A Relational Approach to Defining Document Set Relevance: An Application in Human Genetics

IDI-rapport 7/00 ISSN 0802-6394

Tor-Kristian Jenssen [1], Staal Vinterbo [1]

Knowledge Systems Group, Dept. of Computer and Information Science, Norwegian University of Science and Technology, O.S. Bragstads Plass 2E, N-7491 Trondheim, Norway

Abstract

With the emerging cDNA microarray and oligonucleotide array technologies, it has become possible to study gene expressions of a large number of genes in a single experiment. Searching the literature for references for more than one gene by traditional methods is often time consuming. This particularly in the case where literature pertaining to relationships between genes is sought by, e.g., co-citation of gene labels. For this task, optimal Boolean search expressions cannot be determined a priori, and Boolean ad hoc search formulations often result in either none or far too many hits when using a conjunction or disjunction of search terms, respectively. We address this problem by formulating the search task as a family of set cover problems in an automatically generated gene label–document relation covering the complete MEDLINE database (up to October 1999). Four sets of gene symbols identified in a recent publication were input as search queries to our system. The results indicate that our method is practical and produces manageable sets of document that seem relevant.

Keywords: document relevance; set cover; genes; genetics; literature search; MEDLINE

1 Introduction

Recent advances in biotechnology have made it possible to perform near genome-wide measurements of gene expression levels. Highly parallel methods such as cDNA microarrays [1, 2] and the Affymetrix oligonucleotide arrays [3] have initiated a shift in biomedical research. Researchers are now able to do experiments that involve thousands of genes at the same time, while earlier techniques only allowed gene expression studies with a few genes. The large number of genes that can be screened represents a tremendous opportunity but also a challenge. A single researcher can only have detailed knowledge of a limited number of genes. As more and more documents are written about each gene, it becomes difficult to keep up-to-date on the literature for even a single gene. The interpretation of such data will therefore require extensive literature searches in order to find information about genes outside one's own field of expertise.

The MEDLINE citation database from the National Library of Medicine (NLM) is a natural choice when searching for biomedical literature. MEDLINE is extensively indexed through the MeSH (Medical Subject Heading) system, and most central journals in the field are covered. The MeSH index is useful for qualifying searches beyond free-text search, but due to the general nature of MEDLINE and the MeSH index, in addition to the facts that the MeSH index is built manually and

[1] Equal authorship implied

that the gene nomenclature is constantly revised, all gene-specific information cannot be included in the MeSH thesaurus. In particular, many recent genes are not included in the MeSH system and give no hits in MEDLINE when using MeSH searches. Some genes are thoroughly indexed, and some of the vast literature for, e.g., the TNF gene, is easily retrieved by the query 'TNF [MESH]'. When searching for documents for a single gene, unconstrained search queries frequently result in lists with several thousand documents. A search in MEDLINE with 'TNF' returned more than 35000 documents. Within a specific biomedical context, the search can be narrowed down by qualifying the search by requiring additional keywords relevant to a specific context (disease, physiological process, etc.) to be present. This is difficult if such a context is not known *a priori*, which often may be the case in the high-throughput screening experiments allowed by the array technologies.

Gene-expression studies will often reveal possible interactions between two, or more, genes. Instead of having an *a priori* hypothesis about a biological context, a literature search may then be qualified by searching for literature describing interactions between the genes in question. Unfortunately, such literature searches are not very well supported by conventional search methods. The standard Boolean operators provided by MEDLINE, and similar databases, are not appropriate for this kind of search. Combining search terms with the AND operator is likely to result in zero documents even for as few as 2 or 3 genes, and also even if each gene separately has many references. The OR operators give even more documents than single-gene search.

Much information retrieval and document relevance research focuses on the relevance of single documents to a set of requirements. From an abstract point of view this corresponds to finding a function from a set of documents to some value set. Most approaches use an ordered value set as ranking of documents is considered as being "at the core of information retrieval systems." [4]. Even unordered sets of documents will be considered as ordered from a perspective of the user [5]. Such a view of a function from a set of documents to a relevance value set fits nicely into a classification framework where the objective is to construct such a function called a classifier. Classically, information retrieval models can be classified into probabilistic, Boolean and vector based. Some extended approaches incorporate natural language processing to cope with semantical problems associated with homonyms and synonyms [6], others again are concerned with incorporating user feedback [7, 8]. If, on the other hand, the objective is to rank the relevance of *sets* of documents where there is an assumed synergy effect on relevance achieved by elements in each set, a document ranking approach is not well suited. An approach that takes this into account is by Syu and Lang [9]. This approach is related to ours in that it also uses a covering formulation of article relevance and uses a mechanism for handling synonyms.

