Original Paper

Journal of Biomedical Science

J Biomed Sci 2004;11:864–873
DOI: 10.1159/000081834

Received: March 31, 2004
Accepted: June 16, 2004

TransMiner: Mining Transitive Associations among Biological Objects from Text

Vijay Narayanasamy[a] Snehasis Mukhopadhyay[b] Mathew Palakal[b]
David A. Potter[c]

[a]School of Informatics, [b]Department of Computer and Information Science, and [c]Departments of Medicine and Biochemistry and Walther Oncology Center, Indiana University School of Medicine, Indiana University Purdue University Indianapolis, Ind., USA

Key Words
Association · Discovery · Graph · Hypotheses generation · Text mining · Transitive closure

Abstract
Associations among biological objects such as genes, proteins, and drugs can be discovered automatically from the scientific literature. TransMiner is a system for finding associations among objects by mining the Medline database of the scientific literature. The direct associations among the objects are discovered based on the principle of co-occurrence in the form of an association graph. The principle of transitive closure is applied to the association graph to find potential transitive associations. The potential transitive associations that are indeed direct are discovered by iterative retrieval and mining of the Medline documents. Those associations that are not found explicitly in the entire Medline database are transitive associations and are the candidates for hypothesis generation. The transitive associations were ranked based on the sum of weight of terms that co-occur with both the objects. The direct and transitive associations are visualized using a graph visualization applet. TransMiner was tested by finding associations among 56 breast cancer genes and among 24 objects in the calpain signal transduction pathway. TransMiner was also used to rediscover associations between magnesium and migraine.

Introduction

Many hypotheses are formed by extrapolating the current knowledge: for example, if we know that apoptosis in breast cancer is mediated by calpain [10, 14, 24, 25], we can ask if apoptosis in other related cancer (e.g. ovarian cancer) is also mediated by calpain [2].

The query 'breast cancer' on Medline returned 126,543 documents on March 31, 2003. This shows the large volume of scientific information available in the form of text. It is impossible or impractical for anyone to read through all of these documents to find the relevant information. Also, it is even more difficult to capture the knowledge in those documents. Researchers and scientists are challenged by this increasing knowledge gap. Associations among biological objects such as genes, proteins,

KARGER
Fax + 41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

1021–7770/04/0116–0864$21.00/0
Accessible online at:
www.karger.com/jbs

Snehasis Mukhopadhyay
Department of Computer and Information Science, Indiana University
Purdue University Indianapolis, 723 W. Michigan St. SL 280
Indianapolis, IN 46202 (USA)
Tel. +1 317 274 9732, Fax +1 317 274 9742, E-Mail smukhopa@cs.iupui.edu

molecules, processes, diseases, drugs and chemicals are one such form of underlying knowledge.

Swanson [22] found an association between magnesium and migraine headache that was not explicitly reported in any article, based on associations extracted from medical journal titles before it was discovered experimentally. The goals of this paper are to discover both existing and new, potentially meaningful associations from bibliographic databases such as Medline, and (2) to visualize the associations.

Association rules were originally developed for the purpose of market-basket analysis. An association is an implication of form A→B where 'A' is a set of antecedent items and 'B' is the consequent item [27]. The intuitive meaning is that the baskets that contain 'A' tend to contain 'B'. In text mining, the baskets are individual sentences or the entire documents and the items are the words. In text mining the associations are of the form WordA→WordB and/or vice versa (e.g. Smoking→Lung Cancer, Loss of apoptosis→Cancer).

The 'PubGene' project of Jenssen et al. [9] aims at information extraction about human genes from the published literature and making the information available to computerized analyses of gene expression data, for example. The 'PubGene' gene-to-gene network is constructed from the Medline citation database by linking two genes that have been mentioned in the same article. This is based on the assumption that if two gene symbols appear in the same Medline record, the genes are likely to be related (principle of co-occurrence). The number of documents in which the gene pair appears is used to assess the strength of the relationship between those genes. The bottleneck of this approach is that it cannot identify genes that are functionally related, when they are not mentioned together in any Medline abstract.

Stephens et al. [19] also used the co-occurrence principle for association discovery. They used a thesaurus of genes and tf·idf algorithm for document representation. The relative importance of each gene as well as the strength of their joint occurrences plays an important role in discovering associations in their method. They calculated the association between two gene terms as the sum of the product of weights for those two genes over the entire document collection. An additional thesaurus with possible relationship terms was used to find the nature of relationships between the gene pairs based on co-occurrence of the genes and the relationship term.

Stapley and Benoit [18] described the possible significance of the co-occurrence of two gene terms in a Medline document and the information value of such occurrences. Two gene names can occur in the same document if there is a physiological relationship between the two genes (direct physical interaction between the genes or abstract functional link) or if there is an evolutionary relationship between the two genes or due to genomic proximity.

