords, phrases, or gene names are only useful to a researcher if they can be linked (back) to existing biological resources.

- **Statistically sound integration of data**: In a bag-of-words approach part of the relations are obliterated. Therefore a term-based approach to summarize complex information can be considered naive. Several results, however, support the use of 'literature data', where wielding numerical representations of text can give new insights into biological problems through for example the application of cluster analysis techniques [7, 4, 24]. The advantage of such an approach is that literature can be closely associated with the data analysis process and combining information can be done in ways that are statistically better grounded. Prioritizing genes based on a computer-encoded text representation, as in the third case, provides not only a semi-automated alternative to the problem of gene selection, but also allows us to accommodate for statistical pitfalls such as multiple testing. In earlier published work we have illustrated how to statistically include literature within a meta-clustering approach [8].

The approach we presented, is far from complete and open to several extensions and improvements. The use of controlled vocabularies is only a first step in the direction of more structured text mining. Up to now we neglected the structure information encoded in GO. The idea to connect the structure of ontologies more deeply to the information extraction process has recently been demonstrated by Muller *et al.* [19].

A second improvement involves the incorporation of contextual or grammatical information to extract relationships and offer improved interpretability. Good examples of such approaches can be found in Hoffmann *et al.* [12] and Blaschke *et al.* [3].

Finally, curated literature annotations as we use them are not perfect, and abstracts describing genetic properties, sequencing efforts, or irrelevant mutational analysis regularly occur. Document classification strategies as in Leonard *et al.* [18] and Raychaudhuri *et al.* [24, 23] already accommodate well for this problem.

Acknowledgements

P.G., B.C. and S.V.V. are research assistants of the K.U.Leuven. Y.M. is a postdoctoral researcher of FWO-Vlaanderen and assistant professor at the K.U.Leuven. B.D.M. is a full professor at the K.U.Leuven. Research supported by: Research Council KUL [GOA-Mefisto 666, GOA AMBioRICS, IDO (IOTA Oncology, Genetic networks), several PhD/postdoc and fellow grants]; Flemish Government [FWO: PhD/postdoc grants, projects G.0115.01 (microarrays/oncology), G.0240.99 (multilinear algebra), G.0407.02 (support vector machines), G.0413.03 (inference in bioi), G.0388.03 (microarrays for clinical use), G.0229.03 (ontologies in bioi), G.0241.04 (Functional Genomics), G.0499.04 (Statistics), research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, STWW-Genprom (gene promotor prediction), GBOU-McKnow (Knowledge management algorithms), GBOU-SQUAD (quorum sensing), GBOU-ANA (biosensors)]; Belgian Federal Science Policy Office [IUAP P5/22 (Dynamical Systems and Control: Computation, Identification and Modelling, 2002-2006)]; EU-RTD [FP5-CAGE (Compendium of Arabidopsis Gene Expression); ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours]. Part of this work was adopted from research on gene prioritization performed together with Stein Aerts.

References

[1] S. Aerts, B. Coessens, D. Lambrechts, Y. Moreau, and B. De Moor. Computational candidate gene prioritisation by genomic data fusion. submitted, 2004.

[2] K.G. Becker, D.A. Hosack, G. Dennis, R.A. Lempicki, T.J. Bright, C. Cheadle, and J. Engel. Pubmatrix: a tool for multiplex literature mining. *BMC Bioinformatics*, 4:61, 2003.

[3] C. Blaschke and A. Valencia. The frame-based module of the Suiseki information extraction system. *IEEE Intelligent Systems*, 17:14–20, 2002.

[4] D. Chaussabel and A. Sher. Mining microarray expression data by literature profiling. *Genome Biol*, 3(10):1–16, 2002.

[5] A.J. Enright and C.A. Ouzounis. BioLayout–an automatic graph layout algorithm for similarity visualization. *Bioinformatics*, 17(9):853–4, 2001.

