Text-Based Gene Profiling with Domain-Specific Views

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

Departement Elektrotechniek, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

Abstract. The current tendency in the life sciences to spawn ever growing amounts of high-throughput assays has led to the situation were the interpretation of data and the formulation of hypotheses lag the pace with which information is produced. Although the first generation of statistical algorithms scrutinizing single, large-scale data sets found their way into the biological community, the great challenge to connect their results to the existing knowledge still remains. Despite the fairly large number of biological databases that is currently available, we find a lot of relevant information presented in free-text format (such as textual annotations, scientific abstracts, and full publications). Moreover, many of the public interfaces do not allow queries with a broader scope than a single biological entity (gene or protein). We implemented a methodology that covers various public biological resources in a flexible text-mining system designed towards the analysis of groups of genes. We discuss and exemplify how structured term- and concept-centric views complement each other in presenting gene summaries.

1 Introduction

The availability of the complete sequence of the human genome, along with those of several other model organisms, sparked a novel research paradigm in the life sciences. In 'post-genome' biology the focus is shifting from a single gene to the behavior of groups of genes interacting in a complex, orchestrated manner within the cellular environment. Recent advances in high-throughput methods enable a more systematic testing of the function of multiple genes, their interrelatedness, and the controlled circumstances in which these observations hold. Microarrays, for example, measure the simultaneous activity of thousands of genes in a particular condition at a given time. They enable researchers to identify potential genes involved in a great variety of biological processes or disease-related phenomena. As a result, scientific discoveries and hypotheses are stacking up, all primarily reported in the form of free text. A recent query in PUBMED[1] (the key bibliographic database in the life sciences) for the keyword *microarray* showed that almost a third (i.e., about 1000) of the publications

[1] http://www.ncbi.nlm.nih.giv/PubMed/

related to this technology is dated after January 2003. However, since the data and information, and ultimately the extracted knowledge itself, lack usability when offered in a raw state, various specialized database systems are designed to provide a complementary resource in designing, performing, or analyzing large-scale experiments. To date, we essentially distinguish two types of databases: the first type holds essential information, such as genomic sequence data, expression data, etc. without any extras (e.g., Genbank[2], ArrayExpress[3]); the second type offers curated annotations, cross-links to other repositories and multiple views on the same problem (e.g., LocusLink[4], SGD[5]). Although meticulous upkeep of such databases is still struggling for due credit within the community, it is indispensable for the advancement of the field [1].

The process of successfully gaining insight into complex genetic mechanisms will increasingly depend on a complementary use of a variety of resources, including the aforementioned biological databases and specialized literature on the one hand, and the expert's knowledge on the other. We therefore consider the knowledge discovery process as cyclic, (i.e., requiring several iterations between heterogeneous information sources to extract a reliable hypothesis). For example, to date, linking up analyzed microarray data to the existing databases and published literature still requires numerous queries and extensive user intervention. This process of drilling down into the 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 depicts how this cyclic nature applies to the analysis of gene expression data.

Moreover, until now, it has been largely overlooked that there is little difference between retrieving an abstract from PUBMED and downloading an entry from a biological database [2]. Fading boundaries between text from a scientific article and a curated annotation of a gene entry in a database is readily illustrated by the GeneRIF feature in LocusLink, where snippets of a relevant article pertaining to the gene's function are manually extracted and directly pasted as an attribute in the database. Conversely, we witness the emergence of richly documented web supplements accompanying a scientific publication that allow a virtual navigation through the results presented (see for example http://www.esat.kuleuven.ac.be/neurdiff/ [3]). Additionally, through the use of hypertext, electronic publications will be able to offer more structured views. Hence, we should not expect the growing amount of free text to be halted by the advent of specialized repositories.

The broadening of the biologist's scope, along with the swelling amount of information, results in a growing need to move from single gene or keyword-based queries to more refined schemes that allow a deeper interaction between the user- and context-specific views of text-oriented databases.

