BMC Bioinformatics

Research article

Identifying functional relationships among human genes by systematic analysis of biological literature

Yong-Chuan Tao*[1] and Rudolph L Leibel[2]

Address: [1]Life Science Informatics Unit, Functional Genomics Area, Novartis Pharmaceuticals, Summit, NJ 07901, USA and [2]Division of Molecular Genetics, Department of Pediatrics, Columbia University, New York, NY 10032, USA

E-mail: Yong-Chuan Tao* - yong-chuan.tao@pharma.novartis.com; Rudolph L Leibel - rl232@columbia.edu

*Corresponding author

Published: 7 June 2002

BMC Bioinformatics 2002, **3**:16

Received: 5 April 2002
Accepted: 7 June 2002

This article is available from: http://www.biomedcentral.com/1471-2105/3/16

Abstract

Background: The availability of biomedical literature in electronic format has made it possible to implement automatic text processing methods to expose implicit relationships among different documents, and more importantly, the functional relationships among the molecules and processes that these documents describe.

Results: A computational strategy that identifies functionally related human genes by detecting the implicit relationships among the publications cited under each gene in the Online Mendelian Inheritance in Man (OMIM) was implemented. The implementation was based on a substantially modified version of the kernel document method. The improvements include assigning a calculated weight for a document to indicate its importance in establishing the relationship between two documents, and using multiple kernel documents to reflect the multiple functions of the same gene. An example of using this strategy to identify genes related to the apoptosis pathway in human was given.

Conclusions: The results showed that this method can indeed produce meaningful results when applied to human genes.

Background

Among the many approaches to identifying functional relationships among genes, the use of bibliographic data to group genes that are functionally related has recently attracted great attention. The huge repository of biological literature, which is still growing at a rapid pace, makes it increasingly difficult for any individual to monitor exhaustively the constituent items related to a specific biological process. Therefore, automated data mining systems for biological literature are becoming a necessity.

The availability of biomedical literature in electronic format has made it possible to implement automatic text processing methods to expose implicit relationships among different documents, and more importantly, the functional relationships among the molecules and processes that these documents describe.

Shatkay *et al*[1] proposed a method, which we denote as the "kernel document method", and applied it to the identification of functional relationships among yeast genes. Briefly, for each gene, a kernel document is carefully selected to establish a one-to-one correspondence be-

tween a gene and a kernel document. A set of "related documents" associated with each kernel document is identified using statistical information retrieval methods. The extent to which the two sets of related documents corresponding to each of a pair of kernel documents overlap reflects the relevance of these two kernel documents, and hence the possible functional relatedness of the corresponding genes.

The utility of this method relies heavily on the quality of the kernel documents. In this context, a good kernel document should focus on the functions of a gene, instead of on other topics such as the methods or techniques used to identify or study the gene. With carefully selected kernel documents, the relatedness of this gene to others can be made reliant on functional rather than, e.g., structural characteristics. For example, if the topic of one kernel document is "studying gene A by method X", and the topic of the other kernel document is "studying gene B by method X", two functionally unrelated genes A and B could be related to one another simply because they have both been studied by method X. Avoiding such "false positives" is a challenge in applying this method. The selection of functionally-descriptive kernel documents is, therefore, a key to the success of this algorithm.

In the original kernel document method, all documents that are related to two kernel documents are weighted equally in establishing the qualitative and quantative aspects of relationship between these two kernel documents. A better practice is to give each document a weight reflecting the relative uniqueness of this document's relationship to the kernel documents. A document that is related to only a few kernel documents is given a greater weight than one that is related to many kernel documents. This argument can be illustrated with an intuitive example: if you are asked to identify two people from a crowd, it is not very helpful if the only information you are given is that each of the two has a nose. However, if you are told that each of the two has a mole on the forehead, it will not be too difficult to single them out. This is because "having a nose" is a feature common to almost everybody. But the description that each of two people has a mole on the forehead, an uncommon feature, is an important piece of information that can be used to establish a link between the two people.

The kernel document method was initially applied to yeast genes. Intense, relatively long-standing analysis of yeast genetics has resulted in a large number of PubMed entries on these genes. Whether the kernel document method could be applied to other less abundantly represented genes, such as human genes, was not known. Here we will apply this method to human genes, and show that this method can indeed produce meaningful results when applied to human genes.