[6] T.L. Ferea, D. Botstein, P.O. Brown, and R.F. Rosenzweig. Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc Natl Acad Sci USA*, 96:9721–9726, 1999.

[7] P. Glenisson, P. Antal, J. Mathys, Y. Moreau, and B. De Moor. Evaluation of the vector space representation in text-based gene clustering. *Pac Symp Biocomput*, pages 391–402, 2003.

[8] P. Glenisson, J. Mathys, Y. Moreau, and B. De Moor. Meta-clustering of gene expression data and literature-extracted information. *SIGKDD Explorations, Special Issue on Microarray Data Mining*, 5(2):101–112, 2003.

[9] D. Hanisch, J. Fluck, H.T. Mevissen, and R. Zimmer. Playing biology's name game: identifying protein names in scientific text. *Pac Symp Biocomput*, pages 403–414, 2003.

[10] R. Hoffmann and A. Valencia. Life cycles of successful genes. *Trends Genet*, 19:79–81, 2003.

[11] R. Hoffmann and A. Valencia. Protein interaction: same network, different hubs. *Trends Genet*, 19:681–3, 2003.

[12] R. Hoffmann and A. Valencia. A gene network for navigating the literature. *Nat Genet.*, 36:664, 2004.

[13] L. Hood and R.M. Perlmutter. The impact of systems approaches on biological problems in drug discovery. *Nat Biotechnol*, 22:1215–1217, 2004.

[14] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai. Revealing modular organization in the yeast transcriptional network. *Nat Genet*, 31:370–377, 2002.

[15] Veltman JA, Jonkers Y, Nuijten I, Janssen I, van der Vliet W, Huys E, Vermeesch J, Van Buggenhout G, Fryns JP, Admiraal R, Terhal P, Lacombe D, van Kessel AG, Smeets D, Schoenmakers EF, and van Ravenswaaij-Arts CM. Definition of a Critical Region on Chromosome 18 for Congenital Aural Atresia by ArrayCGH. *Am. J. Hum. Genet*, 72:1578–1584, 2003.

[16] T.K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet*, 28:21–28, 2001.

[17] A. Kasprzyk, D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox, and E. Birney. Ensmart: a generic system for fast and flexible access to biological data. *Genome Res.*, 14(1):160–169, 2004.

[18] J.E. Leonard, J.B. Colombe, and J.L. Levy. Finding relevant references to genes and proteins in MEDLINE using a bayesian approach. *Bioinformatics*, 18:1515–1522, 2002.

[19] H.M. Muller, E.E. Kenny, and P.W. Sternberg. Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.*, 2(11):E309, 2004.

[20] B. Palsson. Two-dimensional annotation of genomes. *Nat Biotechnol*, 22:1218–1219, 2004.

[21] C. Perez-Iratxeta, P. Bork, and M.A. Andrade. Association of genes to genetically inherited diseases using data mining. *Nat Genet.*, 31:316–319, 2002.

[22] M.F. Porter. An algorithm for suffix stripping. *Program*, 14:130–137, 1980.

[23] S. Raychaudhuri, J.T. Chang, P.D. Sutphin, and R.B. Altman. Associating genes with Gene Ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res*, 12:203–214, 2002.

[24] S. Raychaudhuri, H. Schutze, and R.B. Altman. Using text analysis to identify functionally coherent gene groups. *Genome Res*, 12:1582–1590, 2002.

[25] A.E. Shrimpton, E.M. Levinsohn, J.M. Yozawitz, D.S. Packard Jr., R.B. Cady, F.A. Middleton, A.M. Persico, and D.R. Hootnick. A hox gene mutation in a family with isolated congenital vertical talus and charcot-marie-tooth disease. *Am J Hum Genet.*, 75(1):92–96, Jul 2004.

[26] L. Tanabe, U. Scherf, L.H. Smith, J.K. Lee, L. Hunter, and J.N. Weinstein. MedMiner: An internet text-mining tool for biomedical information, with application to gene expression profiling. *BioTechniques*, 27(6):1210–1217, December 1999.