---

[2] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide
[3] http://www.ebi.ac.uk/arrayexpress/
[4] http://www.ncbi.nlm.nih.gov/LocusLink/
[5] http://www.yeastgenome.org/

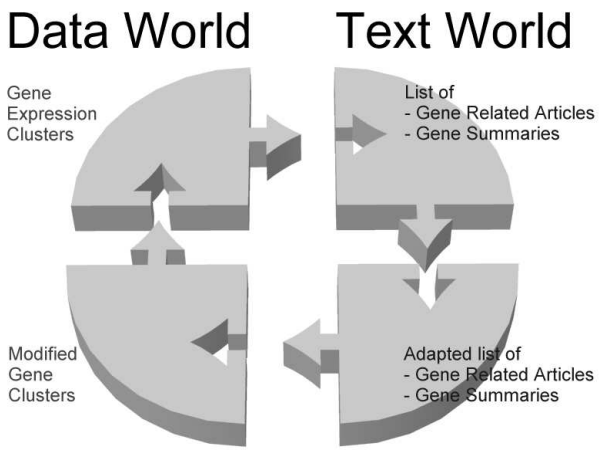

Fig. 1. Cyclic nature of the knowledge discovery process. It shows a high-level view of how it is embodied in microarray cluster analysis: starting from a cluster of genes resulting from a gene expression analysis (the 'Data World'), the corresponding literature profiles are queried and analyzed (the 'Text World'), resulting in either the addition of extra genes of interest or the omission of irrelevant genes. This updated cluster can subsequently be reanalyzed in expression space, which concludes a first cycle.

To facilitate such integrated views, controlled vocabularies that describe all properties of the underlying concepts are of great value when constructing interoperable and computer-parsable systems. A number of structured vocabularies have already arisen (most notably the Gene Ontology[6]) and, slowly but surely, certain standards are being adopted to store and represent biological data.

We can conclude that there is a certain urge towards a semantic biology web and although far from mature, some semantic web ideas have found their way into the bioinformatics community as means to knowledge representation and extraction.

Our general goal is to develop a methodology that can exploit and summarize vast amounts of textual information available in scientific publications and curated biological databases to support the analysis of *groups* of genes (e.g., resulting from gene expression analysis). As discussed above, the complexity of the domain at hand requires such a system to provide flexible views on the problem, as well as to extensively cross-link to other systems. As a result, we created a pilot text mining system, named TextGate, on top of a prevalent biological resource (LocusLink [4]) that aims, in the end, at implementing the interactive (or cyclic) nature of the knowledge discovery process.

A conceptual overview of the system is shown in Figure 2. We essentially indexed two sources of textual information. Firstly, we downloaded the entire

[6] http://www.geneontology.org

LocusLink database[7] and identified those fields that contain useful free-text information. Secondly, we collected all PUBMED abstracts that were linked to by LocusLink. We indexed both information sources with two different domain vocabularies (one based upon Gene Ontology and one based upon the unique gene names found in the HUGO nomenclature database[8]). The resulting indices are used as basis for literature profiling and further query building on the set of genes of interest.

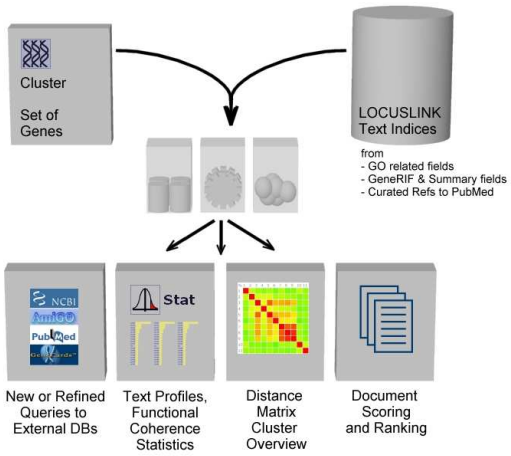

Fig. 2. Conceptual overview of the methodology behind the TextGate application. Indexing of textual gene information from the LocusLink database and abstracts from PUBMED resulted in indices for respectively genes and documents. Starting from a gene or group of genes, the most relevant documents can be retrieved by comparing indices. Afterwards, statistical analysis and further queries can be performed.

Our work is related to several other reported and available systems. PubGene[9] [5] is a database containing cooccurrence and cocitation networks of human genes derived from the full PUBMED database. For a given set of genes it reports the literature network they reside in together with their high scoring MESH headings[10]. MedMiner [6] retrieves relevant abstracts by formulating expanded queries to PUBMED. They use entries from the GeneCard database [7] to fish up additional relevant keywords to compose their query. The resulting filtered abstracts are comprehensively summarized and feedback loops are provided. GEISHA is a tool to profile gene clusters, again using the PUBMED

[7] as of April 8 2003
[8] http://www.gene.ucl.ac.uk/hugo/
[9] htpp://www.pubgene.org
[10] MESH headings are a set of keywords attached by a manual indexer to each PUBMED abstract.

engine, with an emphasis put on comprehensive summarization within a statistical framework [8]. This list of systems is not exhaustive and certainly does not encompass the spectrum of text-mining methods in genomics. Nevertheless, we believe that they well represent the first-generation systems oriented towards the considerations presented above.