Swanson [21] proposed the idea of using the scientific literature for generating new hypothesis that should be later verified by traditional scientific experiments. Swanson made seven medical discoveries by analyzing the medical literature. Swanson's discovery approach is based on two concepts: (1) complementary literature and (2) non-interactive literatures. If one set of articles X reports an association between concepts A and B, and a different set of articles Y reports an association between B and C, but nothing has been reported about a possible association between A and C, then X and Y are called complementary literature. If the readers of literature X are not acquainted with literature Y, then X and Y are non-interactive. This is often the case because of overspecialization. Swanson's concept of undiscovered public knowledge says that though literatures X and Y are available, the potentially new relation between A and C remains undiscovered and is a good source for hypothesis generation and hence new discoveries. Swanson developed software called ARROWSMITH that helps in automating some of the steps in analysis [23]. Swanson's discovery strategy has two steps. In the first step, for a given query or concept of interest 'A' (e.g. a disease) a set of document titles in Medline were analyzed and a list of terms 'B' (or phrases) were generated after removing useless words by using filters such as stop word list. 'A' and 'B' co-occur in the retrieved document titles. Then for each selected 'B' term, Medline document titles were retrieved and analyzed to produce a list of 'C' terms. Those 'C' terms that co-occur with 'A' are eliminated such that, 'B' and 'C' co-occur in the retrieved document titles but not 'A' and 'C'. The remaining 'C' terms represent possible new association with 'A', with 'B' as the intermediary concept linking 'A' and 'C'. The goal of the second step is to narrow down on the 'B' terms to find the most possible A-B-C link [8].

TransMiner uses the co-occurrence principle to discover associations among the objects from the Medline database. The novelty of TransMiner is the graphical representation of the discovered associations and the application of a transitive closure principle on the graph to find the potential transitive associations that may be candidates for discovering new associations, which are not discovered by conventional association discovery methods. The direct associations discovered by the association discovery technique as that of Stephens et al. [19] would not

be the complete set of associations, as the discovery was not made on all the Medline documents but on a limited subset. TransMiner finds these undiscovered direct associations by iterative retrieval of Medline documents and association discovery for each potential transitive association.

TransMiner also discovers transitive associations (new associations) that are not explicitly found in Medline, which are candidates for hypotheses generation. The current implementation of TransMiner assumes that Medline has both the set of complementary literatures. TransMiner finds many ranked AC associations at a time by using the transitive closure principle, while Swanson's approach involves lot of user interaction and it aims to find only few A-B-C type relations at a time. Swanson's approach aims to find only new associations whereas TransMiner helps in discovering both the existing explicit (direct) associations and also potential new associations (transitive associations).

TransMiner differs from tools based on Swanson's approach such as Arrowsmith [23] and LitLinker [15] in that user-defined dictionaries of terms are used in TransMiner and the direct (existing) associations and the transitive (potentially new) associations are discovered among these terms of interest as opposed to these tools where either the starting concept (A) or the starting and target concept (A and C) only are decided by the user. The linking concept (B) is obtained by Medline search. Further these tools do not represent the associations as graph and hence do not take advantage of graph properties such as transitive closure. Also neither Arrowsmith [23] nor LitLinker [15] provide a way to visualize the new associations discovered together with the known associations which improves the understanding of the molecular interactions. For those AC connections discovered by TransMiner, the Arrowsmith tool can be used to find the interlinking B concepts. (Note that one or more interlinking B concepts are already present in the user-defined dictionary which can be easily visualized using the graph visualization applet.)

Natural-language-processing-based text mining tools for extracting molecular connections from the biomedical literature, e.g. GENIES [6] and MedScan [4], can be complemented by TransMiner. They can use the principle of transitive closure on the extracted connections represented as a graph to discover potentially new connections (hypotheses).

Methods

Dictionary

A list of objects that the user is interested in and their interrelationships forms the dictionary. The dictionary can contain units of the same object, e.g. genes, or different objects together, e.g. genes, proteins, and drugs.

Data Source and Data Acquisition

The text data source for TransMiner is the Medline database. TransMiner uses a HTML wrapper (a parser) to retrieve title and abstract of each article from Medline for a given query. Documents related to objects of interest were collected from Medline by constructing the URL query with all objects in the dictionary using the 'OR' operator. So the wrapper will retrieve any document that contains any of the objects.

Document Representation

Salton's [17] vector space model was used for document representation. Document representation converts documents into numerical structures (document vectors) for efficient processing without loss of vital content. TransMiner uses a user-defined dictionary for representing each document. The dictionary helps to capture the essence of the document and to reduce the size of the document vectors. The vector space model attempts to compute the importance of terms on the basis of term frequencies within a document and within an entire document collection. The tf·idf (term frequency multiplied with inverse document frequency) algorithm is used for calculating term weights. Thus, each document vector consists of tf·idf weight of the terms in the dictionary given by the following formula.

$$W_{ik} = T_{ik} \cdot I_k = T_{ik} \cdot \log(N/n_k) \quad (1)$$

where T_{ik} is the number of occurrences of term T_k in document i, $I_k = \log(N/n_k)$ is the inverse document frequency of term T_k in the document set, N is the total number of documents in the document set, and n_k is the number of documents in the set that contain the given term T_k. The document vector is a weight vector whose size is the same as the number of terms in the dictionary and whose elements are the tf·idf weights of the corresponding terms.