[27] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church. Systematic determination of genetic network architecture. *Nat Genet*, 22(7):281–285, 1999.

[28] S. Zuchner, I.V. Mersiyanova, M. Muglia, N. Bissar-Tadmouri, J. Rochelle, E.L. Dadali, M. Zappia, E. Nelis, A. Patitucci, J. Senderek, Y. Parman, O. Evgrafov, P.D. Jonghe, Y. Takahashi, S. Tsuji, M.A. Pericak-Vance, A. Quattrone, E. Battaloglu, A.V. Polyakov, V. Timmerman, J.M. Schroder, J.M. Vance, and E. Battologlu. Mutations in the mitochondrial gtpase mitofusin 2 cause charcot-marie-tooth neuropathy type 2a. *Nat Genet.*, 36(5):449–451, May 2004.

Tables

Table 1: Term-centric profiles of the transcriptional modules from Ihmels *et al.*. Top: 'Glutamate' is a module-specific term whereas 'isocitrate dehydrogenase' and 'TCA cycle' also rank highly when profiling the entire pathway. Using these three terms (plus 'Saccharomyces cerevisiae') to query PUBMED we are able to retrieve highly specific abstracts related to reduction in respiration capacity (or hypoxia). Note that the first reference we retrieve is the one used by Ihmels *et al.* to explain their results. Bottom: Although the terms 'carbon' and 'succinate dehydrogenase' are shared with profiles from the more general TCA pathway, they rank particularly high in this cluster and when used in combination with 'glyoxylate cycle' or 'gluconeogenesis' to query PUBMED, we are again able to retrieve very specific MEDLINE abstracts. Interestingly, the first abstract contains references to a gene whose symbol scores high in our gene-centric profile (see Table 2).

Gene symbol	Term-centric profile	Retrieved PMID	MEDLINE title
Module 1			
IDH2 CIT ACO1 IDH1	**isocitr_dehydrogenas** **tricarboxyl_acid_cycl** nad isocitr	10490611	A transcriptional switch in the expression of yeast tricarboxylic acid cycle genes in response to a reduction or loss of respiratory function.
	glutam mitochondri enzym	8672488	Expression and gene disruption analysis of the isocitrate dehydrogenase family in yeast.
	oxid carbon retrograd	1091851	Isocitrate dehydrogenases and oxoglutarate dehydrogenase activities of baker's yeast grown in a variety of hypoxic conditions.
Module 2			
MDH1 FUM1 SDH1 SDH2 SDH3 SDH4 LSC2 LPD1 KGD1 KGD2	**carbon** **succin_dehydrogenas** glucos enzym mitochondri succin growth metabol **glyoxyl_cycl** gluconeogenesi	10328823	Yeast mutants of glucose metabolism with defects in the coordinate regulation of carbon assimilation.
		3882052	Pyruvate carboxylase deficiency in yeast: a mutant affecting the interaction between the glyoxylate and Krebs cycles.
		11514507	Three target genes for the transcriptional activator Cat8p of Kluyveromyces lactis: acetyl coenzyme A synthetase genes KlACS1 and KlACS2 and lactate permease gene KlJEN1.

Table 2: Gene-centric profile of module 2 from Ihmels *et al.*. We show the top 30 'terms' and append their biological function according to GO. Genes in **bold** belong to the original query. Gene names marked with an asterisk (*) are member genes of module 1 of the TCA pathway (see Figure 3). We marked ACN9 and CAT8 in *italic*, as they also appear in our term-based analysis presented earlier. We demonstrate here that our ranking of co-linked genes via a gene-centric vocabulary can offer primary clues to interpret and further unravel sub-networks from expression clusters. Contrary to a whole network analysis, our approach comprehensibly summarizes, in a first approximation, the most important gene links from literature.