The rest of this paper is organized as follows. In Section 2, we describe LocusLink and PUBMED as our information sources and how the indexed information is used to query the information space we work in. In Section 3, we discuss the construction of our two domain vocabularies and their rationale. Section 4 describes the web-based application built upon the described methodology. In Section 5 the possibilities for query expansion and cross-linking to external data sources are explored. Finally, in Section 6, we provide two illustrative biological examples of a term-based summarization and a co-linkage analysis.

2 Information Selection

2.1 LocusLink as Gene Information Source

LocusLink [4] was used as the source of textual information about genes. LocusLink is a database that organizes information from collaborating public databases and from other groups within the National Center for Biotechnology Information[11] to provide a locus-centric[12] view of genomic information from human, mouse, rat, zebrafish, Drosophila melanogaster, and HIV-1.

Each LocusLink entry (one for each locus and 225,614 in total) has a unique LocusID and consists of a number of fields with information about a gene. Examples of fields include the originating organism, summary information about the gene, official and preferred gene symbols and names, OMIM[13] [9] and PUBMED identifiers, and Gene Ontology annotations.

Although indexing these LocusLink entries can be done on all fields at once, we identified the subset that was most informative in a text-mining context. From this subset of fields we identified (possibly overlapping) groups of fields that constitute either a more specific or a more general *view* on the database. The basic aim of this design choice is that, although we wish to create a free-text index of each entry, we still want to preserve some of LocusLink's logical field structure.

2.2 PUBMED as Document Information Source

As introduced before, PUBMED is the largest bibliographic database containing over 12,000,000 citations in the biomedical literature from 1960 to present. Its great value arises from the fact that most citations have their abstract included.

[11] http://www.ncbi.nlm.nih.gov/

[12] A locus is a specific position on the chromosome.

[13] OMIM is a catalog of human genes and genetic disorders.

We downscaled the PUBMED collection to the subset of 73,172 documents used explicitly by the LocusLink curators to annotate their entries. We assume this set to be reasonably trusted and gene-specific, and therefore it constitutes a good resource for conducting our experiments.

2.3 Textual Information in the Vector Space Model

In the vector space model [10], a text body is represented by a vector (or text profile) of which each component corresponds to a single (multi-word) term from the entire set of terms taken into account (i.e., the vocabulary, see Section 3). For every component a value denotes the presence or importance of a given term, represented by a weight. Indexing is the calculation of these weights:

$$\boldsymbol{d_i} = (w_{i,1}, w_{i,2}, \ldots, w_{i,N}). \tag{1}$$

Each $w_{i,j}$ in the vector of document i is a weight for term j from the vocabulary of size N. This representation is often referred to as *bag-of-words*. In this paper we confine the discussion to the IDF weighting scheme, as it turned out to be a reasonable choice for modeling pieces of text comprising about 500 terms. The underlying assumption is that term importance is inversely proportional to frequency of occurrence. Let D be the number of documents in the collection and D_t be the number of documents containing term t, IDF is defined as:

$$\text{idf} = \log\left(1 + \frac{D}{D_t}\right). \tag{2}$$

Since, in principle, we can index the textual information from both LocusLink and PUBMED abstracts with the same vocabulary, we can represent both *genes* and *documents* as vectors of term weights [11]. We distinguish two cases:

Combining multiple documents into a single gene profile
Since each gene can have one or more curated PUBMED references associated to it in LocusLink, we combine these abstracts by taking the *mean* profile. This is illustrated in Figure 3.

Combining multiple gene profiles into a group profile
To summarize a cluster of genes and explore the most interesting terms they share, we compute the mean and variance of the terms over the group. Although simple, these statistics already reveal information on interesting terms characterizing the gene group.

The vector representation of a gene or gene group can be used as a query to retrieve documents and vice versa. The similarity of one document to another, or of a document d_i to a query q, can be calculated using the cosine distance:

$$\text{sim}_{\cos}(d_i, q) = \frac{\sum_j w_{i,j} w_{q,j}}{\sqrt{\sum_j w_{i,j}^2}\sqrt{\sum_j w_{q,j}^2}}. \tag{3}$$

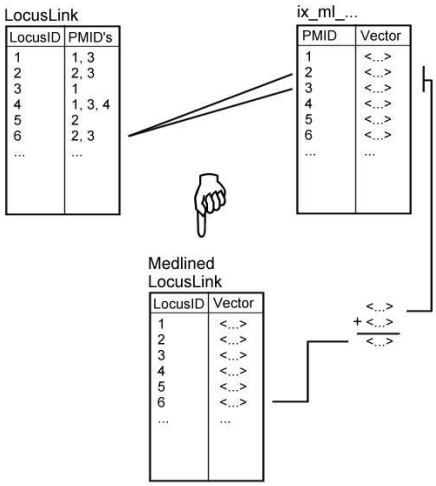

Fig. 3. Generating profiles for LocusID's via PUBMED abstract text profiles. As described in Section 2, some indices are generated using the linked abstracts as sole source of information.

