Maciel et al. BMC Genomics 2011, 12(Suppl 4):S1
http://www.biomedcentral.com/1471-2164/12/S4/S1

BMC Genomics

PROCEEDINGS Open Access

# Can the vector space model be used to identify biological entity activities?

Wesley D Maciel[1,2], Alessandra C Faria-Campos[3], Marcos A Gonçalves[3], Sérgio VA Campos[3*]

*From* 6th International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-meeting 2010)
Ouro Preto, Brazil. 15-18 November 2010

## Abstract

**Background:** Biological systems are commonly described as networks of entity interactions. Some interactions are already known and integrate the current knowledge in life sciences. Others remain unknown for long periods of time and are frequently discovered by chance. In this work we present a model to predict these unknown interactions from a textual collection using the vector space model (VSM), a well known and established information retrieval model. We have extended the VSM ability to retrieve information using a transitive closure approach. Our objective is to use the VSM to identify the known interactions from the literature and construct a network. Based on interactions established in the network our model applies the transitive closure in order to predict and rank new interactions.

**Results:** We have tested and validated our model using a collection of patent claims issued from 1976 to 2005. From 266,528 possible interactions in our network, the model identified 1,027 known interactions and predicted 3,195 new interactions. Iterating the model according to patent issue dates, interactions found in a given past year were often confirmed by patent claims not in the collection and issued in more recent years. Most confirmation patent claims were found at the top 100 new interactions obtained from each subnetwork. We have also found papers on the Web which confirm new inferred interactions. For instance, the best new interaction inferred by our model relates the interaction between the adrenaline neurotransmitter and the *androgen receptor* gene. We have found a paper that reports the partial dependence of the antiapoptotic effect of adrenaline on *androgen receptor*.

**Conclusions:** The VSM extended with a transitive closure approach provides a good way to identify biological interactions from textual collections. Specifically for the context of literature-based discovery, the extended VSM contributes to identify and rank relevant new interactions even if these interactions occcur in only a few documents in the collection. Consequently, we have developed an efficient method for extracting and restricting the best potential results to consider as new advances in life sciences, even when indications of these results are not easily observed from a mass of documents.

## Background

In a biological system there are entities of different types such as diseases and drugs performing important biological activities. The action of an entity can mediate or interfere with the action of other entities developing a complex network of interactions. Frequently entities perform more than one activity in the system, some which are known and integrate the current knowledge in life sciences. Other activities are not so well documented or remain unknown for long periods of time and are generally discovered by chance. Drugs, for instance, have a primary pharmacological activity and secondary activities responsible for side effects. However, drug side effects can be explored as new uses for the treatment of different diseases. A remarkable example is the impotence drug sildenafil citrate (Viagra®)

* Correspondence: scampos@dcc.ufmg.br
[3]Computer Science Department of the Universidade Federal de Minas Gerais, Belo Horizonte, 31270-901, Brazil
Full list of author information is available at the end of the article

that was originally designed for the treatment of angina and hypertension. Viagra® clinical trials revealed, nevertheless, the drug ability of increasing erectile function as its side effect [1].

On the other hand, research achievements in the post genomic age have promoted an enormous and continuous increasing on biological knowledge. These achievements often describe biological entity activities and have been published around the world aiming to assist, increase and speed up the number of discoveries in life sciences. A similar process has occurred since the inception of the World Wide Web and the rise of digital libraries. Web pages have been continuously and rapidly published given rise to a enormous amount of interlinked information. This allowed the conduction of many studies about methods for extracting and analysing the information published in this ocean of information. In many of these studies the vector space model (VSM) [2-4] has been recognized as an important tool to extract the most relevant information in a given context.

In this work we have developed an inference model based on the VSM in order to predict new interactions between biological entities of distinct categories such as ecosystems, organisms, organs, tissues, cells, organelles, genes, proteins, diseases and drugs. Our model constructs a network of known entity interactions from a textual collection. The documents in this collection describe the current knowledge in life sciences. Known entity interactions represent entity co-occurrences in at least one document of the textual collection. After finding all known interactions, our model traverses and analyzes the network predicting new entity interactions. Our objective is to use the known interactions to infer new (unknown) ones and to rank all found interactions. The ranking of interactions allows researchers to focus in the most promising activities, thus promoting further advances in life sciences.

The prediction of new interactions is performed using the VSM along with a transitive closure similar to that used in literature-based discovery [5]. The transitive closure relies on the fact that "**IF** an entity $x$ interacts with entities $y$ and $w$ **AND** another entity $z$ also interacts with entity $y$, **THEN** $z$ probably also interacts with entity $w$". Different from previous work, in our model we adapt this transitive closure in order to exploit the primary and secondary activities performed by entities of distinct biological categories. In the context of our model, $x$ and $z$ are entities of the same biological category, $y$ and $w$ are also entities of the same biological category. However, the category of entities $x$ and $z$ is different of that of entities $y$ and $w$.

We have implemented a system called *BioSearch* [6] as a proof of concept of our model. The system deals with 4 types of distinct entities: diseases, drugs, genes, and targets. The textual collection used in the system encompasses a sample of 17,830 patent claims gathered from the United State Patent and Trade Mark Office (USPTO) [7]. We have used the patent claim because it is an important section in patent specifications, presenting the invention and defining the scope of patent protection [7,8]. From 266,528 possible interactions between entities in our network, the system has found 1,027 known interactions in the patent claim collection and has inferred 3,195 new interactions. Thus, based on our model, the system has constructed a network with 4,222 interactions that can be further analyzed in order to promote new advances in science and technology.