A potential limitation of the original kernel document method is that only one kernel document is chosen for each gene. Many genes encode multi-functional proteins, and one kernel document might relate only to a certain aspect of the gene's many functions. We addressed this problem by selecting multiple kernel documents for a gene, so that any known function of the gene would be discussed in at least one of these kernel documents.

Jenssen *et al*[2] took a different approach. They analyzed the titles and abstracts of MEDLINE records to look for co-occurrence of gene symbols. The results are available at PubGene [http://www.pubgene.org] . This approach is based on the assumption that if two gene symbols appear in the same MEDLINE record, the genes are likely to be related. Furthermore, the number of papers in which the pair of genes both appear is used to assess the strength of relationships between the two genes. Jenssen *et al* manually examined 1,000 randomly selected pairs from the network of genes that had been created using this method: the proportion of incorrect (biologically meaningless) pairs were 40% for the low-weight category and 29% for the high-weight category. The main advantage of this method in comparison with the kernel document method is that it avoids the difficulty of selecting an appropriate kernel document. However, this method cannot identify genes that are functionally related, but are not mentioned together in any MEDLINE abstract. Such implicit relationships between genes are inherently more interesting in the context of mechanism/pathway discovery by computation.

In this paper, we employ a method that is based upon the kernel document concept, with several enhancements. First, instead of choosing one kernel document for each gene, we employ all of the reference articles cited for each gene symbol in OMIM. Admittedly, not all of these articles are good candidates for kernel documents. However, the reference articles cited under each OMIM entry are a set of documents selected by investigators familiar with the gene and are, therefore, related to the gene in some way. Furthermore, by a simple examination of the titles of the articles for keywords alluding to methods or techniques, many articles that would be likely to constitute false positives in this context are excluded. Second, instead of weighing each related article equally, a weight is calculated for each article that is related to two or more kernel documents. We call these articles "base vector documents", because eventually a kernel document will be represented by a vector whose elements are determined by whether it is related to a base vector document. The more

kernel documents a base vector document is related to, the less its weight.

Methods

The calculations described were performed on a Dell Precision 620 running the Linux operating system. Data were stored in a MySQL relational database. Data storage and retrieval were automated with the aid of scripts written in PERL. The most computationally intensive part of the code, which is responsible for the calculation of similarity scores between documents, was written in C. This part of the calculation took about 12 hours.

Data Preparation

1. Download the list of OMIM genes

The OMIM gene list can be downloaded from NCBI [http://www.ncbi.nlm.nih.gov/Omim/Index/genetable.html] . This list is inserted into a relational database table, which consists of only two fields: the symbol of a gene, and the corresponding OMIM identification number (OMIMID). However, due to inconsistencies in gene naming and use conventions, several gene symbols may correspond to the same OMIMID.

2. Download the references cited under each OMIMID

The reference papers listed under each OMIMID are then downloaded. Each distinct reference paper has a unique PubMed identification number (PMID). The titles of all such PubMed papers are also obtained. The data are stored in another table consisting of four fields, OMIMID, PMID, TITLE and KEEP. The first three fields are self-explanatory. KEEP is a flag indicating whether a particular PubMed paper should be treated as a kernel document. As indicated earlier, methodology papers are generally not good candidates for kernel documents. To reduce the number of such false positives, a list of keywords/phrases that include the commonly used methods and techniques is compiled. If the title of a paper includes any of the phrases in the list, the KEEP flag of the paper is turned off (set to zero).

3. Download the related documents

We treat each reference paper whose KEEP flag is on as if it were a kernel document. The documents related to each of these reference papers can be obtained from NCBI [http://www.ncbi.nlm.nih.gov/entrez/utils/pmneighbor.fcgi?pmidfepmid=PMID] . A detailed description of the computational methods used by NCBI to identify related documents is available at [http://www.ncbi.nim.nih.gov/PubMed/computation.html] .

The related documents (or neighbors) of a particular paper are listed according to their relevance to the paper. Documents that appear on the top of the list are more similar to the query than those appear near the bottom of the list. We keep only the PMIDs of the first 100 related documents in the list and the data are stored in another table, consisting of three fields, PMIDK (PMID of the kernel document), PMIDN (PMID of the related document or the neighbor), and RANK, a number from 1 to 100, indicating the place a document appear in the list of related documents. Obviously, for any PMIDK, RANK = 1 if PMIDN = PMIDK, this is because a document is always most similar to itself.