3 Domain Vocabulary as Canvas to the Literature

Depending on the vocabulary chosen, the derived vector space model will be useful only within a given scope. Both the scale and diversity of the information contained in the PUBMED database form a barrier to a fast, functional interpretation of groups of genes. A well-selected corpus, together with a domain- or problem-oriented vocabulary, already alleviates this problem in a first approximation. As explained above, the PUBMED abstracts referred to in LocusLink constitute an acceptable, noise-free, and domain-specific collection. However, the information covered in this subset is still immensely vast. Although a corpus-derived vocabulary might be the first logical choice in a vector-based text mining approach, we constructed a tailored vocabulary in the light of the following issues:

Phrases
: Are additional (statistical or Natural Language Processing) algorithms needed to extract multi-word terms or are external lists available?

Synonyms
: Do we need synonym detection algorithms or can we resort to external lists?

Concept nomenclature
: Genes, proteins, diseases, chemical substances, and so on are all possible concepts of interest to the user. Hence, concept-centric views or representations might be required instead of term-centric ones. Again the question comes up whether such lists are available or need to be generated.

Database integration
Can the choice of the vocabulary enhance interoperability with other databases or systems?

Structured representation
In which way can we ultimately model dependencies between the vector components?

These issues gave rise to the construction of two vocabulary types. The first type is term-centric. It was derived from Gene Ontology (GO) [12] and comprises 17,965 terms. GO is a dynamic controlled hierarchy of (multi-word) terms with a wide coverage in life science literature, and in genetics in particular. We considered it as an ideal source to extract a highly relevant and relatively noise-free domain vocabulary. Moreover, since GO is increasingly used to annotate databases, we envision an improved interoperability with other systems. We note that, at this time, we chose to neglect the structure defining the relations between the objects, as well as the limited amount of synonym information. Genes, however, are not only referred to by their symbols (e.g., TP53), but often also by their full name, typically constituting a phrase (e.g., tumor protein p53, Li-Fraumeni syndrome) that *can* bear an indication of its function. We extracted this information and merged it with the terms from GO.

A second vocabulary type is rather concept-centric (here, gene-centric) and was constructed with the screening of cooccurrence and colinkage in mind. In our setup *cooccurrence* denotes simultaneous presence of gene names within a *single* abstract, as in [5]. *Colinkage* is a weaker form of cooccurrence and screens for simultaneous presence in the *pool* of abstracts that are linked to a given group of genes. To this end, we derived from the HUGO database [9] a vocabulary of all uniquely defined human gene symbols and their synonyms. Since these official gene symbols are requested and used by scientists, journals and databases, it can be assumed that they will occur in scientific literature with high specificity. In total this vocabulary consists of 26,511 gene symbols.

4 The TextGate Application

As many combinations of restricted views and weighting schemes (Section 2), as well as representations (Section 3) are possible, we created a database of various literature indices. Within the scope of this paper this serves the goal of offering a comprehensive interface to various views on the LocusLink database and the textual information captured inside. In a broader sense, this literature index database is part of an experimental platform to test and evaluate (combinations of) settings on a variety of biological annotation databases.

Different combinations of indexing schemes (by taking different fields of the LocusLink entries into consideration) and vocabularies show interesting possibilities towards analysis of genes and gene groups (as shown in Section 6 where three biological analysis cases are discussed).

Figure 4 shows the server architecture of the TextGate application. The different functionalities can be accessed via a browser or more directly by invoking the appropriate SOAP web service.

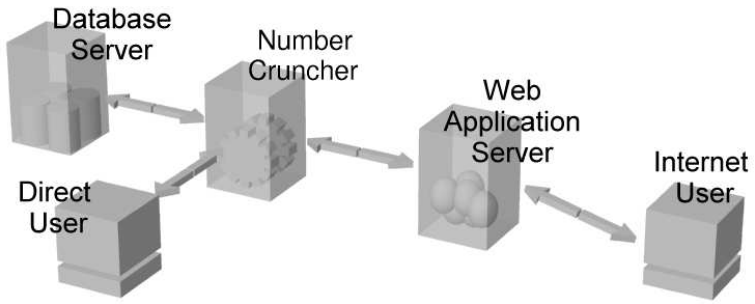

Fig. 4. Architectural overview of the TextGate knowledge discovery tool.

