Software

TXTGate: profiling gene groups with text-based information

Patrick Glenisson[*], Bert Coessens[*], Steven Van Vooren[*], Janick Mathys[*], Yves Moreau[*†] and Bart De Moor[*]

Addresses: [*]Departement Elektrotechniek (ESAT), Faculteit Toegepaste Wetenschappen, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Heverlee (Leuven), Belgium. [†]Current address: Center for Biological Sequence Analysis, BioCentrum, Danish Technical University, Kemitorvet, DK-2800 Lyngby, Denmark.

Correspondence: Bert Coessens. E-mail: bert.coessens@esat.kuleuven.ac.be

Published: 28 May 2004

Genome ***Biology*** 2004, **5:**R43

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/6/R43

Received: 24 November 2003
Revised: 3 February 2004
Accepted: 27 April 2004

Abstract

We implemented a framework called TXTGate that combines literature indices of selected public biological resources in a flexible text-mining system designed towards the analysis of groups of genes. By means of tailored vocabularies, term- as well as gene-centric views are offered on selected textual fields and MEDLINE abstracts used in LocusLink and the *Saccharomyces* Genome Database. Subclustering and links to external resources allow for in-depth analysis of the resulting term profiles.

Rationale

Recent advances in high-throughput methods such as microarrays enable systematic testing of the functions of multiple genes, their interrelatedness and the controlled circumstances in which ensuing observations hold. As a result, scientific discoveries and hypotheses are stacking up, all primarily reported in the form of free text. However, as large amounts of raw textual data are hard to extract information from, various specialized databases have been implemented to provide a complementary resource for designing, performing or analyzing large-scale experiments.

Until now, the fact that there is little difference between retrieving an abstract from MEDLINE and downloading an entry from a biological database has been largely overlooked [1]. The fading of the 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 [2], where snippets of a relevant article pertaining to a gene's function are manually extracted and directly pasted as an attribute in the database. The broadening of biologists' scope of investigation, along with the growing amount of information, result in an increasing need to move from single gene or keyword-based queries to more refined schemes that allow comprehensive views of text-oriented databases.

As gene-expression studies typically output a list of dozens or hundreds of genes that are co-expressed, a researcher is faced with the assignment of biological meaning to such lists. Several text-mining approaches have been developed to this end. Masys *et al.* [3] link groups of genes with relevant MEDLINE abstracts through the PubMed engine. Each cluster is characterized by a pool of keywords derived from both the Medical Subject Headings (MeSH) and the Unified Medical Language System (UMLS) ontology. Jenssen *et al.* [4] set up a pioneering online system to link co-expression information from a microarray experiment with the cocitation network they constructed. This literature network covers co-occurrence information of gene identifiers in more than 10 million MEDLINE abstracts. Their system characterizes co-expressed genes using the MeSH keywords attached to the abstracts about those genes. Shatkay *et al.* [5] link abstracts to genes in a probabilistic scheme that uses the EM algorithm to estimate the parameters of the word distributions underlying a