Association Discovery

The method of Stephens et al. [19] was used to find the object-object association. The goal is to discover pairs of objects from a collection of documents such that the objects in each pair are associated in some manner. They described the differences between gene association discovery and association rule discovery in databases [19]. Association rule discovery is frequently based on transaction records stored in specific formats (such as relational databases), whereas the gene associations are discovered from natural language text. Commonly, database association rule discoveries are based on frequencies of individual items as well as the joint frequencies of pairs. Stephens et al. [19] considered both the relative 'importance' of each gene as well as the strength of their joint occurrences. After the vector representation of all documents is computed, the association between the two object terms k and l is computed as follows:

$$\text{association}\,[k][l] = \sum_{i=1}^{n} W_{ik} \cdot W_{il}\; k,l = 1...m \quad (2)$$

where n is the total number of documents and m is the number of objects in the document vector. W_{ik} denotes the weight of the k^{th}

object term. The computed association value is used as a measure of the degree of the relationship between the k^{th} and l^{th} object terms. This will result in an association matrix. For any pair of object terms co-occurring in even a single document, the association [k][l] will be non-zero and positive. The association matrix is a symmetric matrix. The non-zero and non-diagonal values from the matrix were used for creating the undirected association graph.

Transitive Association Discovery

Relations are ways in which things can stand with regard to one another or to themselves [7]. Relation R is transitive if R (x, y) and R (y, z) imply R (x, z). In symbols, R is transitive if and only if xyz ((Rxy Ryz) Rxz). For example, *if* nutrient A deficiency affects physiological process B *and* B causes disease C, *then* nutrient A deficiency and disease C are associated.

Transitive Closure

The transitive closure of a graph G is the graph G* (fig. 1) such that there is an edge from vertex A to vertex C in G* if there is a path from A to C in G. The traditional Warshall algorithm was used to compute the transitive closure of the association graph [26]. Given a directed graph G = (V, E) where, V is the set of vertices and E is the set of edges, represented by an adjacency matrix A[i, j], where A[i, j] = 1 if (i, j) is in E, compute the matrix P, where P[i, j] is 1 if there is a path of length greater than or equal to 1 from i to j. This algorithm extends paths by joining existing paths together. The transitive closure of a symmetric matrix (undirected graph) is also a symmetric matrix (undirected graph).

Mining Direct and Transitive Associations from Potential Transitive Associations

The newly discovered potential transitive associations must be checked to see if those associations are indeed 'direct' (explicitly found in any of the Medline document). We used an automated way (Algorithm 1) to find those associations that are direct, and that are transitive by submitting the two nodes (objects) of a potential transitive relationship to the Medline database with 'AND' operator in the query iteratively for all potential transitive object pairs. The documents will be retrieved only if both the objects are present in the document. For any pair of objects representing a potential transitive relationship, if the document set retrieved is non-zero, then by the principle of co-occurrence we can conclude that there exists a possibility of association between this object pair and that the association is direct. The association strength of these newly discovered 'direct' associations are given by the product of tf·idf weight of both nodes (objects) summed over all the documents retrieved for the object pair. The rest of the potential transitive associations with zero strength are implicit or transitive. These transitive associations are candidates for hypothesis generation. For these transitive associations, there are no documents in Medline at present that have both the objects in its content.

Algorithm 1: Transitive Association Discovery

(1) Potential transitive associations are the difference between the transitive closure (G*) and the initial association graph (G).

(2) Find the object pair for each potential transitive association and construct the Medline URL query using 'AND' operator.

(3) Retrieve documents for this object pair and calculate the association strength between the object pair.

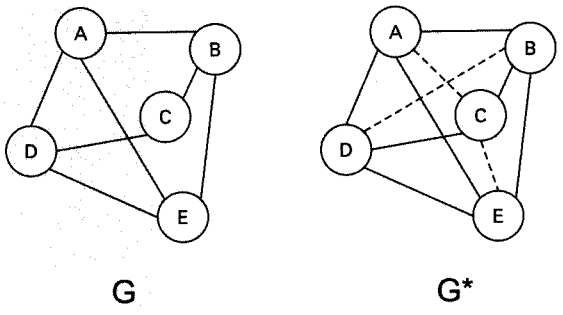

Fig. 1. Diagrammatic representation of transitive closure G* of a graph G. The transitive associations in G* are shown as dashed lines.

(4) If the association strength is not zero, the association is direct. Keep the object pair in G*.

(5) If the association strength is zero, the association is transitive. Remove the object pair from G*.

(6) Repeat steps 2, 3, 4, and 5 for all the potential transitive associations discovered to get G′ that contains the initial direct associations G and the newly discovered direct associations.

Validation of Associations

The goal of validation is to see if the associations discovered based on term co-occurrence in documents have biological or medical sense. Manual expert evaluation of all the direct associations discovered by TransMiner was performed by reviewing the abstracts for each association pair and determining if there exists a known biological association. Linking words and phrases including 'binds', 'involved in', 'function as', 'expressed in', 'is a subunit of', 'related to', 'required for', 'located in', 'interacts with', 'regulates'... were used. Normally the users (e.g. laboratory investigators) validate the results as they have expertise in the domain of their research.