Gene 'term'	Weight	GO annotations (biological process)
SDH4	0,1235	
SDH2	0,1140	
MDH1	0,0947	
IKI1	0,0924	regulation of transcription from Pol II promoter
ACO1(*)	0,0921	
ACN9	0,0905	carbon utilization by utilization of organic compounds
		gluconeogenesis
ICL1	0,0884	
MDH2	0,0867	gluconeogenesis
		glyoxylate cycle
		malate metabolism
MLS1	0,0849	
TCM62	0,0847	protein complex assembly
IDH2(*)	0,0834	
SOR1	0,0822	fructose metabolism
		mannose metabolism
IDH1(*)	0,0775	
FBP1	0,0774	
SDH3	0,0739	
CIT1(*)	0,0725	
PCK1	0,0712	
RTG2	0,0705	intracellular signaling cascade
ELP6	0,0606	regulation of transcription from Pol II promoter
SIP4	0,0597	positive regulation of gluconeogenesis
		regulation of transcription from Pol II promoter
CAT8	0,0594	gluconeogenesis
		positive regulation of transcription from Pol II promoter
KGD2	0,0593	
HSP60	0,0592	mitochondrial matrix protein import
		protein folding
SKO1	0,0589	negative regulation of transcription from Pol II promoter
SFC1	0,0584	fumarate transport
		succinate transport
ACS1	0,0584	acetate fermentation
		acetyl-CoA biosynthesis
KGD1	0,0569	
IDP2	0,0537	
LPD1	0,0514	

Table 3: Extracting a dismorphology profile from case reports based on a dismorphology domain vocabulary.

Patient description	Defined keywords	Patient profile
This 11 years and 11 months old boy was born at 37 weeks of gestation. Birth weight was 2250g (3rd centile). Craniofacial dysmorphic features and an inguinal hernia were noted at birth. Because of feeding difficulties he was tube feeded. At the age of 1 month weight, length and OFC were all below the 3rd centile. At the age of 4 months he was hospitalised because of growth retardation and a chromosomal disorder was suspected. A partial interstitial deletion of the long arm of chromosome 2 was confirmed by routine cytogenetic investigation. He developed epileptic seizures soon after this diagnosis. At 11 months of age length was 69 cm (3rd centile), weight 6.8 kg (¡3rd centile) and OFC 44 cm (3rd centile). Craniofacial dysmorphic features were present such as flat occiput, thin white hairs, downward slanting palpebral fissures, strabismus, bilaterally epicanthic folds, ptosis of the right eyelashes, prominent nasal bridge, long philtrum, thin lips, small and high palate and dysplastic ears. The penis was small and there was a sacral dimple. There was tendency for opistotonus. Surgical intervention for ptosis of the eyelids was done at the age of 2 years. At the age of 8 years 7 months he was re-examined. Anamnesis showed that he suffered from frequent colds and otitis media. He was still tube-feeded. There were no behavioural problems. Length was 114 cm (6cm ¡3rd centile), weight 19,5 kg (3rd to 10th centile) and OFC 48,8 cm (1cm ¡3rd centile). He had microcephaly, small face, downward slanting palpebral fissures, beaked nose, short philtrum and small but high palate. There was hypotonia and he was severely mentally retarded. At the age of 11 years and 11 months length was 137 cm (3rd to 10th centile), weight 27 kg (3rd to 10th centile) and OFC of 50 cm (1cm ¡3rd centile). He still had his first teeth and an X-ray of the oral cavity showed absent adult teeth. His voice was rather specific and resembled 'Donald Duck' speech. He had contractures of the knees.	Craniofacial dysmorphic features inguinal hernia feeding difficulties growth retardation epileptic seizures flat occiput thin white hairs downward slanting palpebral fissures strabismus bilaterally epicanthic folds ptosis prominent nasal bridge long philtrum thin lips small and high palate dysplastic ears penis small sacral dimple opistotonus frequent colds otitis media microcephaly small face downward slanting palpebral fissures beaked nose short philtrum small high palate hypotonia mentally retarded absent adult teeth contractures of the knees	200300_Male genitalia, general abnormalities 5080500_Palpebral fissures, general abnormalities 030105_Microcephaly 090107_Convex/beaked profile of nose 060106_Dysplastic ears 180109_Inguinal hernia 250500_Knee, general abnormalities 320122_Seizures/abnormal EEG 080602_Epicanthic folds 130100_Teeth, general abnormalities 110303_Short philtrum 320112_Hypotonia 120403_High palate 080200_Eyelashes, general abnormalities 030401_Flat occiput 110301_Long philtrum 140100_Voice, general abnormalities 160204_Sacral dimple/sinus 100113_Small face 090302_High/prominent nasal bridge