We have developed a method for searching for literature about a given list of human genes. Conceptually, the approach can be described as a two step procedure. The first step is to compute a gene-document relation by analyzing titles and abstracts from MEDLINE records. The relation, covering the complete MEDLINE from 1966 up to present, is compiled by an automated pattern matching procedure. When receiving a query, the search is formulated as a minimal set-covering problem computed on the pre-computed relation. The system is flexible in that the user can adjust the number of documents to retrieve as well as tune parameters defining relevance used when ranking documents. A prototype system can be accessed over the Web [10].

The conducted experiments, described in Section 3, indicate that the set-covering formulation on the document-gene relation is more suitable for retrieving manageable sets of relevant documents than using standard Boolean operators.

2 Methods

Conceptually, our approach may be de-composed into two distinct and independent steps: 1) construction of a bipartite graph associating human genes with documents (MEDLINE document

records), and 2) computation of set covers, in the form of document sets covering a given input set of genes, defining relevant literature references. The first step creates a common basis to all subsequent searches, whereas the second step is particular to each query.

2.1 Construction of the Association Graph

The construction of the gene-literature index may be further decomposed into five distinct substeps:

1. Collecting data into a gene database, in particular the gathering of gene symbols and names, collectively referred to as gene labels.
2. Composing a list of primary symbols, essentially defining the list of gene records in the database.
3. Associating every label to a gene record.
4. Indexing the MEDLINE documents with respect to occurrences of gene labels, as found by string matching in titles or abstracts.
5. Computing the gene-literature index by gathering references from labels to documents in order to define the list of references for that gene.

A brief word on formal definitions and notation is expedient at this point. Let A and G be two sets, and let $R \subseteq A \times G = \{(a,g) | a \in A, g \in G\}$ be a *relation from A to G*. For two sets $A' \subseteq A$ and $G' \subseteq G$, we define:

$$\begin{aligned} R(A') &= \{g \mid a \in A' \wedge (a,g) \in R\}, \text{ and} \\ R^{-1}(G') &= \{a \mid g \in G' \wedge (a,g) \in R\} \end{aligned}$$

as the *image* of A' under R, and the *preimage* of G' under R, respectively. As a notational convenience, we will use $R(a)$ for $R(\{a\})$ and $R^{-1}(g)$ for $R^{-1}(\{g\})$ as abbreviations for singleton sets. We will call R *separating* if $R(a) \cap R(a') \neq \emptyset$ implies $a = a'$. Furthermore, if $|R(a)| \leq 1$ for all $a \in A$, R is called *single valued*. Also, R is called *injective* if it is separating and single valued, *surjective* if $R(A) = G$, *total* if $R^{-1}(G) \supseteq A$, and *partial* if it is not total. A total, injective and surjective relationship is called a *bijection*.

Returning to the problem at hand, letting A denote the set of documents available, and G the set of genes of interest, we desire a relation R from A to G that associates with each document the genes mentioned in it.

Several practical difficulties associated with the representation of genes, and construction of R crop up. Let L denote the union of gene labels found in public gene databases, and let $L' \subseteq L$ be a set of primary gene labels such that there exists a bijection between G and L'. The existence of this bijection enables us to put a unique label from L' on each gene. The public gene databases induce a natural relation G_L from G to L. This relation is defined by $(g,l) \in G_L$ if the label l has been related to gene g in these databases.

Now, let $L_L \subseteq L \times L$ be the relation defined by

$$L_L = \{(l_1, l_2) \mid G_L^{-1}(l_1) \cap G_L^{-1}(l_2) \neq \emptyset\}.$$

Intuitively, L_L relates labels that refer to the same gene. For a label $l \in L'$, $L_L(l)$ is the set of all labels that refer (among others) to the same gene as l. We are now able to formulate an approximation R' of the relation R. Define the relation $R' \subseteq A \times L'$ by

$$R' = \{(a,l) \mid a \in L_A(L_L(l))\}$$

where $L_A \subseteq L \times A$ is the relation relating gene labels to the documents in which they occur. The relation L_A is computed by an automated document indexing procedure.

Aspects of material presented in this section will be further discussed in Section 4.2.

2.2 Document Selection by Set Covering

Let R be the relation from A to G presented above. A set $A' \subseteq A$ is said to *cover* a set $G' \subseteq G$ if and only if $G' \subseteq R(A')$. A formulation of the literature search problem in terms of a set covers is presented together with a greedy algorithm. Aspects of the material presented in this section will be discussed in Section 4.1

2.2.1 Measures of Relevance

Several measures can be associated with two fixed sets $A' \subseteq A$ and $G' \subseteq G$ using a relation R from A to G:

approximation As a measure of degree of cover of a non-empty G' by A' we define:

$$a_R(A', G') \quad = \quad |R(A') \cap G'|/|G'|.$$

overshoot As a measure of elements in a non-empty $G - G'$ that are covered by A', we define:

$$o_R(A', G') \quad = \quad a_R(A', G - G').$$

separation As a measure of how well the elements of the cover of G' given by A' are separated, we define:

$$s_R(A', G') \quad = \quad \sum_{g \in G'} |R^{-1}(g) \cap A'|.$$

This measure has range $\{i, \ldots, s\}$, where $i = |R(A') \cap G'|$, and $s = |R(A') \cap G'||R^{-1}(G') \cap A'|$. Maximizing the separation of the cover is achieved by minimizing s_R.