To validate our results we have conducted an experiment over the patent issue dates. We have reconstructed the interaction network in a range of 30 years. We have observed that new interactions found in a given past year were confirmed by patents issued in a more recent date. For instance, we have 1 patent claim issued in 2005 specifying the interaction between the disease heart attack and the gene *ppar-gama*. When we removed this patent claim from the textual collection, 61 patent claims indicated this interaction as a possible new interaction in 2004. We have also found scientific papers that confirm some of the new inferred interactions. For instance, the best result found in our model specifies a new interaction between the adrenaline neurotransmitter and the *androgen receptor* gene in the 2-dimensional subnetwork *gene* × *target*. No patent claim in our collection indicates this interaction. However, Sastry et al. [9] reported in 2007 that the antiapoptotic effect of adrenaline partially depends on *androgen receptor*.

### Related work

In this work, our objective is to present a model that employs the VSM in order to identify biological entity activities from a textual collection. In our approach, known entity activities represent entity co-occurrences in the textual collection. On the other hand, new activies are predicted from the known ones. Jenssen et al. [10] show that co-occurrence reflects biologically meaningful relationships, thus providing an approach to extract and structure known biological knowledge. Accordingly, we have developed a strategy based on the VSM that constructs a network of biological entity interactions from the life science literature and ranks these interactions. Our strategy combines the VSM ability to extract knowledge from text along with some underlining principles of literature-based discovery [11]. Don R. Swanson has pioneered the work in the field of literature-based discovery using the syllogism $x \rightarrow y$ AND $y \rightarrow z$ THEN $x \rightarrow z$ in order to discover new

biological entity activities [12,13]. In this syllogism $x \rightarrow y$ and $y \rightarrow z$ are known interactions stated in the literature. On the other hand, $x \rightarrow z$ is a new interaction not explicitly found in the literature and inferred from previously known interactions. Afterwards, Smalheiser et al. [14] have implemented this syllogistic construction in a software called ARROWSMITH. In addition, Weeber et al. [15] have contributed to literature-based discovery introducing a model based on natural language processing (NLP) techniques in order to find concepts in the biomedical literature and reduce the search space. None of these techniques associates weights with these biological interactions in order to rank them.

As mentioned, a challenge we face when dealing with literature-based discovery is how to rank a large number of inferred interactions in a way that can facilitate new discoveries by prioritizing the ones with the largest potential. In order to tackle this challenge, Swanson et al. [16] have proposed and tested strategies to rank and filter the output of the ARROWSMITH system. Hristovski et al. [5,17-19] presented a method for literature-based discovery based on association rules and implemented it in a system called BITOLA. Moreover, Wren et al. [20] have considered the construction of networks from the biomedical literature describing a method based in the syllogism proposed by Swanson. They have defined areas of research interest such as genes and diseases, and model and rank the interactions using the fuzzy set theory. The ranking strategies used in these works consider that entities co-occurring frequently in a textual collection are more likely to represent biologically meaningful relationships [10]. Therefore, these strategies promote new interactions which are predicted from a large number of indications. However, in literature-based discovery there are many distinct scenarios and in some situations a great number of indications may not reveal the most relevant new interactions. For instance, many indications may lead to a set of new interactions that were already studied but were not published because they are not feasible or they are unwanted in practice. On the other hand, there exist situations in which new interactions predicted from a few number of indications are in fact the ones with the best potential. In this sense, ranking strategies for new interactions predicted from few indications are an important tool for literature-based discovery because they help in the identification of relevant interactions not easily observed and extracted from textual collections. In this scenario the VSM provides a great aid to the literature-based discovery. The TFIDF weighting strategy exploited in the VSM promotes interactions with many occurrences in few documents in the collection and penalizes interactions commonly occurring in many documents of the collection. Consequently, the VSM fosters rare interactions over the trite ones.

In literature-based discovery we must avoid the inference of interactions already stated in the literature. Kostoff [21,22] has discussed this problem and issues related to the quantity and quality of interactions. Kostoff et al. [23] have presented a generic methodology for literature-based discovery and have used this methodology to identify interactions concerning Raynaud's phenomenon [24], cataracts [25], Parkinson's disease [26], multiple sclerosis [27] and water purification [28]. Kostoff et al. [29] have also compiled the lessons learned in these experiments and presented guidelines for further research. However, in this series of works the authors have not used any numerical filter to rank the new interactions found.

We have also to cope with the coverage problem when looking for biological entity activities by searching several information sources such as experimental data [30,31], drug labels [32], scientific papers [5,12-20] and patents. Patents are very important instruments of knowledge transfer and researchers commonly resort to this literature because its great value as a source of strategic, technical and business-related information [33,34]. Trippe [35], for example, described *patinformatics* as the science of analyzing patent information to discover relationships and trends. Mukherjea et al. [36] developed a system to retrieve information from biomedical patents. Larkey [37] described the patent retrieval and classification system developed for the USPTO. Fall et al. [38] evaluated the best ways to deal with patent classification and presented a comparison of the classification effectiveness of several algorithms in this task. Tseng et al. [34] described and evaluated several text mining techniques to create patent maps and improve patent analysis tasks such as classification and knowledge sharing. Particularly, the claim section is considered the most important section in patent specifications [7,8]. Thus, Shinmori et al. [8] proposed a framework to represent the structure of the patent claim section and a method to automatically analyze it. Accordingly, here we also explore patents, more specifically, the patent claim section, along with our proposed model in order to discover new biological entity interactions of potential interest.

### Main contribution

We have created a model to construct networks of entity interactions from the biological literature with the objective of finding known and new entity activities in a biological system. In our model we have used VSM to identify already known entity interactions. In addition, we have extended the VSM with a transitive inference process capable of predicting new entity interactions.