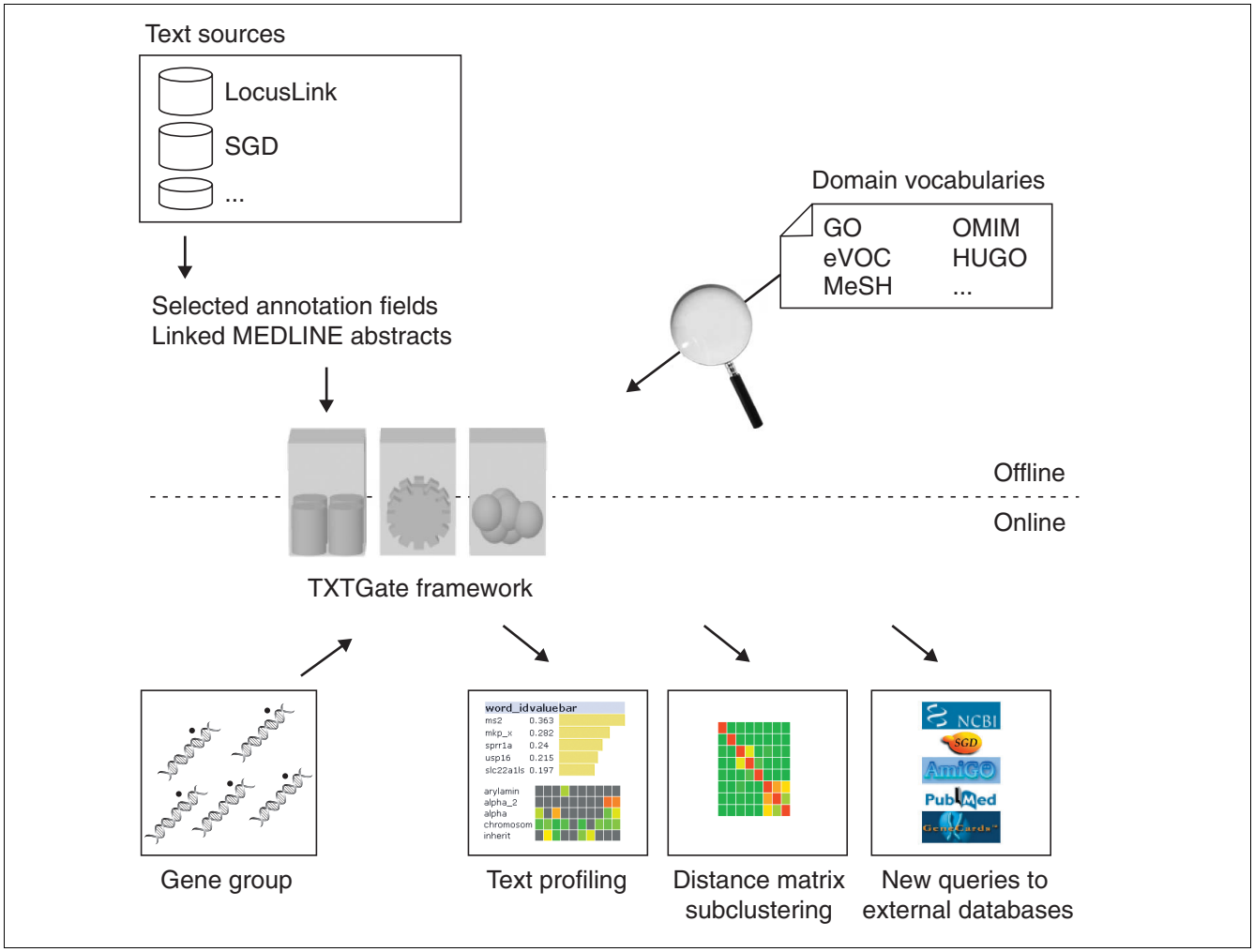

Figure 1
Conceptual overview of TXTGate. We indexed two different sources of textual information about genes (LocusLink and SGD) using different domain vocabularies (offline process). These indices are used online for textual gene profiling and clustering of interesting gene groups. TXTGate's link-out feature to external databases makes it possible to investigate the profiles in more detail.

'theme'. Genes are identified as similar when their corresponding gene-by-documents representations are close. Chaussabel and Sher [6] and Glenisson *et al.* [7] provide a proof of principle on how clustering of genes encoded in a keyword-based representation can further discern relevant subpatterns. Finally, Raychaudhuri *et al.* [8] developed a method called neighborhood divergence, to quantify the functional coherence of a group of genes using a database that links genes to documents. The score is successfully applied to both gold-standard and expression data, but has the slight drawback that it does not give information on the actual function. Their method is indeed geared to the identification of biologically coherent groups, rather than their interpretation.

Our system is built taking into account three main considerations, in an attempt to improve the quality and interpretability of term profiles. First, the construction of a sound linkage between genes and MEDLINE abstracts is often problem-dependent and constitutes a research track on its own that requires advanced document-classification strategies as, for example, proposed by Leonard *et al.* [9] or Raychaudhuri *et al.* [10]. Despite some shortcomings, therefore, curated gene-literature references are helpful resources to exploit. Second, the information contained within curated gene references is sometimes diverse and can range from sequence to disease. In addition, the research questions that scientists are addressing when they scrutinize gene groups from high-throughput assays are similarly diverse. Therefore, considering all the terms occurring in a large set of documents (that is, a general vocabulary) might be detrimental to the extraction of terms that are relevant to the question at hand. The construction of separate vocabularies according to gene name, disease and function seems a logical choice to provide increased insight. Third, as mentioned previously,

annotations offered by curated gene databases are often in semi-structured form and encompass keywords, sentences or paragraphs. To facilitate integration of such annotations with existing knowledge, controlled vocabularies that describe conceptual properties are of great value when constructing interoperable and computer-parsable systems. A number of structured vocabularies have already arisen (Gene Ontology (GO) [11], MeSH [12], eVOC [13]) and, slowly but surely, certain standards are systematically being adopted to store and represent biological information [14].

Armed with these insights, we developed TXTGate [15], a platform that offers multiple 'views' on vast amounts of gene-based free-text information available in selected curated database entries and scientific publications. TXTGate enables detailed functional analysis of interesting gene groups by displaying key terms extracted from the associated literature and by offering options to link out to other resources or to sub-cluster the genes on the basis of text. This way, we address on the one hand the need for easy means to validate gene clusters arising from, for instance, microarray experiments, and on the other hand the problem of using scientific literature in the form of free text as a source of functional information about genes. The strength of TXTGate is its use of tailored vocabularies to visualize only the information most relevant to the query at hand. TXTGate is implemented as a web application and is available for academic use [15].

Related software

This work extends the general ideas of textual profiling and clustering presented in Blaschke *et al.* [16] and Chaussabel and Sher [6], where the utility of literature indices for profiling gene groups in yeast and humans is proven. TXTGate implements the vector-space model for gene profiling [7] and provides indices for MEDLINE abstracts and selected functional annotations from two public databases. Various engineered domain-specific vocabularies (term- as well as gene-centric) act as perspectives to the literature and the tool provides direct links to external resources. In what follows, we compare TXTGate to other reported biological text-mining software.

MedMiner [17,18] retrieves relevant abstracts by formulating expanded queries to PubMed. It uses entries from the GeneCards database [19] to fish for additional relevant keywords to expand a query. The resulting filtered abstracts are summarized in keywords and sentences, and feedback loops are provided. Nevertheless, the system is directed at querying terms and specific gene-drug or gene-gene relationships, rather than at scrutinizing gene clusters. MedMOLE [20,21] is also a system to query MEDLINE more intelligently and detects Human Genome Organization (HUGO) names in abstracts via a natural language processing (NLP)-based gene-name extractor. The retrieved abstracts can be clustered, and top keywords are presented. However, the application scales less well, is not effective at profiling groups of genes, and the summaries provide much less detail than MedMiner and TXTGate. GEISHA [16,22] is a tool for profiling gene clusters with an emphasis on summarization within a shallow parsing framework. This system was implemented for *Escherichia coli* but is no longer updated. PubGene [4,23] is a database containing gene co-occurrence and cocitation networks of human genes derived from the full MEDLINE database. For a given set of genes it reports the literature network they reside in, together with their high-scoring MeSH terms. As not all relevant information can be captured by gene symbols or MeSH terms, the functionalities offered by TXTGate provide complementary views to interpret groups of genes. Although our colinkage feature (being a weaker form of co-occurrence that spans only the set of 73,152 MEDLINE abstracts used in LocusLink) is less elaborate than the possibilities offered by PubGene, we will show its utility and added value through its integration in the broader TXTGate framework. MedGene [24,25] and G2D [26,27] are specialized databases that, in contrast to TXTGate, are geared at ranking genes by disease. They accept user-defined queries scrutinizing gene-disease, disease-disease or gene-gene relationships extracted from the literature. Finally, MeKE [28,29] is an application listing gene functions extracted by an ontology-based NLP system. Its current setup is directed more towards a functional knowledge base, rather than comprehensibly profiling information coming from groups of genes, as offered by our software.