Visualization of Associations

The associations were visualized in the form of a graph. A graph is a mathematical structure used to describe the relationships among different objects. Graph layout helps to understand the relationships among objects easily, which are not easy to interpret from raw data or by other means of visualization. Each object is represented as a node and the relationship as an edge between two objects. The classic network-mapping applet from Sun Microsystems was used to visualize the direct and transitive associations [20]. Several modifications were made to the original applet. Java Graphics2D API was used to draw edges with different thickness to represent the varying association strengths. All potential transitive edges are represented by dashed lines to indicate that these relationships are not found in the initial analysis of the document collection. The direct associations and the transitive associations were colored differently (fig. 2).

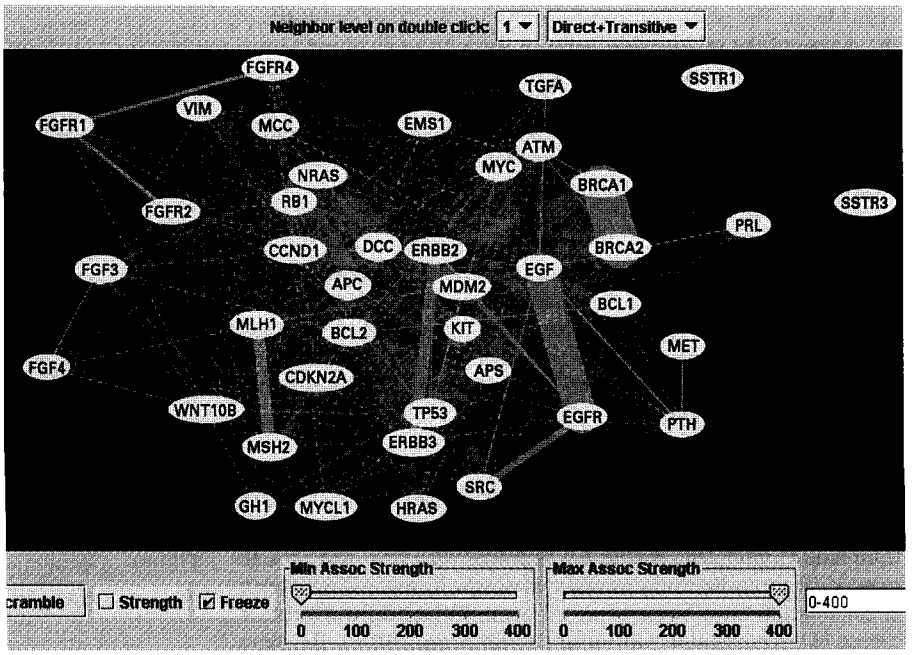

Fig. 2. The initial direct associations among 56 gene symbols based on 5,000 Medline documents, the direct associations discovered from potential transitive associations based on the presence of non-zero association in the Medline database and the transitive associations.

Association Filtering

Not all associations are interesting to all the users. Some users may be interested in gene-gene interactions while others may be interested in protein-protein interactions or protein-drug interactions. In TransMiner, the user will define the dictionary or 'objects' that he/she is interested in finding the associations (for example all objects involved in apoptosis). Similarly not everyone considers a given association interesting. The 'interesting' association is more subjective. Some users may be interested in associations that are obvious (associations with higher strength) and some users may be more interested in least obvious associations or outliers (associations with very less strength). Yet, other users may be interested in associations with moderate strength. One way to customize associations presented to different users includes use of the GUI (graphical user interface) to show only those associations that are above and below a certain association strength that the user prefers. The graph visualization applet has two sliders for selecting the upper and lower association strength values (fig. 2). Neighbor identification capability was implemented using a tree-scan algorithm as implemented by Mrowka [11]. Neighbor identification helps to see only those genes that are neighbors to a particular object. Multiple levels of neighbors can also be selected.

Ranking Transitive Associations

Ranking of the transitive associations that are new potentially meaningful associations will help the user to select associations (hypotheses) that can be further investigated in detail. Transitive association strength cannot be calculated directly as done in the case of direct associations, as there is no co-occurrence in any document between the nodes 'A' and 'C' of a transitive association. The transitive association strength is defined as the sum of weight of all words 'B' that co-occur with both nodes 'A' and 'C' of a transitive association (intersection of words that co-occur with A and words that co-occur with B). This is based on the idea that if there is a strong link in the form of A-B-C then the possibility of the AC association becoming true is increased.

Results

Three experiments were conducted to evaluate the performance of TransMiner. These experiments show how TransMiner can efficiently discover new associations by using the transitive closure principle.