Construction of Base Vectors Documents

Using the data obtained in the previous section, the base vector documents are defined. These are the documents that are related to at least two other documents and are among the 50 top-ranking related documents of any document. The result is inserted into another database table that consists of three fields: 1. PMID, the PubMed identifier of the base vector document; 2. LINKED2, the number of kernel documents of which the specified document is a neighbor; and 3. WEIGHT, which is an indication of the importance of a base vector document in revealing the relevance between two kernel documents. The weight w_i for a base vector document b_i is calculated using the following equation:

$$w_i = \log_2 \frac{N}{n_i}, \qquad (1)$$

where n_i is the number of related documents for b_i and N is the total number of kernel documents. This weight measurement method is based upon information theory [3], and is similar to the weight measure employed by Wilbur *et al*[4] to evaluate the significance of a specific keyword in determining the relatedness of two papers.

Vector Representation of a Kernel Document

Assuming that there are M base vectors documents, b_1, b_2, . . ., b_M, and the weight of b_i is w_i, then any kernel document **K** can now be represented by a vector $(k_1, k_2, ..., k_M)$, with

$$k_i = \begin{cases} w_i & \text{if } b_i \text{ is a neighbor of } \mathbf{K}; \\ 0 & \text{otherwise.} \end{cases}$$

The norm $||\mathbf{K}||$ of a kernel document **K**, i.e., the length of the corresponding vector, can be calculated as follows:

$$\|\mathbf{K}\| = \left(\sum_{i=1}^{M} k_i^2 \right)^{1/2}. \qquad (2)$$

Calculation of Similarity Scores

The cosine similarity score S_{ij} of any two kernel documents $\mathbf{K}_i$, and $\mathbf{K}_j$ can now be calculated:

$$S_{ij} = \frac{\mathbf{K_i} \cdot \mathbf{K_j}}{\|\mathbf{K_i}\| \cdot \|\mathbf{K_j}\|}, \qquad (3)$$

where

$$\mathbf{K_i} = \left(k_1^i, k_2^i, ..., k_M^i\right), \qquad (4)$$

$$\mathbf{K_j} = \left(k_1^j, k_2^j, ..., k_M^j\right), \qquad (5)$$

and

$$\mathbf{K_i} \cdot \mathbf{K_j} = \sum_{l=1}^{M} k_l^i \cdot k_l^j \qquad (6)$$

is the dot product of the two vectors $\mathbf{K}_i$ and $\mathbf{K}_j$.

S_{ij} is between 0 and 1, i.e., $0 \le S_{ij} \le 1$. The closer S_{ij} is to 1, the more similar two kernel documents $\mathbf{K}_i$ and $\mathbf{K}_j$ are.

This is the most computationally intensive part of the calculation and the code is implemented in C. Once the similarity scores for all possible pairs of PMIDs are calculated, the scores are stored in a relational database table, and it is not necessary to recalculate the scores for subsequent queries.

Gene Relationship

The score S_{ij} calculated for two kernel documents $\mathbf{K}_i$ and $\mathbf{K}_j$ does not directly reflect the relevance of two genes. To assess the functional relationship between two genes, gene symbols must be related to PMIDs.

In order to identify the set of genes that are relevant to a query gene *G*, the PMIDs of all reference papers listed under the OMIMID for the query gene are obtained. Each of these reference papers, except any paper whose KEEP flag is turned off, is treated as a kernel document.

There are several considerations that support this approach to selection of kernel documents:

• The reference papers listed under each OMIMID were selected specifically because of their relevance to the corresponding gene;

• The titles of these papers were screened to exclude those that describe commonly used methods or techniques in order to reduce the number of "false positives";

• The process can be fully automated to avoid manually selecting kernel documents.

An interface is provided to allow the user to "fine-tune" the query by manually selecting only some of the reference papers as kernel documents.

Next, all documents (represented by their PMIDs) that are related to each kernel document with a score higher than a specified threshold are identified. The OMIMIDs that have cited papers with any of these PMIDs are collected.