Application overview

A conceptual overview of the system is shown in Figure 1. Various literature indices were created based on selected annotation fields and linked MEDLINE information, both present in the curated repositories LocusLink and the *Saccharomyces* Genome Database (SGD). Several tailored vocabularies derived from public resources (GO, MeSH, Online Mendelian Inheritance in Man (OMIM), eVOC and HUGO) act as a

Table 1

Overview of the indexed resources of textual information in TXTGate

Resource	Information fields	Domain vocabularies used
LocusLink	Linked MEDLINE abstracts	GO, MeSH, eVOC, OMIM, HUGO gene symbols
	GeneRIF annotations	GO
	Functional summaries	GO
	GO annotations	GO
SGD	Linked MEDLINE abstracts	GO-pruned, SGD gene symbols
	GO annotations	GO-pruned

In the second column we specify which fields of the resource were used. The third column lists the domain vocabularies with which the information was indexed.

perspective on the textual information. A user-defined query on any of these indices by providing a group of genes of interest results in a summary keyword profile which can be used for further query building for a variety of other databases. Currently, TXTGate smoothly accommodates queries of around 200 genes. Alternatively, the group can be subclustered on the basis of the selected textual information to discern substructures not apparent in the original summary profile. The operations that can be carried out are described below.

Combining multiple, linked documents into a single gene profile

When a given gene has several curated MEDLINE references associated to it, we combine these abstracts into an indexed gene entry by taking the mean profile. This operation is part of the offline process.

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. This is performed online.

Subclustering gene profiles

We offer the possibility online of subclustering a group of a maximum of 200 genes by means of hierarchical clustering. Ward's method was chosen because of its deterministic nature and the computational advantage of using the same solution when consecutively considering different numbers of clusters *k*. By varying the threshold at which to cut the tree, we can obtain an arbitrary number of clusters.

Text profiling, clustering and the supporting web interface are implemented as a Java web application that communicates with a mySQL database via Java Remote Method Invocation [30]. The literature indices are generated using custom-developed indexing software written in C++. Code is available on request.

Program development

Indexing

The indices are built using the vector-space model [31], where a textual entity 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 (the vocabulary) being used. For each component a value denotes the importance of a given term, represented by a weight. Indexing a document $\vec{d}_i$ is performed by the calculation of these weights:

$$\vec{d}_i = \left(w_{i,1}, w_{i,2}, \ldots, w_{i,N} \right).$$

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'. All textual information is stemmed using the Porter stemmer [32] (stemming is the automated conflation of related words, usually by reducing the words to a common root form) and indexed with a normalized inverse document frequency (IDF) weighting scheme, a reasonable choice for modeling pieces of text comprising up to 200 terms, as observed in database annotations and MEDLINE abstracts. With *D* the number of documents in the collection and D_t the number of documents containing term *t*, IDF is defined as

$$IDF = \log\left(\frac{D}{D_t}\right).$$

We downloaded the entire LocusLink (as of 8 April, 2003) and SGD (15 January, 2003) databases, and identified and indexed subsets of fields (such as GO annotations and functional summaries) that were most sensible in the presented context. Although indexing these database entries could have been performed on all fields at once, we deemed a preservation of selected parts of LocusLink's and SGD's logical field structure more appropriate for functional gene profiling. We indexed not only the textual annotations but also the 73,152 MEDLINE abstracts referred to in all entries of LocusLink, as well as the 24,909 abstracts linked to from SGD. Gene-specific indices were created by taking the average over all indices of MEDLINE abstracts annotated to a certain gene in LocusLink and SGD. The resulting indices are used in TXTGate as a basis for literature profiling and further query building of genes of interest. Table 1 overviews the indexed resources of textual information and connects them to the used domain vocabularies.

Construction of domain vocabularies

We constructed five different term-centric domain vocabularies that provide different views on the gene-specific information we indexed. All vocabulary sources underwent parsing and pruning operations to obtain stemmed words and

Table 2

Overview of the domain vocabularies in TXTGate

Domain vocabulary	Number of terms
Term-centric	
GO	17,965
GO-pruned (yeast)	3,867
MESH	27,930
OMIM	2,969
eVOC	1,553
Gene-centric	
HUGO gene symbols (human)	26,511
SGD gene symbols (yeast)	11,319

The vocabularies are named after the resource they stem from.

phrases, eliminating stop words (such as 'then', 'as', 'of', 'gene') from a handcrafted list. We again applied the Porter stemmer [32]) to avoid information loss due to morphological and inflexional endings. Although stemming is not always desirable, for relatively small documents it has proved advantageous. Where applicable we derived phrases directly from the vocabulary source.

A first vocabulary was derived from the GO [11] and comprises 17,965 terms. GO is a dynamic controlled hierarchy of multi-word terms with a wide coverage of life-science literature, and genetics in particular. We considered it an ideal source from which to extract a highly relevant and relatively noise-free domain vocabulary. We retained all composite GO terms shorter than five tokens as phrases. Longer terms containing brackets or commas were split to increase their detection. For the yeast indices, we pruned the vocabulary, retaining only those terms occurring at least twice and in less than 20% of all MEDLINE abstracts referred to in SGD [33], obtaining a new vocabulary of 3,867 terms.

Two other domain vocabularies are rather similar in scope but differ in size. One is based on the MeSH [12], the National Library of Medicine's controlled vocabulary thesaurus, and counts 27,930 terms. The other is based on OMIM's Morbid Map [34]. This is a cytogenetic map location of all disease genes present in OMIM and their associated diseases. We extracted all disease terms to construct a 2,969-term vocabulary. A fifth domain vocabulary was drawn from eVOC [13], a thesaurus consisting of four orthogonal controlled vocabularies encompassing the domain of human gene-expression data. It includes terms related to anatomical system, cell type, pathology, and developmental stage.