Figures

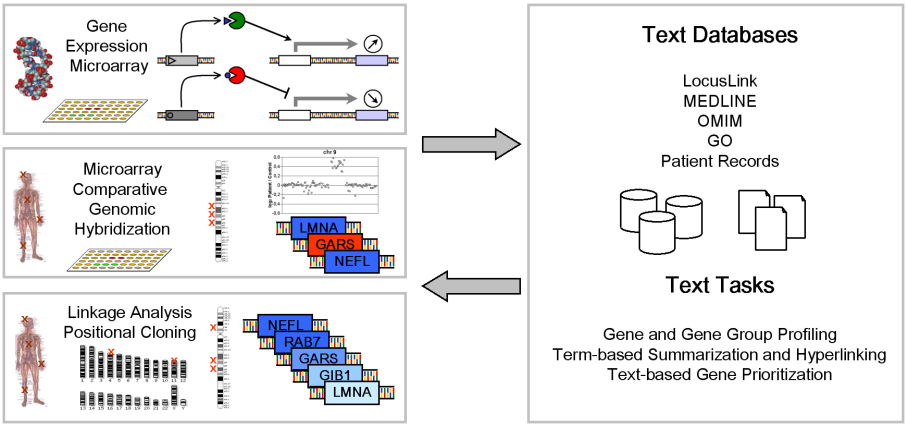


Figure 1: Overview of the proposed literature mining framework that is applied in three fields: molecular biology, biomedicine and genetics. Each of these research areas increasingly rely on a variety of genomic resources and are gradually moving into systemic approaches. Using particular ways to aggregate and index textual information on genes, diseases and patients, we demonstrate in this work how term-based text analysis constitutes a valuable component in the knowledge discovery cycle.

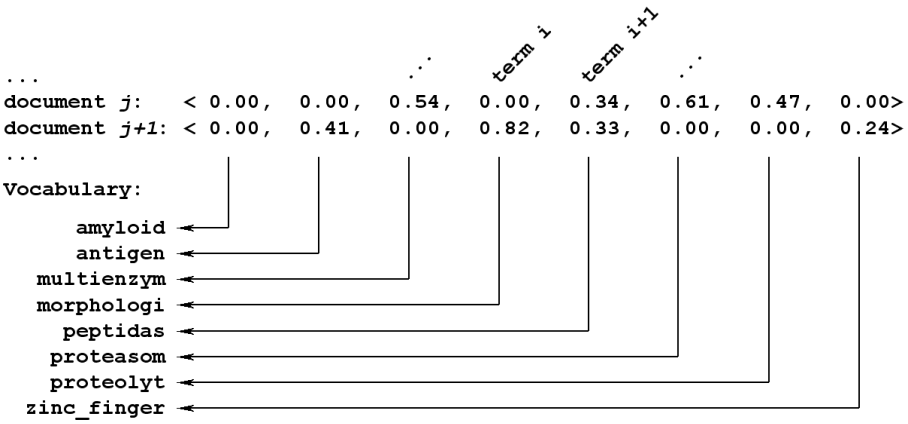


Figure 2: Illustration of term index of a given document j containing terms 'multienzyme', 'peptidase', 'proteasome' and 'proteolytic' (i.e. non-zero weights). The set of all terms is named a vocabulary. Typically stop words such as 'from', 'the', 'often', etc. are removed. Note that keywords are matched according to their stemmed form.