The user can perform a lookup of a single gene or a set of genes. In the case of profiling multiple genes, mean and variance statistics over the terms are displayed. Also, the application offers the possibility to output a distance matrix for a cluster of genes, which visualizes the distances (as calculated with Formula 3) between the text vectors of all genes in a cluster (see Figure 5).

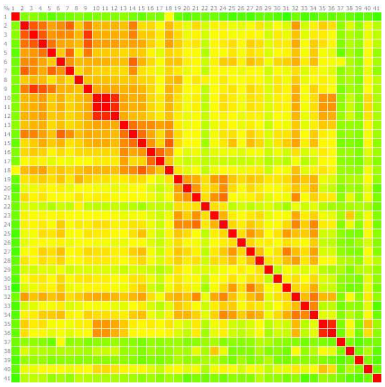

Fig. 5. Example of a distance matrix, visualizing the mutual distances between the text vector representations of a group of genes.

As said before, the functionalities of the application are also available via calls to a SOAP[14] web service. The web service can be invoked by sending the appropriate SOAP request to the TextGate web service router. The SOAP message is interpreted by an Apache Tomcat server and specific requests are sent to a number cruncher that executes the necessary calculations (as can be seen in Figure 4).

This web service architecture allows for an easy integration of the functionalities of our tool with third-party applications. SOAP clients that invoke the service can be written in the programming language of choice. Currently, in our group, we already established an integrated web environment and web service architecture for microarray analysis, called INCLUSive [13], in which TextGate fits naturally.

5 Query Expansion and Hyperlinking

Essentially, TextGate adopts a 'small world' view by scrutinizing only a restricted set of textual information extracted by specific canvases on the literature (determined by the choice of the various representations discussed in Sections 2 and 3). In practice, relevant keywords, phrases, or gene names are only useful to a researcher if they can be linked (back) to existing biological resources.

In a first attempt to strengthen this desired connection, we implemented a query composer for a variety of other databases, among which PUBMED, GeneCards, and the Gene Ontology database are the most prominent, but also OMIM, UniGene, and 15 other sources belong to the list of possible destinations. Figure 6 visualizes this functionality.

6 Example Biological Cases

In this section, we wish to provide two illustrative examples of a term-based summarization and a colinkage analysis. We added the used LocusID's in the discussion for interested readers who want to try out the interface.

6.1 Gene Ontology and Transcriptional Up- and Downregulation

In this experiment, we generated two gene clusters based upon Gene Ontology (GO) annotations of human genes. To construct the first cluster, we retrieved all human genes that are annotated with the concept *transcription activation*. The second cluster are all human genes annotated with the concept *transcription repression*. Both concepts apply to the process of transcriptional regulation in the cell (see Figure 7). Whether a protein complex promotes or inhibits transcription of a gene, depends upon its constitution and environmental conditions. This

[14] SOAP (Simple Object Access Protocol) is an XML-based W3C Proposed Recommendation for exchanging structured information in a decentralized, distributed environment.

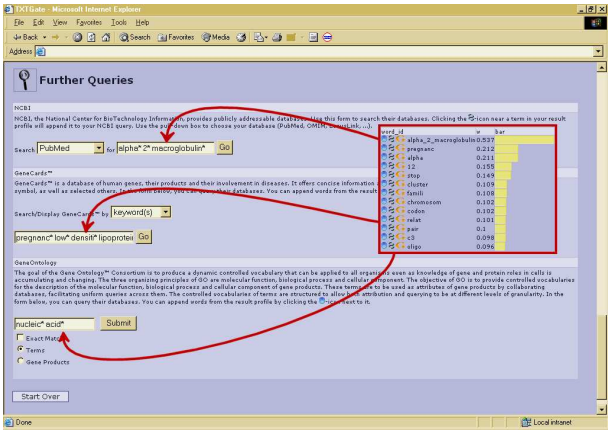

Fig. 6. The cyclic approach to knowledge mining by composing refined queries to a set of public databases.

makes the distinction between both concepts not a trivial task, since a protein can be active in a complex as inhibitor and as activator. The genes in both groups are enlisted in Table 1.