In addition to these term-centric domain vocabularies we constructed two gene-centric vocabularies with the screening of co-occurring and colinked genes in mind. 'Co-occurrence' denotes the simultaneous presence of gene names within a single abstract, as described by Jenssen *et al.* [4]. We define 'colinkage' here as a weaker form of co-occurrence screening for the simultaneous presence of gene names in the pool of abstracts that is linked to a given group of genes.

From the HUGO database [35] we derived a vocabulary consisting of all uniquely defined human gene symbols and their synonyms. In total, this vocabulary consists of 26,511 gene symbols. The second vocabulary consists of all uniquely defined yeast gene symbols found in SGD and contains 11,319 terms. As these official gene symbols are frequently requested and used by scientists, journals and databases, we assume they constitute a good first approximation to detect gene occurrence in MEDLINE abstracts. The domain vocabularies we adopted are listed in Table 2.

Online clustering

The online clustering is done with our own implementation in Java of Ward's method for hierarchical clustering [36]. Ward's method outperforms single, average or complete linkage. The similarity measure used is the cosine distance between two vector representations $\vec{d}_i$ and $\vec{d}_j$. The similarity between a newly formed cluster (r, s) (by linking two existing vectors/clusters) with $(n_r + n_s)$ elements and an existing cluster (t) with n_t elements is given by

$$d[(t), (r, s)] = \alpha_r d[(t), (r)] + \alpha_s d[(t), (s)] + \beta d[(r), (s)]$$

with

$$\alpha_r = \frac{n_r + n_t}{n_r + n_s + n_t},\ \alpha_s = \frac{n_s + n_t}{n_r + n_s + n_t},\ \text{and}\ \beta = \frac{-n_t}{n_r + n_s + n_t}\ .$$

Given the preferred number of clusters k, the linkage tree is cut at the appropriate level to yield k clusters.

Cluster coherence

As a measure of textual coherence, C_G, we calculate the median distance in term space from the profile of the group G of size n_G to the individual profiles, g_i, of all genes in that group:

$$C_G = med_i\left\{dist\left(g_i, \tilde{g}\right)\right\}_{i=1..n_G} \quad with\ \ \tilde{g} = avg\left\{g_i\right\}_{i=1..n_G}. \qquad (1)$$

We assess its significance by computing a background distribution from random gene clusters of different sizes.

To demonstrate how Equation (1) scores groups of functionally related genes, we show its performance on 10 cell-cycle groups of Spellman *et al.* [37]. These involve 126 genes in total, which are identified manually as well as by expression

Table 3

Significance of coherence score C_G

Gene groups	Size	Coherence score
Cell-cycle control	19	1.01E-167
DNA repair	3	3.91E-61
Fatty acids/lipids	25	4.28E-08
Glycosylation	7	6.29E-06
Methionine	5	9.88E-28
Mitotic exit	9	1.50E-82
Nutrition	19	1.76E-18
Pseudohyphae	10	2.79E-05
Secretion	13	1.11E-06
Sporulation	16	1.11E-01

The significance is calulated with respect to 100-fold randomization for 10 cell-cycle related, functional groups selected from Figure 7 in Spellman *et al.* [37]. All groups are functionally coherent according to our score, except for the sporulation group.

analysis. As can be seen in Table 3, all but the sporulation group display *p*-values below the 1-sided 0.025 threshold (that is, a gene group *G* is considered coherent if C_G is smaller than expected by chance). A more detailed analysis can be found in [38], but falls outside the scope of this manuscript. This result corroborates the ability of Equation (1), and more importantly of the vector-space model that underlies TXTGate, to represent biologically relevant functional information. It provides a quantitative foundation that supports the underlying methodology of TXTGate.

TXTGate summarizes and identifies subclusters

TXTGate allows online subclustering and profiling of gene groups via terms extracted from MEDLINE. Below we describe two examples.

Yeast data

We took the reference data set from Eisen *et al.* [39] and used TXTGate to conduct a textual analysis similar to that of Blaschke *et al.* [16]. In Table 4 we show the text profiles of cluster *E* from Eisen *et al.* by subclustering with $k = 2$. Although several of the text-mining settings in Blaschke *et al.* are different from ours (because of the differences in MEDLINE corpus, textual analysis methodology, and the clustering algorithm used), a comparison of the term profiles in both analyses shows that TXTGate also identifies *E1* as being related to glycerol, whereas *E2* is more related to pyruvate metabolism and ethanol fermentation (for more details, see Blaschke *et al.* [16]). Detailed text profiles for each of the clusters {*B*, *C*, *D*, *E*, *F*, *G*, *H*, *J*, and *K*} in Eisen *et al.* are given in Additional data file 1.

Human data

To assess the quality of the indexed MEDLINE abstracts used in LocusLink, we compare the output from TXTGate with results presented in Chaussabel and Sher [6], where the authors describe, among other experiments, the profiling and clustering of nearly 200 genes involved in the 'common transcriptional program' induced in human macrophages upon bacterial infection. We interpreted the results by retrieving the MEDLINE textual profiles of all genes in the clusters and compared TXTGate's best-scoring terms to the cluster terms in Chaussabel and Sher [6]. The results of the first four (non-overlapping) clusters (clusters *a*, *b*, *c* and *d*) can be found in Table 5. The terms 'adipose', 'metastasis' and 'NM' did not show up in the profiles from TXTGate because they are not

Table 4

TXTGate profiling of cluster E from Eisen *et al.* [39]