The networks are formed by subnetworks of interactions between entities of distinct categories. The advantage of using categories is the ability to restrict the research space for interactions between entities of specific categories and promote more accurate results.

Interactions are initially established in a network by entity co-occurrences in a textual collection. These interactions represent known interactions already described in the literature. The known interactions receive a weight corresponding the interaction level between entities based on the similarity value derived from the application of the VSM. The advantage of using the VSM is to explore its well documented algebraic framework for information retrieval from textual collections in order to find the entity co-occurrences and also measure their interaction levels. The VSM contributes for literature-based discovery by helping to predict the best new potential interactions not easily extracted from textual collections. The VSM also helps in situations in which entities rarely co-occurring in a document set are the ones with the potential best contributions for a researcher.

Our model uses the interactions established in the network to predict new interactions based on the transitive closure that we have employed in the inference process. The transitive closure states that "**IF** an entity $x$ interacts with entities $y$ and $w$ **AND** an entity $z$ interacts with entity $y$ **THEN** $z$ may also interact with $w$". Differently from previous work, entities satisfying the transitive closure must always follow a constraint. The constraint imposes that $x$ and $z$ are entities of the same biological category $C_1$, $y$ and $w$ are entities of another category $C_2$, and that $C_1$ and $C_2$ are distinct categories ($C_1 \neq C_2$). This constraint gives rise to the subnetworks that form the network of interactions. The main advantage of using this constraint is to narrow the research space of entity interactions promoting more accurate results. New interactions also receive a value for their interaction levels, based on the interaction levels of the interactions satisfying the transitive closure, as will be detailed later. This makes it possible to rank all entity activities in the network. The main advantage of ranking the network interactions is to reduce the human effort spent in their analysis, by focusing in the ones with the largest potential.

We have implemented the model in a system called *BioSearch* which uses a textual collection formed by patent claims. In our system, users can search all interactions established in the network (Appendix A [additional file]). Searching known entity interactions, users have a representation of the prior knowledge in a given subject that can be extracted from patent literature. These interactions are very important because they present a description of the current knowledge, avoiding patent infringements. On the other hand, users can search new interactions and have a representation of possible new technologies that may yet receive patent protection.

In the present work our goal is not to ensure a comprehensive coverage of the biological literature. Instead, we provide a proof of concept demostrating the applicability of our model in disclosing and ranking new entity interactions. For this, we have used a small textual collection to assess the model. Many new interactions inferred by our model based on this collection may have already been reported in scientific papers, thus validating our results.

## Network construction

Entities in a biological system interact with each other forming an interaction network. We can classify these entities into categories such as diseases, drugs, genes, and targets. In this work we have combined these categories in order to construct a network composed of n-dimensional subnetworks. We have extracted all entity interactions of a subnetwork from a textual collection using the VSM. Given, for instance, the subnetwork with dimensional space *drug* × *disease*, consider that our model indicates we have documents reporting the use of drug $m_1$ in the treatment of diseases $d_1$ and $d_2$ (Figure 1 (a) ). Moreover, suppose the model also indicates we have documents which report the use of drug $m_2$ in the treatment of disease $d_1$. Then, drugs $m_1$ and $m_2$ possibly share some common characteristic responsible for the efficacy of these drugs in the treatment of both diseases $d_1$ and $d_2$. Thus, the model infers a new connection in the subnetwork *drug* × *disease* linking drug $m_2$ and disease $d_2$. The new connection represents a new use of drug $m_2$. Then, in this example, $m_1 \rightarrow d_1$, $m_1 \rightarrow d_2$, and $m_2 \rightarrow d_1$ are known interactions found in the literature. On the other hand, $m_2 \rightarrow d_2$ is a new interaction inferred from the previous three known interactions.

We have represented each subnetwork as a weighted graph whose weights measure the interaction level of the entities based on the textual collection. In this graph, nodes are entities of categories forming the subnetwork dimensional space, edges represent interactions between entities of distinct categories, and the interaction level is a value in the range [0, 1]. We determine the interaction level based on the VSM when we look for the entity co-occurrences throughout the textual collection. In a subnetwork with dimensional space *drug* × *disease*, for instance, suppose that drug $m_1$ treats diseases $d_1$ with interaction level $A$ and $d_2$ with interaction level $B$ (Figure 1 (b)). In addition, suppose drug $m_2$ treats disease $d_1$ with interaction level $C$. Then, the model assigns an interaction level to the new connection

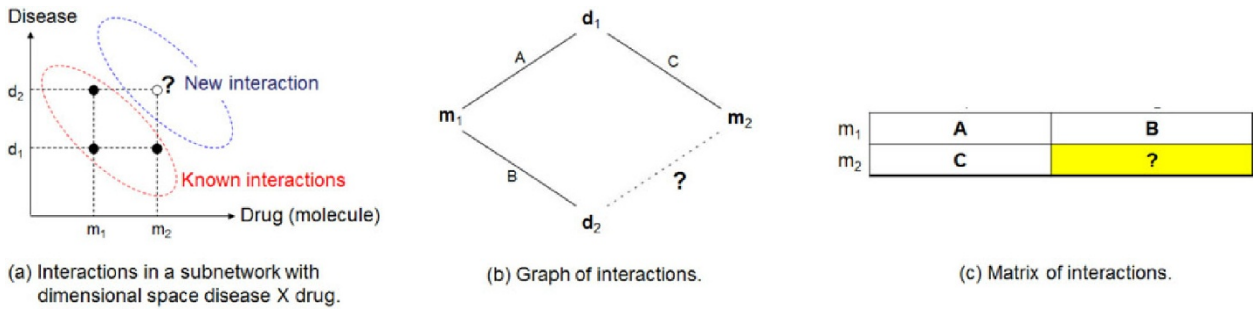


**Figure 1** Representations of entity interactions in a 2-dimensional subnetwork.

linking drug $m_2$ and disease $d_2$ whose value is determined based on *A*, *B* e *C*.