In the first place this indicates that our text-mining approach is reasonably trustable. As our confidence in these kind of methods will grow, one could invert the reasoning and consider this case to give an indication of whether or not the GO curators have made a good choice of splitting the concept of *transcriptional regulation* in *transcription activation* and *transcription repression*: if for those two different clusters TextGate shows that in essence the same terms occur this would mean that there is not really a significant difference between the genes GO associated to *transcription activation* and *transcription repression*. If, however, specific terms linked to activation and repression respectively occur for the activation cluster and the repression cluster, then making two taxons under *transcriptional regulation* was a good choice.

In Table 2, the term ranking and variance are shown for the activation cluster (top of the table) and the repression cluster (bottom). We see an obvious difference in term occurrence. For the activation cluster, **transcript_activ** ranks third place, and for the repression cluster, **repressor** and **repress** rank first and second, respectively.

6.2 Colinkage of Colon Cancer Genes

In Section 3 we discussed how changing the way domain vocabularies and index tables are constructed provides us with a different view on the information. Using only the gene names from the HUGO database [9] as domain vocabulary, we can take a specific stance towards investigating colinkage of genes.

Table 1. Gene symbols and LocusLink identifiers for the two clusters of human genes that are annotated with respectively the Gene Ontology terms *transcription activation* and *transcription repression*.

Activation cluster		**Repression cluster**	
Gene Symbol	*LocusID*	*Gene Symbol*	*LocusID*
BRCA1	672	BTF	9774
BRCA2	675	DMAP1	55929
CGBP	30827	DNMT3L	29947
COPEB	1316	EED	8726
EDF1	8721	EPC1	80314
ELF1	1997	HDAC4	9759
ELF2	1998	HDAC6	10013
EPC1	80314	IFI16	3428
ETV4	2118	LRRFIP1	9208
FOXC1	2296	MBD1	4152
FOXD3	27022	MBD2	8932
HNRPD	3184	NAB1	4664
HOXA9	3205	NRF	55922
HOXC9	3225	NSEP1	4904
HOXD9	3235	PIASY	51588
KLF2	51713	RBAK	57786
MADH1	4086	REST	5978
MADH5	4090	RING1	6015
MITF	4286	THG-1	81628
MYB	4602	UBP1	7342
NSBP1	79366	ZFHX1B	9839
ONECUT1	3175	ZNF24	7572
RREB1	6239	ZNF253	56242
SEC14L2	23541	ZNF33A	7581
SUPT3H	8464	ZNFN1A4	64375
TITF1	7080		
TP53BP1	7158		
TRIP4	9325		
UBE2V1	7335		
ZNF38	7589		
ZNF148	7707		
ZNF398	57541		

Table 2. For the *transcription activation* and *transcription repression* clusters we show the ranking of the 20 terms with the highest mean (left side) and the ranking of the 20 with the highest variance (right side). We note the presence of some noise due to the nature of the term extraction process.

Term	**Mean**	**Term**	**Variance**
transcript_factor	0.205	ovarian	0.011
dna_bind	0.188	thyroid	0.0070
transcript_activ	0.139	site_select	0.0050
nuclear	0.129	h3	0.0050
transcript	0.125	zinc	0.0050
promot	0.117	p53	0.0040
bind	0.113	ey	0.0040
tumor	0.113	hepatocyt	0.0040
domain	0.112	melanocyt	0.0040
famili	0.11	cluster	0.0040
chromosom	0.106	prime	0.0040
site	0.098	bridg	0.0040
pair	0.096	transcript_factor	0.0030
involv	0.095	transform_growth_factor_beta	0.0030
region	0.093	retino_acid_metabol	0.0030
yeast	0.092	tumor_suppressor	0.0030
two	0.09	ubiquitin_conjug_enzym	0.0030
zinc	0.088	leukemia	0.0030
contain	0.088	7	0.0030
map	0.087	pigment	0.0030
Term	**Mean**	**Term**	**Variance**
repressor	0.238	methyl_cpg_bind	0.019
repress	0.205	deacetylas	0.013
dna_bind	0.172	cytosin_5	0.0090
zinc	0.164	repressor	0.0090
transcript_repressor	0.158	histon	0.0080
deacetylas	0.157	polycomb_group	0.0080
transcript_factor	0.151	dna_methyl	0.0060
domain	0.147	ring	0.0060
histon	0.127	zinc	0.0060
transcript	0.123	transcript_repressor	0.0050
yeast	0.116	methyltransferas	0.0050
famili	0.109	silenc	0.0050
gene_express	0.109	hi	0.0050
methyl_cpg_bind	0.105	interferon_gamma	0.0050
region	0.104	stat2	0.0040
nucleu	0.104	cell_structur	0.0040
interact	0.103	leucin_metabol	0.0040
protein_metabol	0.1	polycomb	0.0040
bind	0.1	lrr	0.0040
line	0.095	methyl	0.0040