Association Discovery among Breast Cancer Genes

A list of gene symbols related to breast cancer was made from the Baylor College of Medicine, the Breast Cancer Gene Database and the GeneCards database [1, 16]. Fifty-six official gene symbols were selected to form the dictionary (table 1). Analyzing 5,000 Medline documents discovered 87 direct associations among the 56 gene symbols. The gene pair brca1-brca2 has the highest association strength, as expected. By applying the transitive closure algorithm to the initial association graph, 655 potential transitive gene pair associations were obtained. TransMiner discovered 296 direct associations and 359 transitive associations out of 655 potential transitive associations by iterative retrieval and mining of Medline documents (fig. 2). Based on manual evaluation of the 87 initial gene pair associations discovered by TransMiner, 75

Table 1. Gene symbols of breast cancer genes

APC	APS	ATM	BCL1	BCL2	BRCA1	BRCA2	CCND1
CDKN2A	COL18A1	DCC	EGF	EGFR	EMS1	ERBB2	ERBB3
MSH2	MLH1	FGF3	FGF4	FGFR1	FGFR2	FGFR4	GH1
GRB7	HRAS	IGF1R	KIT	KRAS2	MYCL1	IGF2R	MCC
MDM2	MET	MYC	NF2	NRAS	PGR	PHB	PLAT
PLG	PRL	PTH	PTPN1	RB1	SSTR1	SSTR2	SSTR3
SSTR4	SSTR5	SRC	TGFA	TP53	TSG101	VIM	WNT10B

Table 2. Results of experiments

	Experiment 1	Experiment 2	Experiment 3
Objects	56	24	6
Initial direct associations	87	65	8
Valid initial direct associations	75 (86.21%)	65 (100%)	8 (100%)
Potential transitive associations discovered by transitive closure	655	188	7
Direct associations discovered from potential transitive associations	296	113	2

(86.21%) gene pairs were found to have some valid biological association (experiment 1 of table 2). Similarly, out of 296 direct gene pair associations discovered from potential transitive associations, 237 (80.06%) gene pairs were found to have a biological association based on expert evaluation. The 359 transitive associations were ranked based on the sum of weights of terms that co-occur with both the nodes of a transitive association.

Association Discovery among Objects in the Calpain Signal Transduction Pathway

Objects involved in the hypothetical calpain signal transduction pathway (table 3) were used as terms in the dictionary. Analyzing 5,000 Medline documents discovered 65 direct associations among 19 objects. By applying the transitive closure algorithm to the initial graph, 188 potential transitive associations were obtained. TransMiner discovered 113 direct associations and 75 transitive associations out of 188 potential transitive associations by iterative retrieval and mining of Medline documents. All 65 initial associations (100%) were found to have some valid biological association (experiment 2 of table 2). Out of 113 direct associations discovered from potential transitive associations, 105 (92.92%) object pairs were found to have biological association based on expert evaluation. For example, the protein kinase A (PKA)-diabetes connection was not found by analyzing the initial set of 5,000 documents. But there are documents in Medline that show an association between PKA and diabetes, e.g. '... role of protein kinase A (PKA) signaling events in mediating diabetes associated with obesity. PMID: 11679434'. The 75 transitive associations were ranked based on the sum of weights of terms that co-occur with both nodes of a transitive association.

Table 3. Object terms involved in the calpain signal transduction pathway

SNARE	RAS	SOS	GRB2	CALPAIN
SNAP23	CAMP	AMP	CA	GLUT4
EGFR	ERK	RAF	MEK	INSULIN
SNAP	PKA	AKT	M-CALPAIN	

Rediscovery of the Magnesium-Migraine Association by TransMiner

The ultimate validation of TransMiner would be to assist in making actual scientific discoveries that could be published. This experiment was conducted as an indirect way to prove the validity of TransMiner by repeating Swanson's famous discovery of the relationship between magnesium and migraine headaches by going back in time. Six terms – stress, magnesium, migraine, platelet, depression and calcium – were used as dictionary. Eight initial associations (table 4) were found among six objects by mining the Medline documents published from 01/01/1900 to 31/12/1969. These documents have no co-occurrence of magnesium and migraine.

Table 4. Initial direct associations discovered among six objects related to magnesium and migraine by mining Medline documents published from 01/01/1900 to 12/31/1969 and their association strength

No.	Object-object pair	Association strength	Documents (01/01/1900–12/31/1969)	PubMed sentence with PMID (01/01/1900 to 12/31/1969)
1	Stress-migraine	0.66	10	Plasma serotonin in migraine and stress. 5297855
2	Platelet-migraine	0.73	3	Platelet 5-hydroxytryptamine and adenine nucleotides, serum arginylesterase and plasma 11-hydroxycorticosteroids in migraine. 5653687
3	Platelet-magnesium	1.31	37	Study of platelet adhesiveness and serum magnesium levels in cases of acute myocardial infarction. 5364354
4	Depression-magnesium	1.32	200	Plasma magnesium and calcium in depression. 5358528
5	Calcium-magnesium	115.39	1,970	Plasma magnesium and calcium in depression. 5358528
6	Calcium-stress	1.43	38	Calcium. II. Sterols, steroids and urinary calcium. Is nephrocalcinosis caused by stress? 6078954
7	Calcium-platelet	3.14	89	Platelet aggregation. II. Adenyl cyclase, prostaglandin E1, and calcium. 4310667
8	Calcium-depression	3.18	229	Plasma magnesium and calcium in depression. 5358528

All eight initial object pair associations were found to be valid based on manual evaluation.