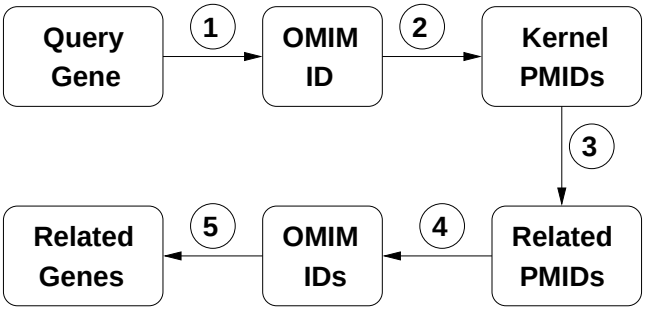

Figure 1
The process of finding the genes related to a query gene. 1) Obtain OMIMID corresponding to the query gene; 2) Obtain the PMIDs of the papers cited under the OMIMID; 3) For each of these PMIDs, find the related PMIDs with a similarity score higher than a given threshold; 4) Find the OMIMIDs under which papers with these PMIDs are cited; 5) Find the genes corresponding to these OMIMIDs.

Finally, these OMIMIDs are connected to their respective gene symbols. The entire process is shown in Figure 1.

User Interface

A user interface is available at [http://gene.cpmc.columbia.edu/cgi-bin/gene.cgi] . Once the gene symbol and a cutoff score (i.e., the cosine similarity score between two kernel documents that correspond to respective genes) are entered, a list of reference papers cited in OMIM for the gene is displayed. Only those papers whose KEEP flag is turned on are shown. The user may select specific paper(s) from the list as kernel documents, or simply check the "Check All" box to use all these papers as kernel documents.

Once the submit button is clicked, the genes with scores higher than the cutoff score are displayed.

Results

Summary of Raw Data

At the time when the raw data were downloaded in July 2001, there were 11251 gene symbols in the OMIM gene list, corresponding to 7192 distinct OMIMIDs. Multiple gene symbols may have the same OMIMID because many genes have aliases, resulting in several symbols referring to the same gene.

Among the 7192 distinct OMIMIDs, 7085 cite reference paper with PMIDs, and 107 (about 1.5%) OMIMIDs do not cite any reference paper, or only cite reference papers whose PMIDs are not specified in OMIM. 54024 reference papers are listed under the 7085 distinct OMIMIDs. Some papers are referenced under several OMIMIDs, therefore, the actual number of distinct PMIDs is 47428.

The title of the corresponding document for each of these 47428 PMIDs is also obtained. After screening the titles using the method described earlier, the KEEP flags of 3680 documents (about 7.8%) were turned off. Ultimately, only those 43748 documents whose KEEP flags are turned on will be used as kernel documents. However, we initially treat all 47428 documents as kernel documents, allowing us to estimate the extent to which these documents whose KEEP flags are turned off contribute to false positives.

For each of the 47428 distinct PMIDs, the related documents ("neighbors") are obtained from NCBI. As indicated earlier, only the first 100 PMIDs of the list of related documents are stored, because they are the ones most related to the kernel document. The highest ranking neighbor of any document is, of course, itself. This search resulted in 4629037 pairs of neighbors, a number that would be much larger if all, instead of only the top 100, neighbors of a document are kept.

Summary of Results of Calculation

The preliminary search identified 437382 base vector documents. Any of these documents is a neighbor of at least two kernel documents. On average, a base vector document is related to 9.1 kernel documents. The average weight of the base vector documents is of 13.13, the maximum weight is 14.53, which corresponds to those base vector documents that are only related to two kernel documents; the minimum weight is 4.66, which corresponds to a base vector document with 1873 neighbors. As described in the Methods section, the weight of a base vector document indicates how much information is conveyed about the relevance of two kernel documents by knowing that both of them are neighbors of this particular base vector document. The more kernel documents a base vector document is related to, the less its weight. Figure 2 shows this relationship. For example, a base vector document that is related to 740 kernel documents has a weight of 6, only half of the weight of a document that is related to 12 kernel documents.

Next, the norm of each kernel document is calculated. There are 95 kernel documents with a norm of zero. These documents do not have any neighbor that is one of the base vector documents. As a result, only 47333 kernel documents are left.