connectivity A measure related to how the cover of G', by the elements of A', overlaps, is the mean of all pairs shortest path in the graph on G' where there is an edge between two elements if they are covered by a single element of A', is used. Let $G' = \{g_i \mid i \in I \subseteq \mathbb{N}\}$, and define $P : G' \times G' \to 2^{2^I}$ as

$$P(g, g') = \{(i_1, i_2, \ldots, i_l) \mid i_j \in I, g = g_{i_1}, g' = g_{i_l}, \exists\, a \in A'\ g_{i_j}, g_{i_{j+1}} \in R(a)\},$$

$l : G' \times G' \to \mathbb{Z}^+$ as

$$l(g, g') = \begin{cases} \min_{s \in P(g,g')} |s| - 1 & \text{if } P(g, g') \neq \emptyset \\ |G'| & \text{otherwise,} \end{cases}$$

and

$$c_R(A', G') \quad = \quad \frac{\sum_{p \in G' \times G'} l(p)}{|G' \times G'|}.$$

The connectivity measure c_R is then the mean all pairs shortest path in the graph described above. The range of c_R is $[1, |G'|]$, where the supremum indicates that no two elements in G' are covered by the same element, and the infimum indicates that any pair $(g, g') \in G' \times G'$ is related by a single element in A'.

The measures of approximation and overshoot both have the unit interval as range. Normalization of the separation and connectivity measures can be achieved by applying the transformation given by:

$$\frac{x-i}{s-i},$$

where x is the measure value, s and i are the supremum and infimum of the range for that measure, respectively.

Note that even though the measures s_R and c_R are related, there is a difference. Maximizing c_R will minimize s_R, but the converse is not true. The two measures also achieve minimum in the case where the cover of G' contains one element that covers G' by itself.

2.2.2 Applying Set Covering to Literature Search

For a given R, $h \in [0,1]$, and a set of genes G', let $\mathcal{C}_h = \{A' \subseteq A | a_R(A', G') \geq h\}$. The collection $\mathcal{C}_h$ contains exactly all answers to the question of which subsets of A cover G' with a degree at least h. For $\delta_{\{\beta,s,o,c\}} \in \mathbb{R}^+ \cup \{0\}$, the minimization of $m_R : 2^A \times 2^G \to \mathbb{R}$ on $\mathcal{C}_h$, where m_R is given by

$$m_R(A', G') = \delta_\beta \beta(A') + \delta_s s_R(A', G') + \delta_o o_R(A', G') + \delta_c c_R(A', G'),$$

where β is a cost function on 2^A, encompasses a family of optimization problems. Among these are [11, 12]: The minimal cover problem, by requiring $\beta(A') = |A'|$, $\delta_\beta = 1$, $\delta_{s,o,c} = 0$, and $h = 1$, and the exact cover problem by requiring $\delta_s = 1$, $\delta_{\beta,o,c} = 0$, and $h = 1$.

Using m_R as a measure of relevance we have now formulated our literature search problem as a set cover problem.

2.2.3 A Sequential Greedy Algorithm

A sequential approach to set covering can be thought of as a stepwise forward or backwards selection approach, where single elements are scheduled for addition to, or removal from, the resulting set conditional on what has been done in the preceding steps. A sequential forward greedy algorithm that returns n disjoint covers of degree at least h can be illustrated by the algorithm skeleton in Figure 1.

This algorithm is based on an auxiliary utility measure on elements of A,u, that can be defined as:

$$\begin{aligned} u(a, A') &= (\delta_\beta \beta(a) + \delta_s \Delta_s(a) + \delta_o \Delta_o(a))(\delta_a(1 - \Delta_a(a)) + \delta_c(1 + \Delta_c(a))) \\ \Delta_*(a) &= *_R(A' \cup \{a\}, G') - *_R(A', G') \end{aligned}$$

The function Δ_* gives the change in the measure $* \in \{s, o, a, c\}$ defined above, each weighted by the corresponding $\delta_{\{s,o,c\}}$.

The algorithm returns a list of covers, sorted ascending on u. As measures $\Delta_{\{s,o\}}$ are independent of the increase in cover a particular element adds, they are multiplied with the measures $\Delta_{\{a,c\}}$ in the greedy utility function. This is done to avoid the possibility of adding a document just because it minimizes $\Delta_{\{s,o\}}$ while in reality not contributing to neither cover nor connectivity.

To use D-Set-Covers to search for a relevant set of documents for a set $L'' \subseteq L'$ of primary gene labels, we can search for the relevant set of documents by calling D-Set-Covers(R',L'',u,h,n) with appropriate parameters.