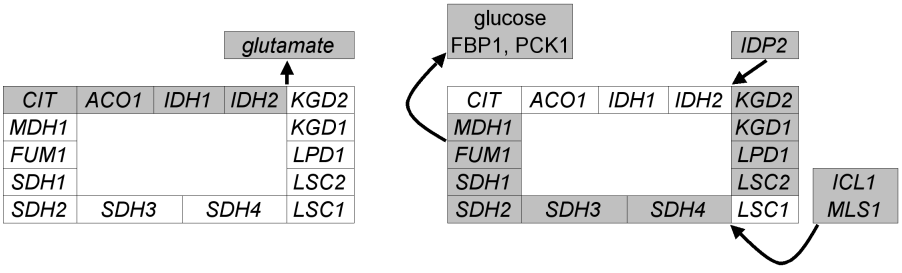


Figure 3: Two modules from the TCA pathway found by Ihmels *et al.*

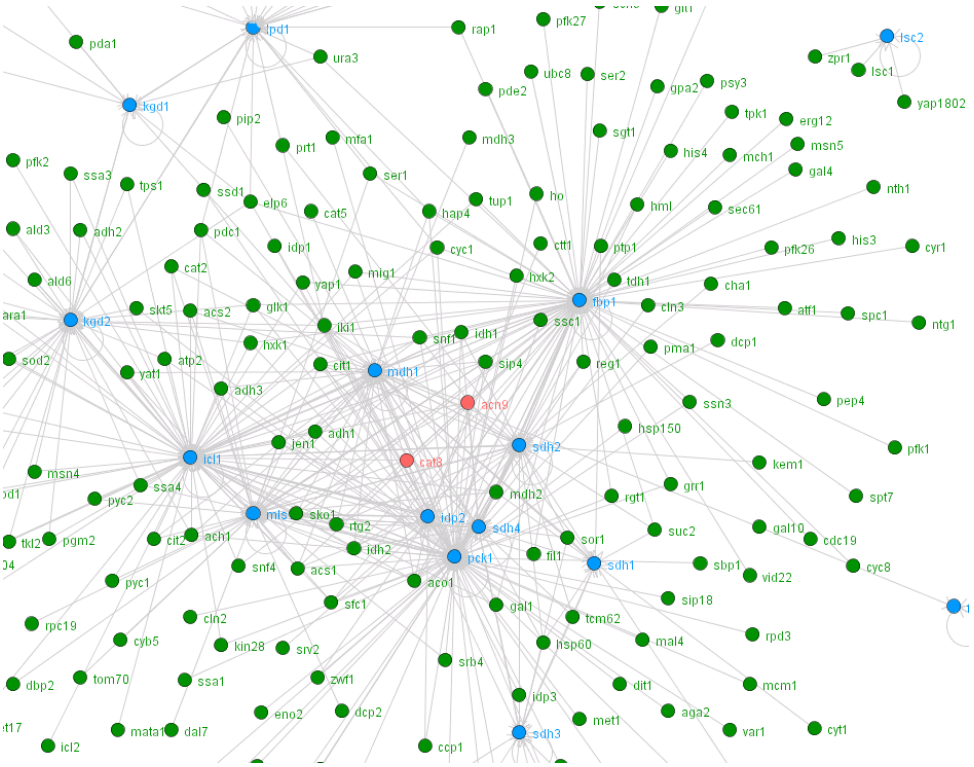


Figure 4: Our co-linkage network for module 2 discovered by Ihmels *et al.*. The query includes the 15 genes of the module and are indicated in blue to show their place in the network. In total there are 212 nodes and 457 edges. As the connectivity tends to grow with the number of genes of interest, such networks quickly become intractable for manual analysis. The text mining method is one appraoch to identify potentially interesting genes in a large network. Relevant genes derived from literature analysis are indicated in red. We also observe that few genes have many links, whereas many genes are sparsely connected. Visualization was done with the BioLayout software [5].