The graph in our model is represented by a matrix that receives biological entities of the subnetwork dimensions in its lines and columns (Figure 1 (c)). We have defined that three interactions in the matrix are in transitive closure when they satisfy the condition $(x, y)$ *and* $(x, w)$ *and* $(z, y) \rightarrow (z, w)$ that means "**IF** entity $x$ interacts with entities $y$ and $w$ **AND** entity $z$ interacts with entity $y$ **THEN** $z$ may also interact with $w$". Then, the model infers a new interaction in the matrix whenever it finds three interactions satisfying this transitive closure.

All cells in the matrix initially receive the value 0 indicating no entity interactions (Figure 2 (a)). We use the entities of a cell in order to form a query. This query represents a conjunction of entities of distinct categories. The conjunction is important because it ensures that documents in which the entities occur are not orthogonal, i.e., they must have occurrences of all entities present in the query. Then, we perform searches in the textual collection in order to find documents satisfying the query of each matrix cell (Figure 2 (b) ).

The VSM assigns weight values for each query entity based on the TFIDF strategy (Equation 1). We use these weights to measure the importance of the entity for a query of the matrix and also for a document of the textual collection.

$$w_{x,i} = tf_{x,i} \times idf_x = \frac{f_{x,i}}{max_{j,i}} \times log(\frac{N}{n_x}) \quad (1)$$

In the TFIDF weight strategy $w_{x,i}$ is the weight of entity $e_x$ in a document $d_i$ in the textual collection, $tf_{x,i}$ is the normalized frequency of entity $e_x$ in document $d_i$, $idf_x$ is the inverse document frequency of entity $e_x$, $f_{x,i}$ is the frequency of entity $e_x$ in document $d_i$, $max_{j,i}$ is the number of times the most frequent entity $e_j$ occurs in document $d_i$, $N$ is the number of documents in the textual collection, and $n_x$ is the number of documents in the textual collection in which entity $e_x$ occurs.

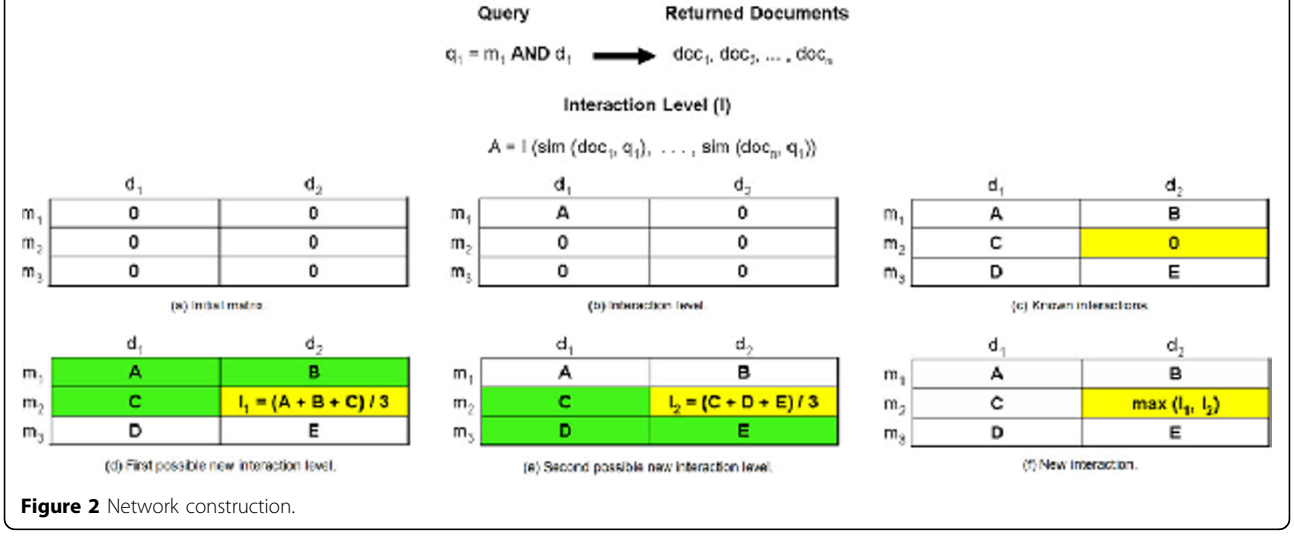


**Figure 2** Network construction.

Each query of the matrix receives a similarity value for each document in the textual collection based on the VSM (Equation 2). For the VSM similarity, $q_j$ is a query $j$ of the matrix representing a conjunction of entities, $t$ is the number of biological entities of the network, $w_{x,i}$ is the weight of entity $e_x$ in document $d_i$, $w_{x,j}$ is the weight of entity $e_x$ in query $q_j$. In our model, the entity weights in a query are always 1 ($w_{x,j}$ = 1). The similarity value stated in the VSM indicates the relevance of a document for a query.

$$sim(d_i, q_j) = \frac{\sum_{x=1}^{t} (w_{x,i} \times w_{x,j})}{\sqrt{\sum_{x=1}^{t} (w_{x,i})^2} \times \sqrt{\sum_{x=1}^{t} (w_{x,j})^2}} \quad (2)$$

We use the similarities returned by equation 2 to determine the interaction level of the query entities. The cell linking the query entities receives this interaction level which represents a known interaction in the subnetwork (Figure 2 (b)). In our current experiments we determine the interaction level of a known interaction in 3 different ways: (i) the arithmetic average of the similarities, (ii) the maximum similarity found, and (iii) the sum of all the similarities.