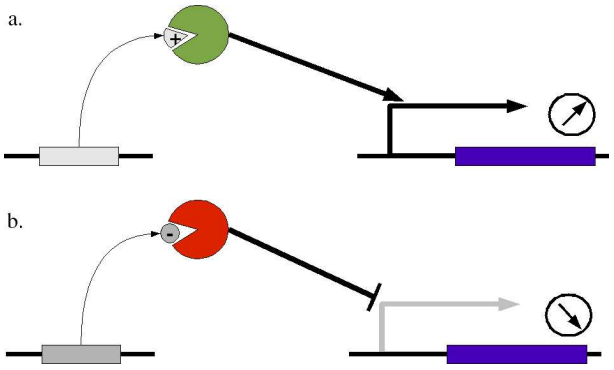

Fig. 7. The activation (a) and repression (b) of the transcription of a gene by DNA-binding protein complexes. The squares represent genes on the DNA. The circles represent protein complexes. In case (a), binding of an activator protein (produced by its corresponding gene) to the complex initiates, and subsequently activates transcription of a given gene while in case (b), binding of a repressor protein (produced by its corresponding gene) inhibits expression of that gene.

For this test case, we constructed a set of genes by consulting a textbook on molecular biology [14] and choosing genes that are related to colon cancer manually. This set was then provided to TextGate using the colinkage index. The set of genes is shown in Table 3. The results are shown in Table 4.

Table 3. A set of seven genes involved in colon cancer.

HUGO Name	LocusID
k-RAS2	3845
NEU1	4758
MYC	4609
APC	324
DCC	1630
P53	7157
MSH2	4436

To validate this result, we verified that these gene names indeed turn up in the literature in relation to colon cancer.

The highest scoring gene is the *CD44 antigen*. This gene is indeed related to colon cancer, as shown in a paper by Barshishat *et al.* [15].

The second ranking gene name is *UBE3A* (ubiquitin protein ligase E3A). At first sight, it is not directly related to colon cancer, but after closer investigation of the available literature, we found that this gene is involved in degradation of

Table 4. For the colon cancer cluster we show the ranking of the 20 colinkage concepts with the highest mean (left side) and the ranking of the 20 colinkage concepts with the highest variance (right side). We note the presence of some noise due to the nature of the concept extraction process.

Gene	Mean	Gene	Variance
cd44	0.446	myc	0.013
ube3a	0.429	pten	0.012
i	0.344	apc	0.01
wwox	0.28	tp53	0.01
sparc	0.27	dcc	0.0090
pax6	0.234	msh2	0.0050
wa	0.232	pax6	0.0040
rieg2	0.223	ra	0.0030
at	0.162	wwox	0.0030
nr4a2	0.156	map	0.0030
ha	0.136	pms2	0.0030
gstz1	0.125	rieg2	0.0030
msh2	0.081	mlh1	0.0030
1	0.081	12	0.0030
3	0.078	ha	0.0020
all	0.077	wa	0.0020
5	0.075	hla	0.0020
kptn	0.066	all	0.0020
tp53	0.065	nr4a2	0.0020
nup214	0.064	gstz1	0.0010

TP53, which plays a crucial role in the regulation of cell division (mitosis) [16]. This explains the detection of frequent co-citation.

7 Conclusion and Future Work

As contemporary biology is evolving towards an information science, integrative views on biological problems will be of increasing importance. Integration is a broad term and is understood differently in the database community than for instance in the field of machine learning. Our perspective on integration was adopted with both the (presumed) cyclic nature of the knowledge discovery process and of a text-mining application in mind. We created various indices on two text-oriented databases (the annotation database LocusLink and the literature repository PUBMED) that enabled text summarization of multiple genes at once. Supported by grateful realizations in the development of annotation standards, nomenclature conventions, and ontologies, TextGate is able to formulate sensible queries to a variety of other resources (including back the GO). However, the system is far from complete, and represents only a first step in the construction of a knowledge discovery platform. Our mid-term challenges include:

Extension to an IR engine
: At this point TextGate uses the index tables in a gene-centric way to summarize and link information. As biological experiments are always carried out in a particular context, allowing term-centric queries (see e.g., the recently established TREC[15] track) would further enhance the usability of the system. This would fully close the cycle between terms, genes, documents, and database annotations.