Finally, the cosine similarity score of each pair of kernel documents is calculated. A document is treated as a kernel document if its KEEP flag is on and its norm is greater than zero. There are 43658 such documents. Out of the 43658(43658-1)/2 = 952988653 possible pairs, only 6596918 (about 0.7%) have a similarity score that is greater than zero, indicating some relationship between

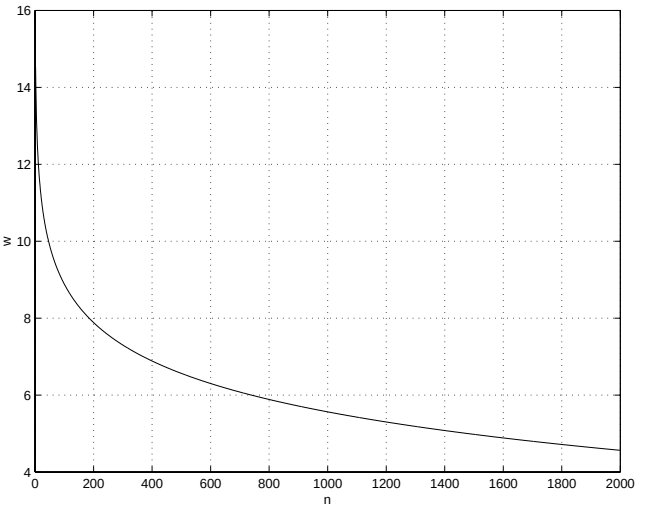

Figure 2
The weight of a base vector document. The weight of a base vector document *w* decreases as the number of kernel documents that it is related to *n* increases.

the two kernel documents of the pair. The average score is 0.04. However, if both documents of a pair are listed as references under the same OMIMID, the average score is 0.14, which is much higher than the overall average score. This difference is expected because the documents listed under the same OMIMID have been selected because they all have some relationship to the gene that corresponds to the OMIMID. Furthermore, this average score also provides an indication of the approximate value of the threshold score that should be used to decide whether two kernel documents are closely related.

Documents that discuss methods or techniques are not included when the similarity scores are calculated, because these documents can lead to false positives – a pair of genes with a high score that are functionally unrelated. To investigate the impact of such documents, we intentionally included them in the calculation of the scores. Excluding these documents when responding to a query is straightfoward, one needs only to check the KEEP flag of a document. The average similarity score of any pair in which both documents have a turned-off KEEP flag is 0.11, much higher than the overall average score 0.04 and close to the average score among a pair of documents referenced by the same OMIMID, i.e., 0.14. This result indicates that these documents should be excluded from calculations designed to find functional relationships.

Although documents that are likely to cause false positive have been excluded by the automated screening process described above, the screened set of documents may still include many that are not optimal kernel document can-

didates. A solution to this is to actually let the users select specific kernel documents from a list of documents.

An Example

As an illustration, we use this computational strategy to identify genes related to the apoptosis (programmed cell death) pathway in human. A brief recent review of this pathway has been given by DeFrancesco [9].

To use this strategy, it is necessary to have a gene to start with. This is usually a gene that is known to be associated with the pathway or function of interest. Usually, such a gene is known to the user who submits the query. If necessary, one can also perform a preliminary search of PubMed for the functions or processes of interest in order to obtain the name of a gene to start with.

We start with APAF1, a gene known to be involved in the apoptosis pathway [8]. A cutoff score of 0.2 is employed, and all reference papers cited in OMIM for this gene are used as kernel documents. The analysis identified the list of related genes displayed in Table 1.

Table 1: Search results for APAF1, using a cutoff score of 0.2

Gene Symbol	Score	Description
DIABLO (SMAC)	0.3167	Mitochondria-derived activator of caspase
XK	0.2574	Amino acid weakly similar to CED8
ABC3	0.2551	Protein sequence similar to CED7
CASP1	0.2455	Apotosis-related cysteine protease
CASP3	0.2094	Apotosis-related cysteine protease
BCL2	0.2091	Protein sequence and structure similar to CED9
CASP2	0.2074	Apotosis-related cysteine protease

CASP1, CAPS2 and CASP3 all belong to the family of apoptosis-related cysteine proteases. Caspase activation is a key regulatory step for apoptosis [10,11].

DIABLO, also known as SMAC (second mitochondria-derived activator of caspase), promotes caspase activation in a cytochrome c-APAF1-CASP9 pathway [5].