Gene symbol		Cluster terms in Blaschke *et al.* [16]	Terms from TXTGate
Subcluster E1	*TPT1 FBA1*	glyceraldehyde-3-phosphate*	glyceraldehyd_3_phosphat_dehydrogenas
	GPM1 TKL1	glyceraldehyde-3-phosphate dehydrogenase*	glycolyt
	PGK1 CDC19	phosphoglycerate kinase*	glucos
	TDH3 HXK2	phosphoglycerate*	enzym
	TDH2 TYE7	mutase*	glycolysi
	ENO2 PFK1	dehydrogenase	carbon
	TDH1 ACS2	enolase	pyruv_kinas
		glycerol-3-phosphate dehydrogenase	ethanol
		osmotic stress	phosphoglycer_kinas
		phospoglycerate	growth
Subcluster E2	*PDC5 PDC1*	alcohol*	pyruv_decarboxylas
	PDC6	transketolase*	pyruv
		catabolite repression	glucos
		decarboxylase	enzym
		ethanol	alcohol
		glucose	decarboxyl
		glucose repression	ethanol
		hexokinases	ferment
		pyruvate	thiamin
		pyruvate decarboxylase	decarboxylas

Profiling is by subclustering ($k = 2$). High-scoring terms are shown for each subcluster E1 and E2. We also show the terms (excluding gene names) resulting from a similar analysis conducted by Blaschke *et al.* [16]. *Terms that were labeled specific to a subcluster by Blaschke *et al.* Although several of their settings are different from ours (because of the differences in MEDLINE corpus, textual analysis and the cluster algorithm used), a comparison of the term profiles in both analyses shows that TXTGate also identifies E1 as related to glycerol, whereas E2 is more related to pyruvate metabolism and ethanol fermentation. Complete data can be found in Additional data file 1.

Table 5

TXTGate profiling of clusters a, b, c, and d from Chaussabel and Sher [6] (GO vocabulary)

Gene symbol		Cluster terms in [6]	Terms from TXTGate	
Cluster a	*LPL*	Lipoprotein		**lipoprotein**
	CD36L1	Density		lipas
	LDLR	Cholesterol		ldl
		Lipid		ldl_receptor
		Adipose		**cholesterol**
				hdl
				scaveng_receptor
				high_densiti_lipoprotein
				low_densiti_lipoprotein_receptor
				low_densiti_lipoprotein
Cluster b	*UPA*	Invasive	Collagenase	**metalloproteinas**
	PLAUR	Invasion	Collagen	**matrix**
	SERPIN	Metastasis	Matrix	metalloendopeptidas
	MMP1	UPAR	MMP	**collagenas**
	MMP10	UPA	Metalloproteinase	extracellular_matrix
	MMP14	Plasminogen	Molecule-1	alpha
	SPARC	Urokinase-type	Adhesion	**upar**
		Urokinase	Vascular	**plasminogen_activ**
		Plasmin	Endothelial	interstiti
		Activator		**invasion**
Cluster c	*AMPD3*	Adenosine		purinerg
	ADA	A2A		**adenosin**
	ADORA2A	A1		deaminas
	ADORA3	Antagonist		p2
	P2RX	Agonist		p2x
	P2RX1	NM		p1
	P2RX7			**agonist**
				receptor
				adenosin_receptor
				ada
Cluster d	*IP10*	Interferon		tumor_necrosi_factor
	MIP1A	IFN-alpha		cytokin
	MIP1B	IFN		induc
	IL8	Interferon-gamma		**interferon**
	STAT4	IFN-gamma		inflammatori
	IL12B	Inducible		antigen
	TNFRSF9			lymphocyt_activ
	TNFSF9			stimul
	SLAM			chemokin
	TNFRSF5			monocyt
	CD83			

Corresponding terms in Chaussabel and Sher [6] and TXTGate are in bold. TXTGate's profiles are comparably informative. Complete data can be found in Additional data file 2.

Table 6

Comparison of the terms in cluster e found by Chaussabel and Sher [6] with those found by TXTGate (OMIM vocabulary)

Gene symbol	Cluster terms in Chaussabel and Sher [6].	Terms from TXTGate
Cluster e		
CKB	Population	deaminas
AMPD3	Frequency	**lipoprotein_lipas**
ADA	Allele	creatin
ADORA2A	Unrelated	lipoprotein
ADORA3	Families	**krabb**
P2RX	Recessive	**epidermolysi_bullosa**
P2RX1	Autosomal	**alagil**
P2RX7	Disorder	bear
GEM	Severe	leukodystrophi
ARHH	Patient	receptor
LPL	Deficiency	down
CD36L1		**corneal_dystrophi**
LDLR		**deaf**
BF		hdl
GALC		nucleosid
LAMB3		retinoblastoma
GJB2		junction
TGFBI		adhesion
JAG1		congenit_heart_defect
DSCR1		**hear_loss**

The diversity of the diseases the member genes are related to makes the relevant terms display high variance, rather than high mean. The terms that were also found by Chaussabel and Sher [6] after manual investigation are marked in bold. Complete data can be found in Additional data file 2.

contained in the GO domain vocabulary. For cluster *e* no common terms were found. Running TXTGate using the OMIM vocabulary, however, we were able to uncover exactly those disease-associated terms that were retrieved by Chaussabel and Sher [6] by manually investigating genes from this cluster in the OMIM database. In Table 6 we highlight these terms in bold. As the set of diseases related to these genes is heterogeneous, the relevant terms display a high variance, rather than a high mean, a reason for also including a variance profile. Moreover, the fact that we retrieve those disease terms only by means of the OMIM vocabulary points out that the use of a variety of vocabularies in TXTGate leads to improved insights, a point discussed further in the next section. We note that all other cluster terms have a comparable equivalent in the TXTGate profiles; the complete analysis is given in Additional data file 2.

Textual information through the eyes of different vocabularies

Another major feature of TXTGate is its ability to present textual information (most importantly MEDLINE abstracts) from different perspectives. This is implemented by offering indices built on GO-, OMIM-, MeSH-, eVOC-, and gene nomenclature-based domain vocabularies respectively. Each configuration is meant to expose a different view of the literature. TXTGate mirrors the dual approach adopted by the external databases it links to, which separate keyword and gene-symbol queries. This, in part, motivated our strategy to construct both term- and gene-centric vocabularies.