After all searches in the textual collection are concluded, we have established all known interactions of the network. However, some cells remain equal to 0 indicating that some entity interactions are not explicitly mentioned in the collection (Figure 2 (c)). These cells with value 0 represent the potential new interactions between the biological entities they relate.

The model infers a new interaction in the matrix whenever it finds three interactions satisfying the transitive closure (Figures 2 (d) and 2 (e)). In our current experiments the interaction level of a new interaction is the arithmetic average of the interaction level of the three interactions satisfying the transitive closure. If many interactions satisfy the transitive closure, the model chooses the one with highest arithmetic average (Figure 2 (f)).

We have applied several iterations of our model on the matrix of a subnetwork in order to infer new interactions from interactions previously inferred. In iteration 0 the model discovers all known interactions reported in the textual collection. In iteration 1 the model discovers new interactions based on the known interactions. In iteration 2 the model discovers new interactions based on interactions discovered in iterations 0 and 1. The model stops iterating when all cells of the matrix receives a value different from 0 or when it is no more possible to find interactions satisfying the transitive closure. Starting from iteration 1, our model divides the interaction level of new interactions by the number of iterations performed. This penalty ensures that interactions found in earlier iterations have higher interaction levels.

## Methods

In our experiments, we have considered a sample of patent claims crawled from the USPTO Web site constituting a textual collection with 17,830 documents. All these patents were issued between 01/01/1976 and 12/31/2005. Besides, in the claim section of all these patents we are able to find at least one entity of the four biological categories considered in our crawling process, namely diseases, drugs, genes, and targets. In the USPTO Web site the query we have used to retrieve these patents is represented as *aclm/"entity" and isd/1/1/1976 → 31/12/2005* where *aclm* specifies the patent claim section, *entity* is the biological entity name, and the *isd* specifies the patent issue date, respectively. The entity names are quoted in order to specify the phrase search mode.

As mentioned, we have considered entities of 4 biological system categories: diseases, drugs, genes, and targets (Table 1). We have chosen these categories based on their importance for life sciences research and the practical applications of their entity interactions for the society. The category disease corresponds to a set of possible states of a biological system (e. g. breast cancer, type 2 diabetes, and atherosclerosis). The category drug corresponds to a set of molecules capable of changing the state of a biological system (e. g. aspirin, diclofenac, and tamoxifen). The categories gene and target correspond to a set of building blocks of the biological system. The category gene is a set of building blocks responsible for generating other building blocks (e. g. *major histocompatibility complex class I*, and *tumor suppressor p53*). The category target is a set of building blocks generated by genes and over which a drug acts (e. g. cachectin, and progesterone receptor).

In order to detect the entity occurrences throughout the collection, we have used exact string matching over the entity names and we have also considered entity related names such as synonyms. For instance, we have considered diabetes mellitus type 2 and type 2 diabetes as the same biological entity of category *disease*. We have formed clusters of related names for each entity (Appendix B [additional file]). A representative single name in each cluster is used to represent the whole cluster during the network construction. Some syntactic variation in entity names are also considered in each cluster (e. g. Alzheimer's disease and Alzheimer disease).

In our experiments all categories forming a subnetwork are disjoint sets. For instance, the categories gene and target do not have entities in common when forming the subnetwork *gene* × *target*. Combining these 4

**Table 1 Categories and web sources of the biological entities**

| Category | Number of Entities | Number of Clusters | Web Source |
|---|---|---|---|
| Disease | 52 | 22 | Karolinska Institute [45], Mayo Clinic [46], Therapeutic Target Database [47], Drug Bank [48], Medline Plus [49] |
| Drug | 44 | 22 | Drugs.com [50], Patient.uk [51], Therapeutic Target Database, Drug Bank |
| Gene | 43 | 20 | Kyoto Encyclopedia of Genes and Genomes [52], HUGO Gene Nomenclature Committee [53], NCBI Entrez Gene [54] |
| Target | 50 | 23 | The Free Dictionary [55], Therapeutic Target Database, Drug Bank |
| Total | 189 | 87 | |

biological categories, we have a network composed by 11 subnetworks (Table 2). Of these, 6 have 2 dimensions, 4 have 3 dimensions, and 1 has 4 dimensions.

In the current implementation of our model we neither use natural language processing (NLP) [5,15] nor heuristics to capture the context in which the entity names are applied in the documents. Notwithstanding, the entity names we have selected were satisfactory for our purpose of validating the model, as we shall see.

## Results

### Network construction

In our experiments the biological network has 266,528 possible interactions. Searching the patent claim collection our model has identified 1,027 known interactions (Table 3). Based on these known interactions our model was able to infer 3,195 new interactions.

We have ranked the subnetworks according to their best new interactions (Table 4). In most cases, subnetworks with few dimensions had the higher interaction levels. This happens because in a subnetwork with many dimensions it is more difficult to find documents in which entities of all dimensions co-occur. However, we find more accurate results in subnetworks with more dimensions because the model is able to better constrain the research space when we increase the number of dimensions of a dimensional space.