The identification of XK and ABC3 is more interesting, because they are not well recognized as components of the apoptosis pathway. In order to identify the process by which XK was included, we retrace the search path to find the two original kernel documents that related APAF1 to XK. They are: "Apaf-1, a human protein homologous to C. elegans CED-4, participates in cytochrome c-dependent activation of caspase-3" (PMID: 9267021), a paper linked to APAF1; and "The ced-8 gene controls the timing of programmed cell death in C. elegans" (PMID: 10882128), a paper linked to XK. XK is a Kell blood group precursor. Stanfield *et al*[6] noted that 458-amino acid CED8 transmembrane protein of C. elegans is weakly similar to the human XK protein. The CED8 and XK proteins share 19% amino acid identity, have similar hydropathy plots, and both contain 10 hydrophobic predicted membrane-spanning segments. CED8 functions downstream of, or in parallel to, the regulatory cell death gene CED9, and may function as a cell death effector downstream of the caspase encoded by programmed cell death gene, CED3. APAF1 is known to share amino acid similarity with CED3 and CED4, a protein that is believed to initiate apoptosis in C. elegans.

The gene ABC3 (ABC Transporter 3) is linked to APAF1 in a manner similar to that which connects XK to APAF1. It is reported that CED7 protein has sequence similarity to ABC transporters. CED7 functions in the engulfment of cell remnants during programmed cell death [7].

There was evidence that BCL2 is a homolog of CED9: CED9 encodes a 280 amino acid protein showing sequence and structural similarity to BCL2 [12]. BCL2 is involved in programmed cell death [9].

A secondary search can be performed with each of the genes in Table 1. Usually, more stringent criteria is required for secondary searches because the genes used for secondary queries often have other functions not related to the one of interest. Kernel documents need to be selected more carefully, and a higher cut-off score might be used.

For example, for XK, if all papers cited in OMIM for the particular gene are used as kernel documents, there are many high-score hits that do not seem to be directly linked to apoptosis. Among the kernel document candidates for XK, the title of only one of the papers mentions programmed cell death. The majority of papers discusses McLeod syndrome, which is associated with XK, but has no recognized relationship with apoptosis.

Therefore, further inspection is necessary to determine whether these hits are really linked to the apoptosis pathway. To simplify the process and obtain better results, instead using all reference papers cited in OMIM for each of these genes, we manually select kernel documents from the list of OMIM reference papers for these secondary searches, using the interface described before. For example, in a list of more than 20 papers cited for XK, we choose only one paper, titled "The ced-8 gene controls the timing of programmed cell death in C. elegans".

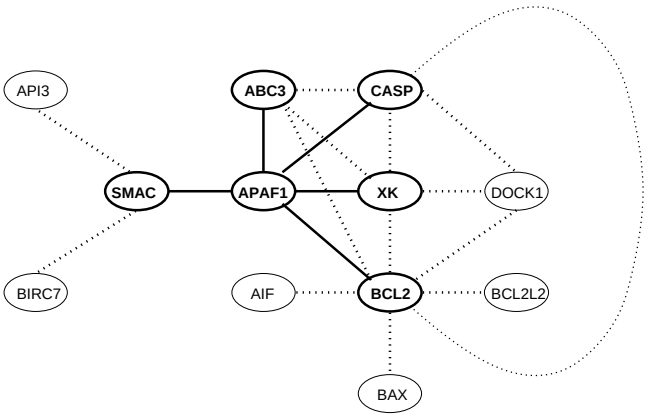

Figure 3
A network of apoptosis related genes. The network was built starting with the APAF1 gene. The node CASP include the Caspase family of genes, including CASP2, CASP3, CASP6, CASP7, CASP8, CASP9, CASP10. Genes found in the first round of searches are shown inside bold ovals, dotted lines indicate links identified by a secondary search.

With the results of the initial and secondary searches, a network of genes nominally associated with apoptosis can be built. The network is shown in Figure 3.

If necessary, further searches can be performed with the hits from a previous search, so that the network can be expanded to include more genes.

Discussion

The similarity score is the only criterion used to determine whether two documents are related. Any two documents with a similarity score above the cutoff score are considered to be related.

Here we discuss how the cutoff score should be determined. To this end, it is necessary to investigate how the distribution of similarity scores differs between related and unrelated document pairs.

To simplify the problem, we assume that any two documents that are listed as references under the same OMIMID are related, and that the distribution between such documents approximates the distribution between two related documents.

For any two documents that are not listed under the same OMIMID, it is reasonable to assume that they are unrelated, because the vast majority of such documents are, in fact, unrelated. Therefore, we assign the score distribution for unrelated documents to such pairs. It should be emphasized that this assumption is an approximation. Indeed, the most interesting documents are those documents that are not listed under the same OMIMID, but are found through analysis to be related. However, this assumption makes finding the distribution of similarity scores among unrelated documents much easier.