D-SET-COVERS(R,G',u,h,n)
Input: $R \subseteq A \times G$, $G' \subseteq G$, $u: A \times 2^A \to \mathbb{R}$, $h \in [0,1]$, $n \in \mathbb{N}$
Output: a list l of subsets of A
$l \leftarrow [\,]$
$T \leftarrow R^{-1}(G')$
while $n > 0$
  $A' \leftarrow \emptyset$
(1) $c \leftarrow 0$
(2) **while** $T \neq \emptyset$ **and** $a_R(A', G') < h$ **and** $a_R(A', G') > c$
(3) $c \leftarrow a_R(A', G')$
    $k \leftarrow \arg\min_{a \in T} u(a, A')$
    $T \leftarrow T - \{k\}$
    $A' \leftarrow A' \cup \{k\}$
(4) **if** $a_R(A', G') < h$
    return l
  $l \leftarrow \mathit{append}(A', l)$
  $n \leftarrow n - 1$
return l

Figure 1: The D-SET-COVERS algorithm.

3 Experiments

The objective of the experiments was to elicit the ability of our method to find relevant documents for relationships between the genes represented by a given set of gene labels.

The measure of success was constructed by a comparison of results obtained by our method and those obtained by using the standard MEDLINE interface as a baseline to compare against.

3.1 Measures

There exist at least two types of information about relationships between genes that can be found using literature search:

- Information given by the semantic content of the documents for which references were returned, and
- Information contained in a co-occurrence relation.

As the former requires domain specific knowledge and cannot be fully evaluated automatically, and thus cannot be incorporated into an automatic search strategy, we chose to assume that a document contained such information about a relationship between a set of genes if references to all elements in this set could be found in this document. This reduces the former type of information into a sub-type of the second.

The measures used to quantify the results from the searches were

- the number of documents returned by each search,

- the degree of cover of the gene symbol set,
- the connectivity of the co-occurrence relation this set of documents induced, and
- the specificity of the result with regard to the set of gene symbols, as given by the overshoot measure.

3.2 Data Material

Based on Iyer et al. [13], describing the response of human fibroblasts to serum, we picked 4 sets of genes corresponding to 4 functional categories. The categories were Signal Transduction, Immediate-Early Transcription Factors, Inflammation, and Angiogenesis. The sets are not disjoint as some of the included genes have pleiotropic roles. A few 'genes' from each category could not be related to any primary symbol and were excluded from the study. Table 1 contains the groups and their gene symbols.

The collection of documents used was the collection of articles indexed in MEDLINE up to October 1999.

3.3 Experiments and Results

For each of the groups we initiated a search using the symbols in this group, using the Boolean combination search method offered in MEDLINE and our method.

3.3.1 MEDLINE Search

We restricted the search to occurrence of the primary symbol in either the title or the abstract, as in our approach. Thus,for all symbols in the union of all groups' symbols, we searched for all articles matching this symbol in either the title or abstract.

For all groups we searched articles matching for the conjunction (AND'ing all symbols) of all symbols, and the disjunction of all symbols. We also searched for articles matching any conjunction of two symbols from within each group.

Due to the semantics of the AND operation, including search terms resulting in 0 hits alone, will trivially force the AND'ed query to result in 0 hits. We therefore decided to omit the genes resulting in 0 hits from the AND-combinations. Furthermore, due to the rapid changes in the terminology, several gene symbols used in the referred article have already been out-dated. In such cases, we used the OR-combination of the new and the old symbol rather than any of the symbols alone.

Table 1 shows the number of articles returned by MEDLINE for each of the individual genes in the four sets.

Table 2 shows the number of articles returned by MEDLINE for each of the four sets when the search terms were combined using the Boolean operators AND and OR.

3.3.2 Set-Cover Search

We constructed the relations G_L, L_L, L_A and R' as described in Section 2.1, the automatic indexing procedure consisting of standard pattern matching in document titles and abstracts. The resulting relation R' was between 1405363 documents and 8986 gene labels. From the symbols for the four groups, the primary symbols were (automatically) identified and substituted for the non-primary.

Set/Gene	#
Signal Transduction	
RORA (ROR1)	9 (14)
KIT	8940
TGFBR3	62
BMPR2	1
PTPN12	-
DUSP7 (PYST2)	1 (3)
EDG1	6
SLC6A5 (NET1)	1 (-)
LTBP3 (LTBP2)	- (3)
SGK	37
DUSP1 (MKP1)	- (5)
Immediate-Early Transcription Factors	
ATF3	22
EGR1	99
DEC1	6
TIEG	14
NR4A3 (MINOR)	0 (74546)
JUNB	543
COPEB/KLF1 (CPBP/EKLF)	2/2 (18/69)
ID3	1649
ID2	95
NFIL3 (NF-IL3A)	4 (3)
ETS2	83
Angiogenesis	
PTGS2 (COX2)	9 (120)
SCYA2 (MCP1)	12 (914)
IL8	3864
IL1B	20
FGF2	646
SDF1	23
VEGF	2915
MME (CALLA/CD10)	95 (1973/2244)
Inflammation	
PTGS2 (COX2)	9 (120)
IL8	3864
GRO2 (MIP2A)	3 (-)
ICAM1	57
IL1B	21
IL6	12033
SCYA2 (MCP1)	12 (914)
SDF1	23