Table 5. Potential transitive associations

No.	Documents	Object-object pair of potential transitive association	Association strength by analyzing documents from 01/01/1900 to 12/31/1969	PubMed sentence with PMID (01/01/1900 to 12/31/1969)
1	0	Migraine-magnesium	0.0	No co-occurrence
2	13	Migraine-depression	0.0	No co-occurrence in title/abstract
3	1	Migraine-calcium	0.0	No co-occurrence in title/abstract
4	25	Magnesium-stress	1.18	Relationship between the changes of potassemia and the plasmatic magnesium level in stress. 5699346
5	12	Stress-platelet	0.66	Blood platelet fluctuation consecutive to trauma and surgical intervention in the rat, a possible stress effect. 5338736
6	122	Stress-depression	0.0	No co-occurrence in title/abstract
7	45	Platelet-depression	0.0	No co-occurrence in title/abstract

Out of seven potential transitive associations, two direct associations were found automatically by iterative retrieval and association discovery for each potential transitive object pair. The remaining five associations are transitive. By manual evaluation both the newly discovered direct associations were found to have some biological association.

Seven potential transitive associations were found by applying transitive closure to the initial association matrix. From these seven potential transitive associations, two direct associations and five transitive associations were found including the association between magnesium and migraine (table 5). For transitive associations other than between magnesium and migraine, documents were retrieved from Medline, but the association strength is zero as none of those documents retrieved has both the objects in the title or abstract (mostly found in medical subject headings, MeSH, terms or other fields). The transitive association between magnesium and migraine was proven to be valid later by scientific experiments [5].

Discussion

The massive growth of textual scientific information in databases such as Medline requires novel ways to extract the underlying knowledge. TransMiner is built upon the association discovery method of Stephens et al. [19] with enhancements such as the transitive association discovery, association filtering and the graph visualization. TransMiner helps to discover the existing associations and to predict potentially new meaningful associations. The approach by TransMiner of using partial knowledge and existing knowledge to infer unpublished conclusions is a promising strategy for the future development of text-based discovery systems.

The fact that TransMiner can find the magnesium-migraine association discovered by Swanson with only six objects in the dictionary suggests the importance of having a concise and 'quality' dictionary that clearly represents the user interest. Extracting important terms from the documents that match the users' interest can indirectly generate this dictionary. These documents can be obtained by using a profile-based information filtering system such as BioSifter [13]. The advantage of using a customized dictionary rather than a comprehensive dictionary (e.g. all genes, proteins, molecules and other object terms) is that it reduces noise and generates associations that are interesting to the user quickly. Also, since the user has knowledge in the domain of search he/she can validate the associations with ease.

TransMiner was used to identify associations among objects in the areas of diabetes and cell motility/cytoskeletal remodeling with the help of a user-defined dictionary. The calpain family of proteases, well studied in the area of ischemic stroke, brain trauma and myocardial infarction, has recently been linked to the fields of type 2 diabetes genetics and cancer-related cell motility. Cellular processes relevant to diabetes and cancer cell invasion that have been linked to calpain have included secretion and motility-associated cytoskeletal remodeling and signaling. In this study, the dictionary terms include known calpain substrates and the signaling proteins that are associated either upstream or downstream of calpain. The discovered direct associations were further filtered with the help of GUI to visualize associations that have an association strength of 2–10 to remove highly studied associations (>10) and preliminary associations (<2). This filtered visualization left us with the areas of epidermal growth factor receptor (EGFR)-mediated calpain activation, calpain activation of NF-κB signaling and SNAP/SNARE-mediated regulation of secretion as active areas of investigation. Transitive associations were found most strongly in the area of SNAP/SNARE-mediated secretion. TransMiner was therefore able to use a set of dictionary terms that were not organized to derive a map of most actively investigated areas and points to the area with potential for future discovery.

One of the transitive associations (A-C) predicted (before March 31, 2003) is that there is a possible connection between MEK and spectrin. No document that has both the terms MEK and spectrin was retrieved from PubMed at that time. A single subsequent article established the validity of this transitive connection between the MEK pathway and calpain from empiric data [3]. '... MEKK1 is required for activation of the cysteine protease calpain and cleavage of spectrin and talin ...'.

TransMiner found the following A-B and B-C connections and predicted A-C.

(1) (A-B) MEK-calpain connection:

Example: '... calpain was stimulated by transfection of constitutively active MEK ...'.

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd = Retrieve&db = PubMed&list_uids = 10644690&dopt = Abstract.

(2) (B-C) Calpain-spectrin connection:

Example: 'Tyrosine phosphorylation regulates alpha II spectrin cleavage by calpain.'

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd = Retrieve&db = PubMed&list_uids = 11971983&dopt = Abstract.

In case of the magnesium-migraine connection, the magnesium-platelet association and platelet-migraine association led to hypothesize that there could be a magnesium-migraine association. Later it has been proven experimentally.

As mentioned in the LitLinker paper [15], evaluating knowledge discovery systems is a challenging task because if they are successful, by definition they are capturing new knowledge that has yet to be proven useful. So we used the idea of going back and forth in time to prove that the knowledge discovered in the past by TransMiner are actually found to be true in the present (as reported in the scientific literature).