Table 2 is a summary of the score distributions of related and unrelated document pairs. Note that for unrelated documents, 75% of the scores are less than 0.03087, while for related documents, only 25% of the scores are less than 0.03027.

Table 2: Summary of similarity scores for document pairs

	Related	Unrelated
Min.	0.00341	0.002794
1st Quartile	0.03027	0.010710
Median	0.07647	0.016630
Mean	0.14150	0.030780
3rd Quartile	0.19860	0.030870
Max.	0.98370	0.943100

The probability $P(S > S_{cutoff})$ of the score S being greater than a cutoff score, S_{cutoff}, can be easily found:

$$P\left(S > S_{cutoff}\right) = 1 - \frac{N\left(S \le S_{cutoff}\right)}{N}, \qquad (7)$$

where $N(S \le S_{cutoff})$ is the number of document pairs whose similarity score is not greater than the cutoff score, and N is the total number of such pairs.

$P(S > S_{cutoff})$ was calculated separately for those pairs in which both documents were listed under the same OMIMID, i.e., the "related documents" according to the assumption above, and for those pairs in which the two documents were not listed under the same OMIMID, i.e., the "unrelated documents" corresponding to our definitions. The results are shown in Figure 4. The solid curve is the probability $P(S > S_{cutoff})$ for related document pairs (true positives), the dotted curve is the probability $P(S > S_{cutoff})$ for unrelated document pairs (false positives). Using a cutoff score of 0.05, about 61% of the related documents will be accepted; these documents are true positives. About 39% of the related documents will be rejected; these are the false negatives. Only 14% of the unrelated documents will be accepted; these are the false positives. And, 86% of the unrelated documents will be rejected, these are the true negatives.

Based on these results, the sensitivity and specificity of the search can be calculated. The sensitivity is the proportion of related document pairs that are about the cutoff score,

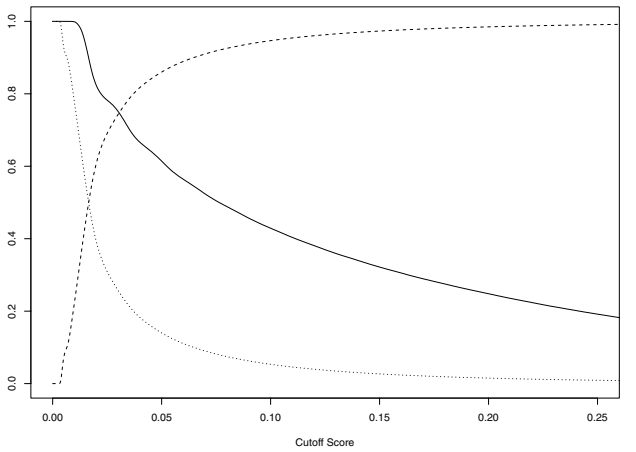

Figure 4
Sensitivity and specificity. Solid curve: the proportion of related document pairs with a similarity score above the cutoff score, this is also the sensitivity curve. Dotted curve: the proportion of unrelated document pairs with a similarity score above the cutoff score. Dashed curve: specificity curve.

and therefore are accepted. Therefore, the solid curve in Figure 4 is also the sensitivity curve. The specificity is the proportion of unrelated documents that are below the cutoff score, and therefore are rejected. Specificity is equal to 1 - $P(S > S_{cutoff})$, where $P(S > S_{cutoff})$ is the proportion of unrelated document pairs that are above the cutoff score S_{cutoff}. In Figure 4, the dashed curve is the specificity curve.

Figure 4 can be used to determine what cutoff score to use for any specific purpose. For example, using a high cutoff score such as 0.2, the specificity will be 0.985, corresponding to a false positive rate of only 1.5%. However, the corresponding sensitivity is 0.248, so that above three quarters of the related documents will also be rejected. On the other hand, choosing a low cutoff score will result in many false positives, while ensuring that most related documents are accepted. Using a cutoff score of 0.03, both the sensitivity and specificity will be around 0.75. However, because there are often many more unrelated documents than related documents, the search result will still contain many false positives. By referring to Figure 4, users can select a cutoff score that is best suited to their needs.