Table 1: Number of MEDLINE hits when searching PubMed for individual genes using official gene symbols as defined by HUGO NC. The symbols and numbers in parenthesis refers to the symbols as given in the fibroblast article along with the corresponding number of hits. Due to recent changes in the nomenclature, some of the symbols used in the fibroblast article have been replaced by the HUGO NC committee. In such cases we have included the old symbol as well as the new symbol when performing MEDLINE search.

Category	OR	AND	Pairs
Signal Transduction	9090	-	0
Immediate-Early Transcription Factors	77012	-	11
Angiogenesis	12954	-	10
Inflammation	15907	-	9

Table 2: Number of MEDLINE hits when searching PubMed for sets of genes using official gene symbols as defined by HUGO NC and combining query terms using standard Boolean operators. The pairs column shows how many pairs from each set that gave a non-empty set of hits when combined with AND. For the genes with several symbols the query was constructed by combining the symbols with OR.

The parameters for the algorithm D-SET-COVERS were $h = 1$, $n = 1$, $\delta_{\{\beta,a,o,c\}} = 1$, and $\delta_s = 0$. The function β was defined as $\beta(A') = |A'|$. The algorithm was then applied for each of the groups with L'' as the set of all primary symbols for that group.

Table 3 shows for each category its size, the size of the returned set of references together with the degree of cover, the number of gene labels overshot, and the connectivity measure.

Category	$\|L''\|$	$\|A'\|$	$a_{R'}(A', L'')$	$No_{R'}(A', L'')$	$c_{R'}(A', L'')$
Signal Transduction	11	10	1	9	10.818
Immediate-Early Transcription Factors	11	8	1	7	10.109
Angiogenesis	8	6	1	7	1.607
Inflammation	8	5	1	10	1.429

Table 3: Set covering search results summary. The number N is the number of genes in $L' - L''$, thus $No_{R'}(A', L'')$ is the number of gene labels covered that are not in L''.

The Table 4 contains for each category the MEDLINE identification numbers returned by the algorithm together with the corresponding document titles. The computation times of each experiment programmed in Perl on a Sun Ultrasparc 4 took 70, 62, 309, and 527 seconds for each of the groups respectively.

4 Discussion

Standard evaluation measures to be maximized in information retrieval are *precision* and *recall*. The former relates to the the relevance of the single documents to the query, while the latter pertains to the totality of relevant documents returned. Precision is not directly applicable in our setting because we are interested in relevance of sets of documents, and not so much relevance of single documents to a given query. Rather, single documents are assigned relevance conditioned on a set of documents already chosen. Recall is not applicable because we try to find minimal sets of documents with maximal relevance.

The automatic nature of the approach makes it in effect suitable for a large range of document collections. MEDLINE is limiting as such, as it does not contain full text entries, thus disallowing analysis of the full text that could be used to determine better estimates for stem frequencies, etc. The automatic approach also has the benefit of flexibility in a changing field, facilitating maintenance.

The set covering approach also has a noise limiting effect in that if one assumes independence of occurrence of gene labels in documents that are noise, e.g., that contain gene labels but with other