Table 7

Various perspectives on textual information in TXTGate

GO	OMIM	MeSH	eVOC
mismatch_repair	colorect	colorect_neoplasm	colorect
tumor	colorect_cancer	mismatch	tumour
dna_repair	tumor	cancer	malign_tumour
mismatch	kinas	colorect	colon
pair	colon	mutat	growth
tumor_suppressor	hereditari	repair	cell
apc	cancer	dna_repair	carcinoma
kinas	colon_cancer	colon	metabol
somat	associ	neoplasm_protein	fibroblast
ra	on	tumor	chain

Here we show how term-centric vocabularies based on GO, OMIM, MeSH and eVOC profile a group of genes involved in colon and colorectal cancer.

To compare our term-based vocabularies we profiled a group of genes involved in colon and colorectal cancer extracted from the OMIM Morbid Map database (see Additional data file 3). Table 7 shows the top 10 terms for each of the retrieved profiles. As can be seen, there is little difference between the MeSH and OMIM profiles, whose terms are mainly medical- and disease-related ('colorect_cancer', 'colon_cancer', 'colorect_neoplasm', 'hereditari'), whereas the scope of the GO profile is focused more on metabolic functions of genes ('mismatch_repair', 'dna_repair', 'tumor_suppressor', 'kinas') and the eVOC profile contains terms more related to cell type and development ('growth', 'cell', 'carcinoma', 'metabol', 'fibroblast'). TXTGate's link-out feature allows a more profound analysis of the retrieved terms. Top-ranking terms can be sent to PubMed to retrieve relevant publications. Because all MEDLINE entries are tagged with MeSH keywords, using terms from the MeSH vocabulary assures a successful query. When using the GO-derived vocabulary, terms can be mapped back directly to the GO tree with AmiGO [40] to investigate the term's neighborhood. Other databases available for querying include LocusLink and OMIM.

We used the same colon cancer case to test the ability of our human gene symbol vocabulary in screening for colinkage of genes. We constructed two different index tables - one with

Table 8

Co-linkage analysis of genes with gene-centric vocabularies

Gene name	Description
hnpcc	Hereditary nonpolyposis colon cancer
apc	Adenomatous polyposis coli protein
p53	Cellular tumor antigen P53 (tumor suppressor P53)
mlh1	DNA mismatch repair protein MLH1 (mutL protein homolog 1)
muts	E. coli mismatch repair gene mutS
p21	Cyclin-dependent kinase inhibitor 1A
msh2	DNA mismatch repair protein MSH2 (mutS protein homolog 2)
bax	BAX protein, cytoplasmic isoform delta
wnt	Wingless-type MMTV integration site family members
pms2	DNA mismatch repair protein PMS2
src	Proto-oncogene tyrosine protein kinase SRC
dcc	Tumor suppressor protein DCC precursor (colorectal cancer suppressor)
mcc	Colorectal mutant cancer protein MCC
braf	Proto-oncogene serine/threonine protein kinase B-RAF
fgfr3	Fibroblast growth factor receptor 3 precursor
hcc	Hepatocellular carcinoma
dra	Chloride anion exchanger DRA
axin2	AXIS inhibition protein 2
pms1	DNA mismatch repair protein PMS1
abl	Abelson murine leukemia viral oncogene homolog 1
bub1	Mitotic checkpoint serine/threonine protein kinase BUB1
ptp	Protein tyrosine phosphatase family
bcl10	B cell lymphoma/leukemia 10
ptp_pest	Protein tyrosine phosphatase family with C-terminal PEST-motif
prlts	PDGF-receptor beta-like tumor suppressor

This table shows the top-25 colinked gene symbols in the pool of abstracts of the colon and colorectal cancer case. Genes that were not in the query list are indicated in bold.

and one without alternative gene symbols; the former was constructed by mapping all synonymous symbols to the primary gene symbol. The first table has the disadvantage of not being able to disambiguate alternative gene symbols that are mapped to different primary gene symbols; the second does not take synonyms into account, as only true occurrences of a symbol were counted. As a consequence, frequently used symbols are ranked highly, while not being the official gene symbols. Examples of this are p21 and dra, whose primary symbols are CDKN1A and SLC26A3, respectively. The top-25 gene symbols using the first index table are given in Table 8. Most of the retrieved gene names are also in the query list. We used TXTGate's link-out feature to investigate the role of the genes that were not in the input list by sending them as a query to LocusLink and GeneCards. This way we were able to determine their function and their relation to colon and colorectal cancer, as can be seen in Table 8.

Application of TXTGate to a real-life research problem

In the framework of an ongoing collaboration with a medical research group, our system was deployed to tackle a current research issue [41,42]. We analyzed 350 genes that were upregulated in a mouse model for human benign tumors of the salivary glands and evaluated the results in a biological context. We had a medical researcher write a summary of pathological and genetic observations, reflecting relevant literature and expert knowledge. From this we derived a list of important terms. This list was cross-referenced with textual profiles retrieved from TXTGate using different domain vocabularies (see Additional data file 4). As pathology and developmental issues were the focus of the summary in this case, the eVOC domain vocabulary proved most appropriate, as can be seen from the occurrence of terms such as 'fibroblast', 'embryo', 'tumor', 'teratoma' and so on (see Table 9).

Table 9

Textual profile of a gene group from a mouse model for human benign tumors of the salivary glands

Terms sorted by mean	Terms sorted by variance
organ	organ
intern	intern
normal	growth
red	development
male	fibroblast
femal	tumour
visual	red
capillari	nucleu
system	normal
optic	embryo
retina	tera
viral	depend
bacteri	stem_cell
adult	kidnei
chain	epithelium
cell	visual
growth	multipl
tissu	skin
development	muscl_cell
metabol	system
embryo	capillari
fibroblast	mammari
tumour	type_ii
depend	bacteri
genet	male

This table shows the 25 top-ranking terms (for both mean and variance) of the textual profile of a group of 350 genes that were upregulated in a mouse model for human benign tumors of the salivary glands processed with the eVOC domain vocabulary.