Extension of the conceptual representations
: Up to now we neglected the structure of GO. Embedding its structure as well as adding additional ontologies for functional genomics[16], or biomedicine[17] would provide more structured views on information.

Finally, since the core functionality of the TextGate system is also provided as a SOAP service, it can seamlessly be integrated with other systems, primarily the expression analysis pipeline currently present in our lab[18].

Acknowledgments

P.G. and B.C. are research assistants of the K.U.Leuven. S.V.V is an intern in fulfillment of the Master in Bioinformatics Program at the K.U.Leuven. Y.M. is a post-doctoral researcher of FWO-Vlaanderen and assistant professor at the

[15] http://trec.nist.gov/
[16] for example: http://www.sofg.org/index.html
[17] for example: http://www.nlm.nih.gov/research/umls/umlsmain.html
[18] http://www.esat.kuleuven.ac.be/inclusive/

K.U.Leuven. B.D.M. is a full professor at the K.U.Leuven. Research supported by Research Council K.U.Leuven: [GOA-Mefisto 666, 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), research communities (ICCoS, ANMMM)]; AWI: [Bil. Int. Collaboration Hungary/Poland]; IWT: [PhD Grants, STWW-Genprom (gene promotor prediction), GBOU-McKnow (Knowledge management algorithms), GBOU-SQUAD (quorum sensing), GBOU-ANA (biosensors)]; Belgian Federal Government: [DWTC (IUAP IV-02 (1996-2001) and IUAP V-22 (2002-2006)]; EU: [CAGE]; ERNSI; Contract Research/agreements: [Data4s, Electrabel, Elia, LMS, IPCOS, VIB]. We acknowledge Peter Antal for starting up this research direction.

References

1. Navarro, D., Niranjan, V., Peri, S., Jonnalagadda, C., Pandey, A.: From biological databases to platforms for biomedical discovery. Trends Biotechnol. **21** (2003) 263–268
2. Gerstein, M., Junker, J.: Blurring the boundaries between scientific papers and biological databases. Nature Online, http://www.nature.com/nature/debates/e-access/Articles/gernstein.html (web debate, on-line 7 May 2001)
3. Dabrowski, M., Aerts, S., Hummelen, P.V., Craessaerts, K., De Moor, B., Annaert, W., Moreau, Y., De Strooper, B.: Gene profiling of hippocampal neuronal culture. J. Neurochem. **85** (2003) 1279–1288
4. Pruitt, K., Maglott, D.: RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res. **29** (2001) 137–140
5. Jenssen, T., Laegreid, A., Komorowski, J., Hovig, E.: A literature network of human genes for high-throughput analysis of gene expression. Nature Genet. **28** (2001) 21–28
6. Tanabe, L., Scherf, U., Smith, L., Lee, J., Hunter, L., Weinstein, J.: MedMiner: An internet text-mining tool for biomedical information, with application to gene expression profiling. BioTechniques **27** (1999) 1210–1217
7. Rebhan, M., Chalifa-Caspi, V., Prilusky, J., Lancet, D.: GeneCards: A novel functional genomics compendium with automated data mining and query reformulation support. Bioinformatics **14** (1998) 656–664
8. Blaschke, C., Oliveros, J., Valencia, A.: Mining functional information associated with expression arrays. Funct. Integr. Genomics **1** (2001) 256–268
9. McKusick, V.: Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders. Twelfth edn. Johns Hopkins University Press (1998)
10. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press / Addison-Wesley (1999)
11. Glenisson, P., Antal, P., Mathys, J., Moreau, Y., Moor, B.D.: Evaluation of the vector space representation in text-based gene clustering. In: Proceedings of the Pacific Symposium on Biocomputing. Volume 8. (2003) 391–402
12. The Gene Ontology Consortium: Gene Ontology: tool for the unification of biology. Nature Genet. **25** (2000) 25–29

13. Coessens, B., Thijs, G., Aerts, S., Marchal, K., Smet, F.D., Engelen, K., Glenisson, P., Moreau, Y., Mathys, J., Moor, B.D.: INCLUSive - a web portal and service registry for microarray and regulatory sequence analysis. Nucleic Acids Res. **31** (2003) 3468–3470
14. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P.: Molecular Biology of the Cell. Fourth edn. Garland Science Publishing (2002)
15. Barshishat, M., Levi, I., Benharroch, D., Schwartz, B.: Butyrate down-regulates CD44 transcription and liver colonisation in a highly metastatic human colon carcinoma cell line. Br. J. Cancer **87** (2002) 1314–1320
16. Levine, A.: p53, the cellular gatekeeper for growth and division. Cell **88** (1997) 323–331