Category	ID	Title
Signal Transduction	94156037	Chromosomal localization of the protein tyrosine phosphatase G1 gene and characterization of the aberrant transcripts in human colon cancer cells.
	96015073	Assignment of human transforming growth factor-beta type I and type III receptor genes (TGFBR1 and TGFBR3) to 9q33-q34 and 1p32-p33, respectively.
	96224012	MKP-3, a novel cytosolic protein-tyrosine phosphatase that exemplifies a new class of mitogen-activated protein kinase phosphatase.
	96430916	Cloning and characterization of a new member of the G-protein coupled receptor EDG family.
	97014933	Expression of transforming growth factor beta isoforms and their receptors during hair growth phases in mice.
	97322244	Cloning of human bone morphogenetic protein type IB receptor (BMPR-IB) and its expression in prostate cancer in comparison with other BMPRs.
	98047057	The glycine transporter GLYT2 is a reliable marker for glycine-immunoreactive neurons.
	98279134	Specific activation of the nuclear receptors PPARgamma and RORA by the antidiabetic thiazolidinedione BRL 49653 and the antiarthritic thiazolidinedione derivative CGP 52608.
	98321652	Abnormal distribution of intestinal pacemaker (C-KIT-positive) cells in an infant with chronic idiopathic intestinal pseudoobstruction.
	98390195	Genomic organization and chromosomal localization of the human SGK protein kinase gene.
Immediate-Early Transcription Factors	90021152	Interdigitating cell sarcoma (ICS). Evidence of interdigitating cell origin, immunocytochemical studies with monoclonal anti-ICS antibodies.
	90158590	Phosphorylation of the ETS-2 protein: regulation by the T-cell antigen receptor-CD3 complex.
	94227121	Fourier amplitude image circumferential profile analysis in the evaluation of the dipyridamole test.
	96152889	Fusion of the EWS gene to CHN, a member of the steroid/thyroid receptor gene superfamily, in a human myxoid chondrosarcoma.
	96243046	Altered AP-1/ATF complexes in adenovirus-E1-transformed cells due to EIA-dependent induction of ATF3.
	97242200	Inducibility of E4BP4 suggests a novel mechanism of negative gene regulation by glucocorticoids.
	98372580	Reduced transduction mechanisms in the anterior accumbal interface of an animal model of Attention-Deficit Hyperactivity Disorder.
	98447675	Cloning the cDNA for a new human zinc finger protein defines a group of closely related Kruppel-like transcription factors.
Angiogenesis	91378887	A highly sensitive E.L.I.S.A. for endopeptidase-24.11, the common acute-lymphoblastic-leukaemia antigen (CALLA, CD-10), applicable to material of porcine and human origin.
	93005105	[FGFB binding sites in cancers of the human breast].
	96039262	Structure and chromosomal localization of the human stromal cell-derived factor 1 (SDF1) gene.
	99022844	[Free radicals and antioxidants: physiology, human pathology and therapeutic aspects].
	99064567	Spatial and temporal expression of angiogenic molecules during tumor growth and progression.
	99309816	IL1B gene polymorphisms influence the course and severity of inflammatory bowel disease.
Inflammation	92265957	Genes for beta-thromboglobulin and platelet factor 4 are closely linked and form part of a cluster of related genes on chromosome 4.
	96039262	Structure and chromosomal localization of the human stromal cell-derived factor 1 (SDF1) gene.
	99141541	Cyclooxygenase metabolites mediate glomerular monocyte chemoattractant protein-1 formation and monocyte recruitment in experimental glomerulonephritis.
	99309816	IL1B gene polymorphisms influence the course and severity of inflammatory bowel disease.
	99375136	Infection of human respiratory submucosal glands with rhinovirus: effects on cytokine and ICAM-1 production.

Table 4: MEDLINE Id's and titles for the articles returned by the set-covering algorithm using the sets of primary symbols from each set as search query.

semantics. The probability of a document being noise quickly becomes small as the number of gene labels it contains increases.

As all queries resulted in sets with degree of cover 1, at least one document for each input gene label was found that contained it. This was not the case for the MEDLINE search as the gene label PTPN12 was not found in any documents. This may be explained by the "thesaurus" functionality included by the indexing via the gene label – gene label relation L_L.

For all the four groups of gene labels, the number of returned documents was less than the number of gene labels in the corresponding group. Relative to the number of gene labels, the most documents were returned for the Signal Transduction group, the fewest for the Inflammation group. The connectivity measure identifies a partition of the set of groups, Signal transduction and Immediate-Early Transcription factors on one hand and Angiogenesis and Inflammation on the other. The first grouping showing few co-citations of labels, while the second showing many. These differences seem to indicate little research into commonalities between the genes in the categories in the first grouping, while more of such in the categories of the second grouping. A possible explanation for this is that the first grouping represents properties of genes, while the second represents particular biochemical processes. Even though connectivity seems to rediscover gene groupings, it comes at a cost as up to 97% of the computation time in our implementation is spent evaluating this measure alone. This means that if connectivity is left out, significant speedups over the achieved results can be expected.

Inspection of the selected documents reveals that 1 (MEDLINE ID 94227121) can be characterized as noise, in terms of not containing a reference to one of the genes in the group. This is however, a weakness of the indexing procedure and is not discussed here. Furthermore, as many as 8 of the remaining 27 documents describes chromosomal mapping, and are as such less interesting with respect to retrieving documents pertaining to regulatory or other kinds of interactions. This can also be alleviated by refining the indexing procedure, or alternatively by imposing a penalty for articles that can be characterized as describing genomic organization.

4.1 Document relevance

The main assumption behind our approach to document relevance is that a description of a relationship between the gene labels in a given set[2] requires all genes in this set to be mentioned together in a single document. This leads to the idea that the most relevant documents must be among the minimal sets of documents that together mention all gene labels of interest. These are exactly the minimal set covers of gene labels by document references collected in the relation R'. This results in an extension of document relevance to relevance conditioned on a set of documents and a particular document-gene label relation.

g_1	g_2	g_3	g_4	g_5	ranking value
1	1	0	0	0	0.5
0	1	1	0	0	0.5
1	0	1	0	1	0.5
0	0	0	1	0	0.25
0	0	0	1	1	0.25

Table 5: Example for the failure of an approach based on document ranking. The table denotes presence or absence of gene symbols in documents, together with a fictitious value (incidentally the cover) assigned to each individual document for a match with the input symptom pattern g_1, g_2, g_3, g_4 all present.