We can conclude that the choice of domain vocabulary depends on the experimental context and focus of the investigation. This supports our strategic choice of offering different domain vocabularies.

As a measure of textual coherence C_G, we calculated the median distance in vocabulary space from the profile of the group G to the individual profiles g_i of all genes in that group:

$$C_G = med_i\{dist(g_i, \tilde{g})\}_{i=1..350} \quad with \quad \tilde{g} = avg\{g_i\}_{i=1..350}. \qquad (2)$$

As background we generated 5,000 random gene clusters of both the same size and random sizes (see Figure 2), and calculated their coherence as in Equation (2). We derived two background distributions modeling the information content for random clusters. This allows the calculation of a p-value for a cluster of genes, expressing the probability that the observed textual coherence occurs by chance. The cluster

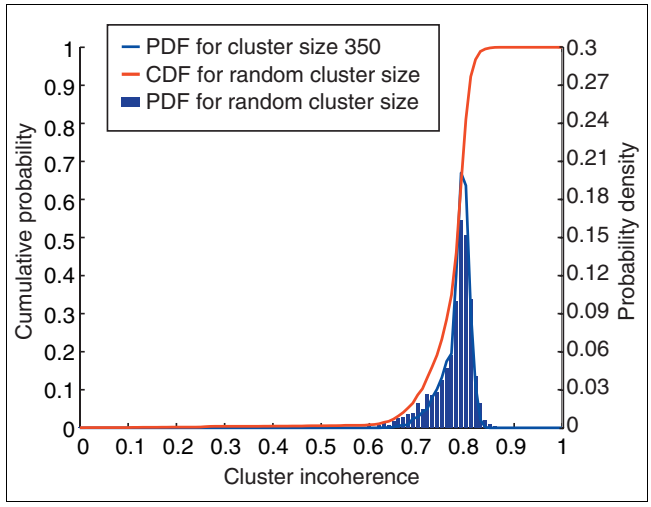

Figure 2
Background distributions for cluster incoherence. Cluster incoherence is defined as the median distance in vector space between the mean cluster profile and all individual gene profiles. Probability density functions (pdf) are shown for random clusters of size 350 (blue curve) and random clusters of random size (blue bars). For randomly sized clusters, the cumulative distribution function (cdf) is also shown (red curve).

profile of the 350 upregulated mouse genes was significant against both the background for random cluster size (p-value 1.8×10^{-3}) and for cluster size 350 (p-value $< 10^{-8}$).

Discussion

We have described a framework for advanced textual profiling of groups of genes. TXTGate is implemented as a web application designed to efficiently process queries of up to 200 genes, although this is not a strict limit. We believe that the application scales well enough to be of use in, for example, microarray cluster validation.

Supported by the work of Stephens *et al.* [43] and more recently that of Chiang and Yu [28], we aimed to complement the limitations of a single, more general, text index by offering different views. Nevertheless, some vocabularies could still be optimized to improve the information content of the profiles. For example, some general or non-informative terms are still scoring high because of our stemming and phrase-detection methods (for example, 'ii', 'protein', 'alpha').

Finally, although the citations in LocusLink and SGD constitute good sources for retrieving relevant gene-related MEDLINE abstracts, weighting the information according to the context and eliminating poorly informative or contaminating annotations (such as sequence-related articles) still need to be taken into account in future incarnations of the software. Document-classification strategies as in Leonard *et al.* [9] or Raychaudhuri *et al.* [10] can be adopted to this end.

As with GO annotations, transfer of literature references according to homology can be used to characterize poorly annotated genes [44,45]. At this stage, the application allows for the study of homologs within all organisms contained in LocusLink, provided the user inputs the corresponding LocusLink identifiers. This type of operation will be increasingly supported with future additions of literature indices from other organisms and databases.

In conclusion, TXTGate's approach to summarizing database annotations and literature via specific vocabularies, along with its options to perform further analysis via clustering or query building, make it a flexible gateway to explore text-based information comprehensively.

Additional data files

The following additional data are available with the online version of this article: the MEDLINE-based text profiles of yeast expression clusters from Eisen *et al.* [39] (Additional data file 1); the MEDLINE-based profiles for the data in Chaussabel and Sher [6] (Additional data file 2); details on the colon and colorectal cancer test case (Additional data file 3); the expert summary and textual profiles of the 350 upregulated mouse genes for different domain vocabularies (Additional data file 4).

Acknowledgements

This research was supported by grants from the Research Council K.U. Leuven (GOA-Mefisto-666, GOA-Ambiorics, IDO), the Fonds voor Wetenschappelijk Onderzoek - Vlaanderen (G.0115.01, G.0240.99, G.0407.02, G.0413.03, G.0388.03, G.0229.03, G.0241.04), the Instituut voor de aanmoediging van Innovatie door Wetenschap en Technologie Vlaanderen (STWW-Genprom, GBOU-McKnow, GBOU-SQUAD, GBOU-ANA), the Belgian Federal Science Policy Office (IUAP V-22), and the European Union (FP5 CAGE, ERNSI, FP6 NoE Biopattern, NoE E-tumours). We acknowledge Peter Antal for starting up this research direction.