The new paradigm in cancer therapeutics is to identify and target pathways or systems rather than single molecules. A recent study showed that a small GTPase involved in actin remodeling, CDC42, is bound to the EGFR through the EGFR-associated c-Cbl protein and a previously identified docking protein, Cool-1, i.e. a (CDC42-Cool-1)-(c-Cbl-EGFR) complex [28]. Cool-1 was previously known to be associated with CDC42. c-Cbl

was previously known to be associated with the EGFR. c-Cbl is associated with Cool-1. What was not known is the link between CDC42 and EGFR. Much of the research in molecular biology is about finding links between signal transduction pathways that cause cross-talk rather than cause-effect relationships. The discovery of these cross-talk pathways is critical because it leads us to a better understanding of cancer progression and new potential therapeutic targets.

TransMiner can be used for analysis of DNA microarray data to study gene expression. In a typical drug experiment, where we look at changes in gene expression related to the drug, we may see 1,000 genes significantly altered in expression. In order to make sense of this massive amount of data you need to have web-based software tools. One such tool is a filter that selects significantly altered gene expression ($p < 0.05$ for quadruplicate measurement and 1.2-fold change up or down) by metabolic or signaling pathway. Once the user has used the filter, e.g. proteases, he/she may still be left with 30 genes that are altered in expression. Invariably, there are genes that are unexpected and the potential for connection with the literature is unknown. We can use LocusLink (Entrez, National Library of Medicine) or Gene Card (Weitzman Institute) to find the full name of the gene and something about the pathways in which the gene lies. It is clear that an additional filter is needed to identify connections between the unexpected genes and the pathway in question. TransMiner is an excellent tool to perform this function. Again, the search has a beginning that is defined, with the pathway used to sort the genes. TransMiner can be used to query each of the unexpected genes and link it back to the pathway.

Quantification of the associations discovered by TransMiner depends on the input objects and the current associations that exist among those objects. One approach is to define the strength of direct associations into three classes: e.g. weak (1–5), moderate (5–10) and strong (>10). This can be achieved by using the slider in the visualization applet. Weak associations are the ones that the investigator may be reluctant to explore because of a lack of corroboration. Moderate associations may be of greatest interest, because there is enough corroboration to suggest that the area of investigation may be useful for the testing of novel hypotheses. Objects that are strongly associated will require judgment of the investigator as to whether there are truly novel hypotheses worth testing. TransMiner could help less experienced investigators steer in unexplored directions.

Though transitive associations are ranked, quantification of transitive associations will be very difficult. Because, for a set of objects given by the user as a query, there might not be any transitive associations at all or they will produce interesting transitive associations. It depends on the input objects and the direct associations that exist among those objects.

Medline converts the 'user query' into 'Medline query' based on MeSH terms and other dictionaries. For example, the query 'TP53 AND EGFR' will be converted into: '(('genes, p53'[MeSH Terms] OR TP53 [Text Word]) AND (('receptor, epidermal growth factor'[MeSH Terms] OR 'genes, erbb-1'[MeSH Terms]) OR EGFR [Text Word]))'. So, sometimes the documents retrieved may not have the gene symbols that the user query has. Also, the query words may not be present in the title or the abstract of the retrieved documents but in other fields such as MeSH terms. This results in some of the inconsistencies observed that could be removed by fine-tuning the user query.

We plan to improve the performance of TransMiner by storing all the documents available at Medline in a local database that updates regularly. This will drastically reduce the amount of time taken during the discovery process especially in the iterative retrieval and analysis stage. Also storing all expert-verified associations for future use will avoid repeating the same process again and again. We plan to use a thesaurus or an ontology-based dictionary to improve the quality of associations discovered as the objects are represented in multiple forms in the literature (e.g. p53 and TP53). We also plan to use natural language processing for the identification of objects without ambiguity (e.g. KIT gene symbol and other occurrence of 'kit' such as Abbott's ER-EIA Kit, DNA Sequencing Kit, etc.). We are also examining the possibility of parallelizing the transitive association discovery process to improve the speed of analysis. Another area of interest is extraction and inference of the nature of a relationship between the objects. Integrating TransMiner with a profile-based information-filtering system such as SIFTER will help to dynamically map the knowledge that the user is interested in [12].

It is important to take note of the limitation of association discovery tools that are based on the principle of co-occurrence. Two objects may co-occur for many reasons and there may not be a biologically meaningful association always. Also, manual evaluation of discovered associations may be subjective.

Like any other informatics approach, TransMiner cannot bypass scientific experiments, but can help to discover

knowledge from scientific literature for generating hypotheses. Recycling of existing data, using approaches such as TransMiner, will help to increase efficiency and productivity. The transitive associations that are not explicitly found in any of the Medline document at present may be discovered by scientific experiments in the future.

Acknowledgments

The research reported in this paper was supported in part by National Science Foundation Information Technology Research grant No. NSF-IIS/ITR 0081944, NIH BISTI grant, NIH-NIGMS P20 GM66402, Clarian Values Foundation Grant, and a grant from the Thoracic Oncology Program of the Indiana University.