Consider the example search given in Table 2. A set of five documents containing references to one or more of five genes is presented, and ranked according to the cover of the input set of gene

[2] In the general case, a gene label could also denote a taxonomic label.

labels $\{g_1, g_2, g_3, g_4\}$. It is impossible to say, based only on the ranks, which combination, if any, will maximize the cover of the input set. One could argue that a combination of documents that have large ranking values will improve our explanation. As we can see in this simple example, however, the two lowest valued documents are the only one that cover label g_4.

In addition to the degree of cover, three additional measures can be associated with a set cover. They are separation, connectivity and overshoot. Overshoot can be related to specificity of a given set of documents with respect to the set of genes given. If we assume that the more genes that are mentioned in a given document, the less central this document is in the literature about each gene, we would like to penalize documents writing about a lot of genes not mentioned in the query. Hence, by incorporating the overshoot measure in the measure of relevance, it is possible to favor documents that are specific with respect to the given list of genes by giving δ_o a high value. Similarly, separation is related to how many times a gene is covered. High separation might be wanted, e.g., if we assume that fewer genes mentioned in a single document is better. Connectivity is a measure related to separability in complementary fashion. It is a measure that can be used to targeting maximal cliques in the gene label graph induced by a set of documents. High connectivity might be wanted if we assume a sort of transitivity of the relation between the genes, and want to minimize the number of transitive "steps" we have to make to "connect" any connectable pair of genes. The values of δ_s and δ_c govern separation and connectivity.

If we substitute lines marked by (1), (2), (3) and (4) in the greedy algorithm given in Section 2.2.3 with

(1') $c \leftarrow 1$
(2') **while** $T \neq \emptyset$ **and** $c_R(A', G') > h$ **and** $c_R(A', G') < c$
(3') $\quad c \leftarrow c_R(A', G')$
(4') $\quad$ **if** $c_R(A', G') > h$

we get a variation of the D-Set-Covers that returns covers with maximal connectivity. The algorithm will add documents to the cover even though all genes have been covered if this decreases c_R. The substitution represents a change in stop criterion, and can be used in combination to, or as an alternative to, setting a high value for the maximum numbers of disjoint covers returned by D-Set-Covers when an increase in the number of document references returned is wanted.

If information about relevance of individual documents by themselves is available, this can be incorporated in the cost function β used in m_R. An example would be knowledge about impact factors[3]of publications in which the documents are published. If we let a document $a \in A$ inherit this impact factor from the publication in which it can be found, and denote it $i(a)$, we can construct the mapping β as:

$$\beta(a) = \frac{\sup(i(A)) - i(a)}{\sup(i(A))},$$

where $\sup(i(A))$ denotes the largest impact factor for any document in A'. For $A' \supseteq A$, $\beta(A') = \sum_{a \in A'} \beta(a)$. Larger impact factors will then be deemed more relevant than smaller. Other properties of documents that can be used in a cost measure of documents are availability, language, and whether it is a primary publication or not.

The difference of separation as a search criterion and as a property of all possible solutions of a document search should be noted. If separation is a property of all possible covers of two gene labels, this indicates a lack of reported relationships between these labels and could be used as an indicator of an area to be investigated. Similarly, connectivity in the result could be taken as an indication of extensive research into the relationship between the genes in question.

In some sense the results from the set-cover algorithm may be interpreted as a guide to formulate

[3]Such an impact factor can be obtained from the Institute for Scientific Information

Boolean criteria for combining the search terms. In fact, a minimal set-cover will be a subset of the set of documents returned from the strictest intersection operation (Boolean query) returning a non-empty set of references. This subset will in many cases be a much smaller proper subset.

The notion of applying set-covering in searching for documents in a keyword-document index is not new. Syu and Lang [9] present a covering approach where document relevance is defined as the product of scores of overlap, overshoot and parsimony. Their factorial formulation disallows weighting of these competing factors. Their measure of overshoot is also different in that it favors multiple covers of singleton elements and not cover of pairs such as connectivity does, and is thus less suited for our particular problem.

4.2 Constructing the Gene-Literature Index

As there is no complete database of human genes available, the construction of the set of gene labels L, the primary gene labels L', and the relation G_L, from genes to labels, was done using several online databases. The list of primary labels L' was selected from the database of the HUGO Nomenclature Committee, and was extended to L by data from the Genome Database (GDB), LocusLink, GENATLAS, as well as UniGene. The relation L_L was constructed from the aliasing information in these databases.

The MEDLINE literature data base as a source for document references was chosen because of availability, and because a large portion of documents published in the field of biomedicine are indexed by MEDLINE.