**Table 2 The subnetworks**

| Subnetwork | Dimensional Space |
|---|---|
| 1 | *disease* × *drug* |
| 2 | *disease* × *gene* |
| 3 | *disease* × *target* |
| 4 | *drug* × *gene* |
| 5 | *drug* × *target* |
| 6 | *gene* × *target* |
| 7 | *disease* × *drug* × *gene* |
| 8 | *disease* × *drug* × *target* |
| 9 | *disease* × *gene* × *target* |
| 10 | *drug* × *gene* × *target* |
| 11 | *disease* × *drug* × *gene* × *target* |

### Validation

Removing patent claims from our collection according to the years in which they were issued and applying our model after each remotion, we observed that new interactions found in a year were confirmed by patent claims removed from the collection and issued in more recent years (Figure 3). For example, in order to better assess the quality of our model, we have analyzed known interactions established in the network in 2005 that became new interactions in 2004 when the patent claims issued in 2005 were removed from our textual collection (Table 5). These known interactions in 2005 represent patents filed in 2005 that our model would have identified in 2004. Thus, we used these known interactions in 2005 as confirmation patent claims for new interactions inferred in 2004. For instance, the interaction between the disease heart attack and the gene *ppar-gama* has 1 patent claim issued in 2005. When we removed this patent claim from the collection, 61 patent claims indicated this interaction as a new one in 2004.

Removing all patent claims issued in 2005, our model predicted 2,930 new interactions based on patents issued up to 2004. Among these new interactions in 2004, we had 32 confirmation patent claims filed in 2005. We then verified the top 100 new interactions found in 2004 for each subnetwork in order to check whether these 32 confirmation patents were among the highest ranked indications of our method (Figure 4).

We have observed the distributions of confirmation patent claims when the known interactions were determined by the average, maximum and sum strategies applied over the similarities returned by the VSM. In the

**Table 3 The subnetworks and their number of known and new interactions.**

| Subnetwork | Dimensional Space | Known Interaction | New Interaction | Total |
|---|---|---|---|---|
| 1 | *disease* × *drug* | 192 | 270 | 462 |
| 2 | *disease* × *gene* | 76 | 184 | 260 |
| 3 | *disease* × *target* | 138 | 346 | 484 |
| 4 | *drug* × *gene* | 38 | 130 | 168 |
| 5 | *drug* × *target* | 105 | 294 | 399 |
| 6 | *gene* × *target* | 50 | 175 | 225 |
| 7 | *disease* × *drug* × *gene* | 71 | 304 | 375 |
| 8 | *disease* × *drug* × *target* | 199 | 958 | 1.157 |
| 9 | *disease* × *gene* × *target* | 55 | 269 | 324 |
| 10 | *drug* × *gene* × *target* | 34 | 76 | 110 |
| 11 | *disease* × *drug* × *gene* × *target* | 69 | 189 | 258 |
| Total | | 1 027 | 3 195 | 4 222 |

subnetwork *disease* × *drug*, for instance, we had 275 new interactions in 2004 (Table 6). This subnetwork had 5 new interactions with confirmation patent claims issued in 2005. When we used the arithmetic average strategy for known interactions we had 3 new confirmed interactions at the top 100 new interactions of this subnetwork ranking. On the other hand, with the maximum and sum strategies we found 4 new confirmed interactions at the top 100 new interactions of this subnetwork. Further, the first confirmed new interaction in this subnetwork is among the top 10 interactions of the ranking and the second one is among the top 20 when we used the arithmetic average strategy. In sum, when we applied the average, maximum and sum strategies, we found 53%, 56%, and 69% of the 32 confirmation patents at the top 100 new interactions of all subnetworks, respectively.

**Table 4 Ranking of subnetworks based on their best new interactions and number of dimensions. The interaction level of known interactions was determined by the arithmetic average of all similarities returned by the vector space model.**

| Subnetwork | Dimensional Space | New Interaction | Level of Interaction |
|---|---|---|---|
| 6 | gene<br>target | androgen receptor<br>adrenaline | 0.9757 |
| 2 | disease<br>gene | HIV<br>transforming growth factor, beta 1 | 0.9738 |
| 5 | drug<br>target | verapamil<br>cyclooxygenase 2 | 0.9597 |
| 1 | disease<br>drug | erectile dysfunction<br>divalproex | 0.9470 |
| 3 | disease<br>target | arrhythmia<br>cyclic-gmp phosphodiesterase | 0.9272 |
| 4 | drug<br>gene | ciclosporin<br>androgen receptor | 0.8211 |
| 8 | disease<br>drug<br>target | alzheimer dementia<br>acetylsalicylic acid<br>adrenaline | 0.8807 |
| 10 | drug<br>gene<br>target | acarbose<br>apolipoprotein a 1<br>lymphotoxin | 0.8723 |
| 9 | disease<br>gene<br>target | parkinson disease<br>apolipoprotein e<br>choline acetylase | 0.8695 |
| 7 | disease<br>drug<br>gene | gout<br>hydrochlorothiazide<br>endothelin 1 | 0.8357 |
| 11 | disease<br>drug<br>gene<br>target | breast adenocarcinoma<br>tamoxifen<br>ppar-gamma<br>hmg-coa reductase | 0.7826 |

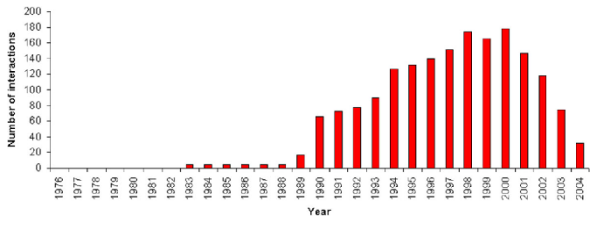


**Figure 3** Interactions with confirmation patent claims by year.

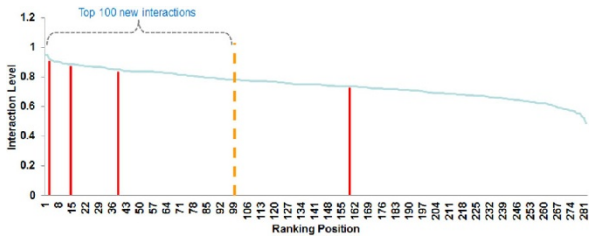


**Figure 4 Distribution of confirmation patent claims filed in 2005 throughout the levels of the ranking constructed with patents issued up to 2004 for the subnetwork *drug* × *target*.** The number of new interactions predicted in this subnetwork in 2004 is 282. We have found 4 confirmation patent claims filed in 2005 for the new interactions predicted in 2004. The position of these 4 confirmation patent claims in the ranking of new interactions predicted in 2004 are respectively 3, 14, 39, and 159. Thus, we have observed that 3 confirmation patent claims were among the top 100 best ranked indications of subnetwork *drug* × *target*.