References

1 Baasiri RA, Glasser SR, Steffen DL, Wheeler DA. The Breast Cancer Gene Database: A collaborative information resource. Oncogene 18: 7958–7965;1999. http://tyrosine.biomed-comp.com/4d.acgi$tsrchname?Name = &topic = BCIR.

2 Bao JJ, Le XF, Wang RY, Yuan J, Wang L, Atkinson EN, LaPushin R, Andreeff M, Fang B, Yu Y, Bast, RC Jr. Reexpression of the tumor suppressor gene ARHI induces apoptosis in ovarian and breast cancer cells through a caspase-independent calpain dependent pathway. Cancer Res 62:7264–7272;2002.

3 Cuevas BD, Abell AN, Witowsky JA, Yujiri T, Johnson NL, Kesavan K, Ware M, Jones PL, Weed SA, DeBiasi RL, Oka Y, Tyler KL, Johnson GL. MEKK1 regulates calpain-dependent proteolysis of focal adhesion proteins for rear-end detachment of migrating fibroblasts. EMBO J 22:3346–3355;2003.

4 Daraselia N, Yuryev A, Egorov S, Novichkova S, Nikitin A, Mazo I. Extracting human protein interactions from MEDLINE using a full-sentence parser. Bioinformatics 20:604–611; 2004.

5 Demirkaya S, Vural O, Dora B, Topcuoglu MA. Efficacy of intravenous magnesium sulfate in the treatment of acute migraine attacks. Headache 41:171–177;2001.

6 Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics 17(suppl 1):S74–S82;2001.

7 Honderich T. The Oxford Companion to Philosophy, Oxford University Press. 1995 http://www.xrefer.com/entry/553381.

8 Hristovski D, Dzeroski S, Peterlin B, Rozic-Hristovski A. Supporting Discovery in Medicine by Association Rule Mining of Bibliographic Databases. Proc Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases. Berlin, Springer, 446–451;2000.

9 Jenssen TK, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. Nature Genetics 28:21–28;2000.

10 Mathiasen IS, Sergeev IN, Bastholm L, Elling F, Norman AW, Jaattela M. Calcium and calpain as key mediators of apoptosis-like death induced by vitamin D compounds in breast cancer cells. J Biol Chem 277:30738–30745; 2002.

11 Mrowka R. A Java applet for visualizing protein-protein interaction. Bioinformatics 17: 669–671;2001.

12 Mukhopadhyay S, Mostafa J, Palakal M, Lam W, Xue L, Hudli A. An Adaptive Multi-level Information Filtering System. Proceedings of the Fifth International Conference on User Modeling, 21–28;1996.

13 Palakal M, Mukhopadhyay S, Mostafa J, Raje R, N'Cho M, Mishra S. An intelligent biological information management system. Bioinformatics 18:1283–1288;2002.

14 Pink JJ, Wuerzberger-Davis S, Tagliarino C, Planchon SM, Yang X, Froelich CJ, Boothman DA. Activation of a cysteine protease in MCF-7 and T47D breast cancer cells during beta-lapachone-mediated apoptosis. Exp Cell Res 255:144–155;2000.

15 Pratt W, Yetisgen-Yildiz M. LitLinker: capturing connections across the biomedical literature. Proceedings of the International Conference on Knowledge Capture 105–112;2003.

16 Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: encyclopedia for genes, proteins and diseases. Weizmann Institute of Science, Bioinformatics Unit and Genome Center (Rehovot, Israel) 1997. http://bioinformatics.weizmann.ac.il/cards.

17 Salton G. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Reading, Addison-Wesley, 1989.

18 Stapley BJ, Benoit G. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. Pac Symp Biocomput 5:529–540;2000.

19 Stephens M, Palakal M, Mukhopadhyay S, Raje R, Mostafa J. Detecting gene relations from Medline abstracts. Pac Symp Biocomput 483–495;2001.

20 Sun Microsystems, Inc. Graph.java demonstration software. Santa Clara, Sun Microsystems, 1995. http://java.sun.com/applets/jdk/1.0/demo/GraphLayout/index.html.

21 Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med 30:7–18;1986.

22 Swanson DR. Migraine and magnesium: eleven neglected connections. Perspect Biol Med 31:526–557;1988.

23 Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artif Intell 91: 183–203;1997.

24 Tagliarino C, Pink JJ, Dubyak GR, Nieminen AL, Boothman DA. Calcium is a key signaling molecule in beta-lapachone-mediated cell death. J Biol Chem 276:19150–19159;2001.

25 Tagliarino C, Pink JJ, Reinicke KE, Simmers SM, Wuerzberger-Davis SM, Boothman DA. Mu-calpain activation in beta-lapachone-mediated apoptosis. Cancer Biol Ther 2:141–152; 2003.

26 Warshall S. A theorem on boolean matrices. J ACM 9:11–12;1962.

27 Wong PK, Whitney P, Thomas J. Visualizing Association Rules for Text Mining. Proceedings of IEEE Information Visualization, 120–123;1999.

28 Wu WJ, Tu S, Cerione RA. Activated Cdc42 sequesters c-Cbl and prevents EGF receptor degradation. Cell 114:715–725;2003.