A major design issue was that the construction of the relation between gene labels and document references should be automated, thus avoiding problems associated with indexing done by humans.

The organization of information about genes and their labels into the relations L_L and G_L, imply that full knowledge of all alternative names for a given gene is not required when searching for relevant literature. A search on any of the labels associated with a gene is automatically mapped to that gene, and the search is conceptually done on that level. Two problems associated with this approach are:

- The existence of a set of primary gene labels such that a bijection to the set of genes can be constructed. In other words, can we uniquely label each gene?
- Whether the restriction of L_L to L' is separating. In other words, are there non-primary gene labels that are associated with different primary labels.

Both these problems are problems of standardization, and are of the domain in general. From a practical standpoint, the first problem was in our case small, as only a handful of genes shared primary symbols, and the last can be resolved by either augmenting the set of genes that need to be covered in the search by the alternative genes associated with this symbol, compute in parallel all possible ways of resolving ambiguities, or require disambiguating input. The latter conflicts with the criterion of automation, the number of parallel solutions can become overly large, so pragmatics dictated augmentation of the target gene set.

Physically creating our own gene-literature index was necessary both because of the limitations of the MeSH system and MeSH index, but also because of the need to have this data locally in order to enable effective computation.

4.3 Future work

The applicability of our approach depends on several factors such as the adequacy of the relevance definition, the quality of the gene label–document relation, the quality of the gene label–gene label

relation, and the efficiency of the implementation.

Document relevance can be extended by additional relations and by specifying additional requirements in terms of (weighted) coverage in these additional relations. One such relation that has already been implemented but not included is a gene label–Mesh term relation. Other relations could come from an automatic analysis of the contexts of the gene label occurrences in the documents. Additionally, a sensitivity analysis of weight selection for the different measures involved in the relevance measure presented is planned, as is the extension of the software to use publication impact factors in the relevance determination.

An extended occurrence context analysis can also be used to filter out noise, i.e., occurrences of gene labels in non-gene related documents, and thus improving the gene label–document relation. Incorporating coming gene label standards will also be done. The incorporation of label–document weights by, e.g., occurrence frequencies for full text data sources is envisioned.

The current implementation is a suite of Perl modules accessible through a CGI interface on the web. Identifying and rewriting the time-critical parts of the system is planned, as is the investigation of the adequacy of approximate heuristic and parallel solutions to the covering problem as we define it.

Finally, the whole system will be subjected to parts of the biomedical community for feedback.

Acknowledgements

We thank Eivind Hovig, Jan Komorowski, and Astrid Lægreid for inspiring discussions. We also thank Holger Billhardt, Stephan Dreiseitl, Aleksander Øhrn, and Bjarte Østvold for helpful comments on the manuscript, and the National Library of Medicine for giving us access to the MEDLINE data. This work was funded in part by grants R01 LM06538-01 from the US National Library of Medicine.

References

[1] M. Schena, D. Schalon, R. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a cdna microarray. *Science*, 270:467–470, 1995.

[2] P. O. Brown and D. Botstein. Exploring the new world of the genome with dna microarrays. *Nature Genetics*, 21(1):33–37, 1999.

[3] R. J. Lipshutz, S. P. A. Fodor, T. R. Gingeras, and D. J. Lockhart. High density synthetic oligonucleotide arrays. *Nature Genetics*, 21(1):20–24, 1999.

[4] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[5] M. Eisenberg and C. Barry. Order effects: a study of the possible influence of presentation order on user judgements of document relevance. *Journal of the American Society for Information Science*, 39(5):293–300, Sept 1988.

[6] R. Marega and M. T. Pazienza. Codhir: an information retrieval system based on semantic document representation. *Journal of Information Science*, 20(6):399–412, 1994.

[7] Y. Y. Yao. Measuring retrieval effectiveness based on user preference of documents. *Journal of the American Society for Information Science*, 46(2):133–45, 1995.

[8] Ching Chi Hsu and Chia Hui Chang. Webyacht: a concept-based search tool for www. *International Journal on Artificial Intelligence Tools*, 8(2):137–56, 1999.

[9] I. Syu and S. D. Lang. A competition-based connectionist model for information retrieval using a merged thesaurus. In *Proceedings of CIKM 1994*, pages 164–170, 1994.

[10] http://www.idi.ntnu.no/~tkj/scare/.

[11] D. S. Johnson. Approximation algorithms for combinatorial problems. *Journal of Computer and System Sciences*, 9:256–278, 1974.

[12] G. Ausiello, P. Crescenzi, G. Gambosi, V. Kann, A. Marchetti-Spaccamela, and M. Protasi. *Complexity and Approximation*. Springer-Verlag, 1999.

[13] V. R. Iyer, M. B. Eisen, D. R. Ross, D. Lashkari, T. Moore, G. Schuler, J. Lee, M. Boguski, D. Shalon, D. Botstein, and P. O. Brown. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283:83–87, 1999.