In addition, we have looked for papers on the Web in order to confirm some of the new interactions found in 2005. For instance, the best result found in our model relates the interaction between the *androgen receptor* gene and the adrenaline neurotransmitter in the 2-dimensional subnetwork *gene* × *target*. Sastry et al. [9] reported that the antiapoptotic effect of epinephrine partially depends on androgen receptor. A modest decrease in the antiapoptotic effect of epinephrine in cells where *androgen receptor* expression was reduced provides evidence that epinephrine reduces sensitivity of cancer cells to apoptosis.

We have found confirmation papers on the Web for the first new interaction of five 2-dimensional subnetworks (Table 7). Out of six 2-dimensional subnetworks, four have had their most relevant new interaction confirmed by later papers issued from 2007 up to 2009. In just one case we have not found a confirmation paper for the first new interaction, in the ranking of the 2-dimensional subnetwork *drug* × *target*. We have not found any confirmation paper for the first new interaction in the ranking of the 3-dimensional subnetworks neither for the 4-dimensional subnetwork. We have not looked for papers on the Web for new interactions in other positions of the rankings in each subnetwork.

## Example session

### *Research space of new interactions*

In our model, subnetworks with more dimensions constrain better the search space for new interactions, thus promoting more accurate results. For instance, consider a researcher using our system who is interested in new interactions related to the drug aspirin. Initially, the researcher decides to analyze the interactions of aspirin with HMG-CoA reductase, cachectin and acetylcholinesterase targets (Figure 5).

Our system shows that the best option would be to conduct research about the interaction between aspirin and acetylcholinesterase, since this interaction has a high interaction level ($l_2$ = 0.9514 where $l_n$ is the interaction level in the $n$-dimensional subnetwork, $n$ = 2, 3, 4...) and the other two are known interactions. Therefore, our model predicts the interaction between aspirin

**Table 5 The top 5 known interactions with high interaction level in 2005 that became new interactions in 2004. The interaction level of known interactions was determined by the arithmetic average of all similarities returned by the vector space model.**

| Subnetwork | Dimensional Space | Interaction | Level in 2005 | Level in 2004 | Patents in 2005 | Patents in 2004 |
|---|---|---|---|---|---|---|
| 5 | disease<br>gene | heart attack<br>ppar-gamma | 0.9999 | 0.8324 | 1 | 61 |
| 1 | target<br>disease | adrenaline<br>cardiac ischemia | 0.9866 | 0.8676 | 1 | 36 |
| 11 | target<br>disease<br>drug<br>gene | hmg coa reduct.<br>breast cancer<br>tamoxifen<br>kennedy disease | 0.9190 | 0.6383 | 1 | 5 |
| 2 | target<br>drug | gp iib/iiia<br>neoral | 0.9137 | 0.8354 | 1 | 103 |
| 4 | disease<br>drug | HIV<br>bonyl | 0.9041 | 0.8825 | 1 | 30 |

**Table 6 The subnetworks and their number of confirmation patent claims at the top 100 new interactions predicted in 2004.**

| Subnetwork | Dimensional Space | New Interactions in 2004 | Confirmations Issued in 2005 | Distribution at the Top 100 New Interactions | | |
|---|---|---|---|---|---|---|
| | | | | AVG | MAX | SUM |
| 1 | *disease* × *drug* | 275 | 5 | 3 | 4 | 4 |
| 2 | *disease* × *gene* | 167 | 2 | 2 | 2 | 2 |
| 3 | *disease* × *target* | 348 | 2 | 1 | 1 | 1 |
| 4 | *drug* × *gene* | 119 | 0 | 0 | 0 | 0 |
| 5 | *drug* × *target* | 282 | 4 | 3 | 3 | 3 |
| 6 | *gene* × *target* | 152 | 3 | 2 | 2 | 3 |
| 7 | *disease* × *drug* × *gene* | 308 | 4 | 1 | 1 | 4 |
| 8 | *disease* × *drug* × *target* | 786 | 9 | 2 | 2 | 3 |
| 9 | *disease* × *gene* × *target* | 242 | 2 | 2 | 2 | 2 |
| 10 | *drug* × *gene* × *target* | 76 | 0 | 0 | 0 | 0 |
| 11 | *disease* × *drug* × *gene* × *target* | 175 | 1 | 1 | 1 | 0 |
| Total | | 2 930 | 32 | 17 | 18 | 22 |
| % | | | | 53 | 56 | 69 |

and acetylcholinesterase as a very promising research topic. However, the researcher can still reach more precise results since the search space is still very large and these entities can interact with several other entities of distinct categories. In other words, the researcher may obtain even more accurate results when considering subnetworks with more dimensions.