References

1. 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].
2. Pruitt K, Maglott D: **RefSeq and LocusLink: NCBI gene-centered resources.** *Nucleic Acids Res* 2001, **29:**137-140.
3. Masys DR, Welsh JB, Fink JL, Gribskov M, Klacansky I, Corbeil J: **Use of keyword hierarchies to interpret gene expression.** *Bioinformatics* 2001, **17:**319-326.
4. Jenssen T, Laegreid A, Komorowski J, Hovig E: **A literature network of human genes for high-throughput analysis of gene expression.** *Nat Genet* 2001, **28:**21-28.
5. Shatkay H, Edwards S, Boguski M: **Information retrieval meets gene analysis.** *IEEE Intell Syst (Special Issue on Intelligent Systems in Biology)* 2002, **17:**45-53.
6. Chaussabel D, Sher A: **Mining microarray expression data by literature profiling.** *Genome Biol* 2002, **3:**research0055.1-0055.16.
7. Glenisson P, Antal P, Mathys J, Moreau Y, Moor BD: **Evaluation of the vector space representation in text-based gene clustering.** *Pac Symp Biocomput* 2003:391-402.
8. Raychaudhuri S, Schutze H, Altman RB: **Using text analysis to identify functionally coherent gene groups.** *Genome Res* 2002, **12:**1582-1590.
9. Leonard JE, Colombe JB, Levy JL: **Finding relevant references to genes and proteins in Medline using a Bayesian approach.** *Bioinformatics* 2002, **18:**1515-1522.
10. Raychaudhuri S, Chang JT, Sutphin PD, Altman RB: **Associating genes with Gene Ontology codes using a maximum entropy analysis of biomedical literature.** *Genome Res* 2002, **12:**203-214.
11. **Gene Ontology Consortium** [http://www.geneontology.org]
12. **Medical Subject Headings** [http://www.nlm.nih.gov/mesh/meshhome.html]
13. Kelso J, Visagie J, Theiler G, Christoels A, Bardien S, Smedley D, Otgaar D, Greyling G, Jongeneel C, McCarthy M, *et al.*: **eVOC: a controlled vocabulary for unifying gene expression data.** *Genome Res* 2003, **13:**1222-1230.
14. **Gene Ontology Annotation** [http://www.ebi.ac.uk/GOA]
15. **TXTGate Portal** [http://www.esat.kuleuven.ac.be/txtgate]
16. Blaschke C, Oliveros J, Valencia A: **Mining functional information associated with expression arrays.** *Funct Integr Genomics* 2001, **1:**256-268.
17. 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* 1999, **27:**1210-1217.
18. **MedMiner** [http://discover.nci.nih.gov/textmining]
19. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D: **GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support.** *Bioinformatics* 1998, **14:**656-664.
20. Calogero R, Iazzetti G, Motta S, Pedrazzi G, Rago S, Rossi E, Turra R: **MedMOLE: mining literature to extract biological knowledge by microarray data.** In *Proc Virtual Conf Genomics Bioinformatics* 2002, **2:**9-14.
21. **MedMOLE at CINECA** [http://www.cineca.it/HPSystems/Chimica/medmole]
22. **DNA Array Analysis with GEISHA** [http://www.pdg.cnb.uam.es/blaschke/cgi-bin/geisha]
23. **PubGene Gene Database and Tools** [http://www.pubgene.org]
24. Hu Y, Hines L, Weng H, Zuo D, Rivera M, Richardson A, LaBaer J: **Analysis of genomic and proteomic data using advanced literature mining.** *J Proteome Res* 2003, **2:**405-412.
25. **MedGene Database** [http://hipseq.med.harvard.edu/MEDGENE]
26. Perez-Iratxeta C, Bork P, Andrade M: **Association of genes to genetically inherited diseases using data mining.** *Nat Genet* 2002, **31:**316-319.
27. **G2D Candidate Genes to Inherited Diseases** [http://www.bork.embl-heidelberg.de/g2d]
28. Chiang J, Yu H: **MeKE: discovering the functions of gene products from biomedical literature via sentence alignment.** *Bioinformatics* 2003, **19:**1417-1422.
29. **MeKE (Medical Knowledge Explorer)** [http://ismp.csie.ncku.edu.tw/~yuhc/meke]
30. **Java Remote Method Invocation (Java RMI)** [http://java.sun.com/products/jdk/rmi]
31. Baeza-Yates R, Ribeiro-Neto B: *Modern Information Retrieval* Reading, MA: Addison-Wesley/ACM Press; 1999.
32. Porter MF: **An algorithm for suffix stripping.** *Program* 1980, **14:**130-137.
33. ***Saccharomyces*** **Genome Database** [http://www.yeastgenome.org]
34. **OMIM - Online Mendelian Inheritance in Man** [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM]
35. **HUGO Gene Nomenclature Commitee (HGNC)** [http://www.gene.ucl.ac.uk/nomenclature]
36. Jain A, Dubes R: *Algorithms for Clustering Data* Upper Saddle River, NJ: Prentice Hall; 1988.
37. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization.** *Mol Biol Cell* 1998, **9:**3273-3297.
38. Glenisson P, Mathys J, Moreau Y, De Moor B: **Scoring and summarizing gene groups from text using the vector space model.** *Technical Report 03-97, ESAT-SISTA* 2003 [ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/glenisson/ reports/genomebiol/TR03-97.pdf]. Leuven, Belgium: K.U.Leuven
39. Eisen M, Spellman P, Brown P, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95:**14863-14868.
40. **AmiGO Gene Ontology browser** [http://www.godatabase.org]
41. Kas K, Voz ML, Roijer E, Astrom AK, Meyen E, Stenman G, Van de

Ven WJ: **Promoter swapping between the genes for a novel zinc finger protein and beta-catenin in pleiomorphic adenomas with t(3;8)(p21;q12) translocations.** *Nat Genet* 1997, **15:**170-174.

42. Voz ML, Mathys J, Hensen K, Pendeville H, Van Valckenborgh I, Van Huffel C, Chavez M, Van Damme B, De Moor B, Moreau Y, Van de Ven WJ: **Microarray screening for target genes of the proto-oncogene PLAG1.** *Oncogene* 2004, **23:**179-191.
43. Stephens M, Palakal M, Mukhopadhyay S, Raje R, Mostafa J: **Detecting gene relations from Medline abstracts.** *Pac Symp Biocomput* 2001:483-495.
44. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.*: **Gene ontology: tool for the unification of biology.** *Nat Genet* 2000, **25:**25-29.
45. Raychaudhuri S, Chang JT, Imam F, Altman RB: **The computational analysis of scientific literature to define and recognize gene expression clusters.** *Nucleic Acids Res* 2003, **31:**4553-4560.