Conclusions

The key to the success of the kernel document method is the selection of the kernel documents. However, this is also the most difficult and tedious part of the implementation. An efficient way to select the kernel documents related to gene function is necessary for a large-scale literature mining effort using this method. We started with all of the reference papers listed in OMIM, and applied a filter to exclude those papers that are likely to focus primarily on methods and techniques. We can either treat the rest of papers as kernel documents, or allow the user to select kernel documents from this small pool of papers (usually contain around a dozen papers).

This process can be fully automated. Furthermore, since we are not limited to one kernel document per gene, a gene can correspond to multiple kernel documents that capture different aspects of its functions. This characteristic of the strategy makes it possible to identify genes that are related to the query gene through a variety of functional mechanisms.

In distinction to the gene co-occurrence method used by Jenssen *et al*, this approach does not require the symbols of two gene to appear in the title or abstract of the same paper in order to establish a relationship between them. As long as similar or related functions of the two genes are described in the literature, the relationship between the two genes is likely to be revealed. Furthermore, it is easier to identify the related functions of these genes because they are precisely those functions that related one gene to another by computation. While the co-occurrence method is biased towards revealing gene relationships that have been explicitly described in the literature, the method we propose is more sensitive to implicit relationships between two genes that have not necessarily been explicitly identified.

The process of selecting kernel documents can also be improved by taking advantage of user feedback in a networked environment. For example, the user can be allowed to select kernel documents from a list of candidate papers. The papers selected most frequently by users can then be used as the bases for subsequent automatic kernel document selection in searches related to a specific gene or pathway.

Finally, it is important to take note of the limitation of literature mining tools: two genes could be found to be related for many reasons, some of which might not be biologically meaningful. The identified relationships could therefore have different biological meanings, if any. Further investigation is always necessary to determine the origin of such relatedness. However, bibiliographic data mining efforts such as ours could shed light on the less obvious relationships between two genes. When considered in conjunction with other data, such as gene expression profiles, the results could lead to biologically meaningful conclusions.

Acknowledgement

We would like to thank Yingyao Zhou for many helpful discussions.

References

1. Shatkay H, Edwards S, Wilbur WJ, Boguski M: **Genes, Themes and Microarrays: using information retrieval for large-scale gene analysis.** *Proceedings of the International Conference on Intelligent Systems for Molecular Biology AAAI Press, San Diego,* 2000, **8:**317-328
2. Jenssen T-K, Lægreid A, Komorowski J, Hovig E: **A literature network of human genes for high-throughput analysis of gene expression.** *Nature Genetics* 2001, **28:**21-28
3. Ash R: **Information Theory.** *Chapter 1. Dover Publishers* 1990
4. Wilbur WJ, Yang Y: **An analysis of statistical term strength and its use in indexing and retrieval of molecular biology texts.** *Comput. Biol. Med.* 1996, **26:**209-222
5. Du C, Fang M, Li Y, Li L, Wang X: **Smac, a mitochondrial protein that promotes cytochrome c-dependent caspase activation by eliminating IAP inhibition.** *Cell* 2000, **102:**33-42
6. Stanfield CM CM, Horvitz HR: **The ced-8 gene controls the timing of programmed cell deaths in C. elegans.** *Mol. Cell* 2000, **5:**423-433
7. Wu Y-C, Horvitz HR: **The C. elegans cell corpse engulfment gene ced-7 encodes a protein similar to ABC transporters.** *Cell* 1998, **93:**951-960
8. Li P, Nijhawan D, Budihardjo I, Srinivasula SM, Ahmad M, Alnemri ES, Wang X: **Cytochrom c and dATP-dependent formation of Apaf-1/caspase-9 complex initiates an apoptotic protease cascade.** *Cell* 1997, **91(4):**479-89
9. DeFrancesco L: **Death in the balance.** *The Scientist* 2001, **15(13):**17
10. Budihardjo I, Oliver H, Lutter M, Luo X, Wang X: **Biochemical pathways of caspase activation during apoptosis.** *Annu. Rev. Cell Dev. Biol.* 1999, **15:**269-90
11. Cohen CM: **Caspases: the executioners of apoptosis.** *Biochem. J* 1997, **326(Pt 1):**1-16
12. Hengartner MO, Horvitz HR: **C. elegans cell survival gene ced-9 encodes a functional homolog of the mammalian proto-oncogene bcl-2,** *Cell* 1994, **76(4):**665-76