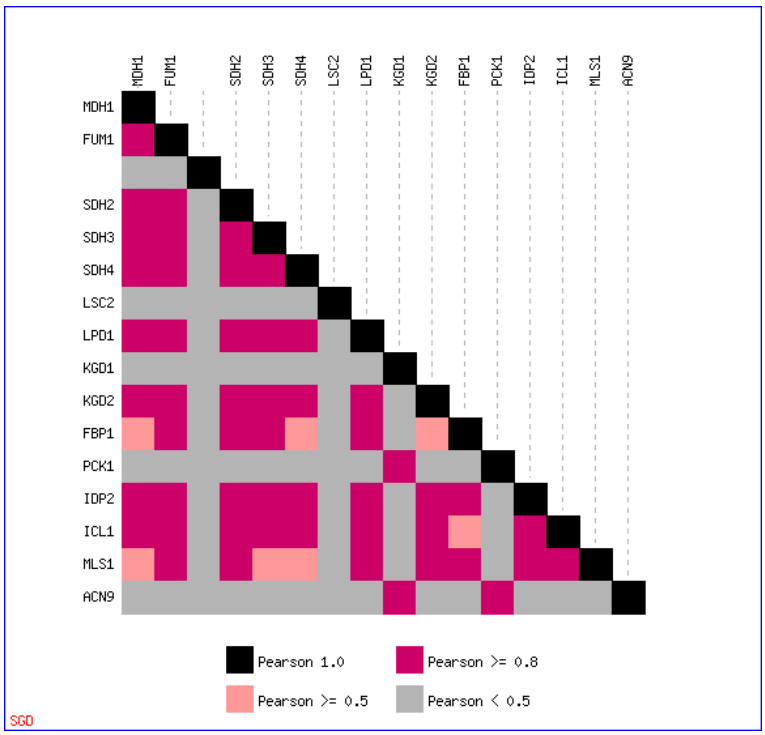


Figure 5: Visualization from SGD's *Expression Connection* of expression correlation between member genes of module 2 and ACN9 in a microarray experiment studying the effect of glucose limitation. ACN9 was independently found by term- and gene-centric analyses and shows here a high expression correlation with KGD1 and PCK1, making it an additional candidate to this module (see also Figure 3). This supports our cyclic views on the analysis process, where modules or clusters from the data world are analyzed in the text world and subsequently yield candidate genes that can be further scrutinized in the data world.

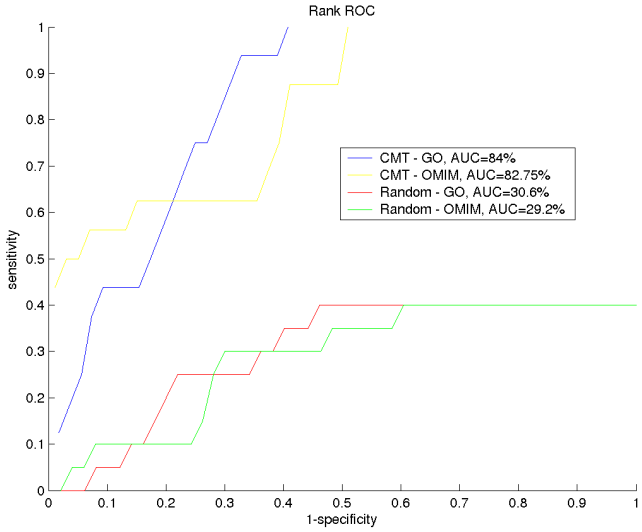


Figure 6: Rank ROC curves visualizing the performance of the prioritization method. The curves are based on information from a leave-on-out cross-validation. The two curves labeled 'CMT' show the performance of our approach with respect to the Charcot-Marie-Tooth model with an area under the curve (AUC) of 0.84 for the GO vocabulary and of 0.83 for the OMIM vocabulary. A leave-one-out procedure with a random model of 20 genes results in much worse AUCs.