Using a 3-dimensional subnetwork, the researcher now considers the dimension gene in the analysis. Then, the researcher discovers that the interaction between aspirin and the acetylcholinesterase becomes less promising because no interaction between these entities is established in the 3-dimensional subnetwork. The researcher realizes that in the 3-dimensional subnetwork

**Table 7 Confirmation papers for the first new interaction predicted in 2005 for each subnetwork.**

| Subnetwork | Dimensional Space | First Interaction | Confirmation Papers |
|---|---|---|---|
| 1 | *disease* × *drug* | impotence<br>divalproex | [56,57] |
| 2 | *disease* × *gene* | acquired immunodeficiency syndrome<br>transforming growth factor, beta 1 | [58,59] |
| 3 | *disease* × *target* | arrhythmia<br>cyclic-gmp phosphodiesterase | [60] |
| 4 | *drug* × *gene* | ciclosporin<br>androgen receptor | [61,62] |
| 5 | *drug* × *target* | verapamil<br>cyclooxygenase 2 | none |
| 6 | *gene* × *target* | androgen receptor<br>adrenaline | [9] |
| 7 | *disease* × *drug* × *gene* | gout<br>hydrochlorothiazide<br>endothelin 1 | none |
| 8 | *disease* × *drug* × *target* | alzheimer's disease<br>aspirin<br>adrenaline | none |
| 9 | *disease* × *gene* × *target* | parkinson's disease<br>apolipoprotein e<br>choline acetylase | none |
| 10 | *drug* × *gene* × *target* | acarbose<br>apolipoprotein a-1<br>lymphotoxin | none |
| 11 | *disease* × *drug* × *gene* × *target* | breast cancer<br>tamoxifen<br>ppar-gamma<br>hmg-coa reductase | none |

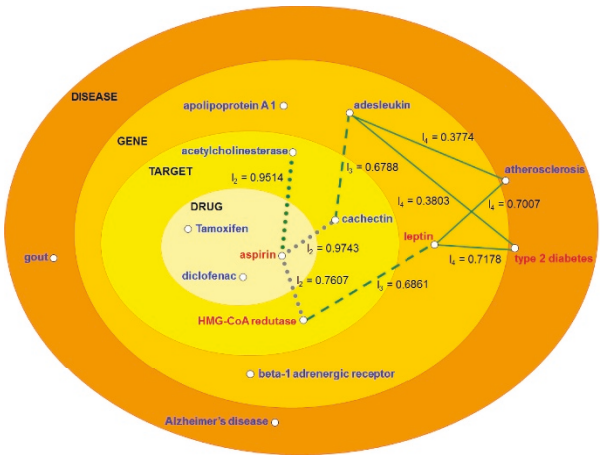


**Figure 5 Research space for some possible interactions with the drug aspirin**. Gray lines are known interactions and green lines are new interactions. The interaction level of known interactions was determined by the arithmetic average of all similarities returned by the vector space model.

*drug* × *gene* × *target* the interaction between aspirin, HMG-CoA reductase and *leptin* with interaction level $l_3$ = 0.6861 becomes the most promising research topic. Finally, going a step further by searching the 4-dimensional subnetwork, the researcher discovers that the interaction among aspirin, HMG-CoA reductase, *leptin* and type 2 diabetes with interaction level $l_4$ = 0.7178 is in fact the most promising interaction for research.

***Interaction history***

The history of how each new interaction may have been established in the network can be followed with the *BioSearch* system. As an example, we observe the history of the new interaction with highest interaction level in the network when we used the arithmetic average strategy to determine the known interaction values. Our model inferred this interaction in 3 steps on the matrix of subnetwork *gene* × *target* (Figure 6).

In the first step the model identifies the possible new interaction between the *androgen receptor* gene and the adrenaline target (Figure 6 (a)). In the second step, the model finds three known interactions in the transitive closure. These known interactions produce an interaction level for the new interaction with value 0.8761 (Figure 6 (b)).

In the third step, the model finds other three known interactions in the transitive closure. In this case, the known interactions produce a new interaction with value 0.9757 (Figure 6 (c)). No more possibilities are found for this new interaction. Thus, the interaction level found at the third step becomes the interaction level of the new interaction because it is higher than that found in the second step.

## Discussion

We have been able to achieve significant results in a strategy that combines the VSM with an inference process in order to predict new biological entity activities. We have used this strategy to model biological systems and to construct a network of biological entity interactions. Modeling biological systems is a complex task for many reasons. For example, we must consider a large number of biological parameters, we must identify entity concentrations and roles in different reactions, and we must bear in mind that biological systems are not linear systems and perturbations commonly give rise to unexpected results. Thus, we have abstracted details and studied biological systems in a higher level in order to decrease their complexities and conduct our analysis [39]. Our abstraction of biological systems is constructed from textual collections that represent a particular view of the technological advances in life sciences reported in patent claims.

In our model, we have focused on retrieving biological entity information from a textual collection consisting of patent claims using the VSM and expressing this information in a transitive closure. This approach has allowed several analysis with important findings. The approach has indicated the VSM as a useful tool to retrieve relevant information in an inference process and how the biological knowledge is interconnected in patent claims.

Texts in patents have a particular writing style characterized by a rich technical terminology and an intentional vagueness in order to promote wide protection to inventions [8,33,34,36,38]. This intentional vagueness may bring a potential benefit to our inference strategy of new interactions. The vagueness in patent texts can indicate some known interactions not easily observed in other literatures characterized by a more strict writing style, as scientific papers. From these known interactions we can infer new ones that are even more innovative than those predicted from texts with strict writing style.

As observed in some contemporary search engines, term co-occurrences is a good way to restrict the documents which can better satisfy an user information need. Thus, term co-occurrences is a good strategy to isolate good hits from a big mass of documents. In addition, previous work in the field of literature-based discovery have indicated many important findings relying on term co-occurrences [12,13,16]. Particularly, Jenssen et al. [10] have shown that co-occurrences reflect biologically meaningful relationships, thus providing an approach to extract and structure known biological knowledge. Accordingly, in our work we have relied on term co-occurrences in order to produce relevant results from a textual collection.

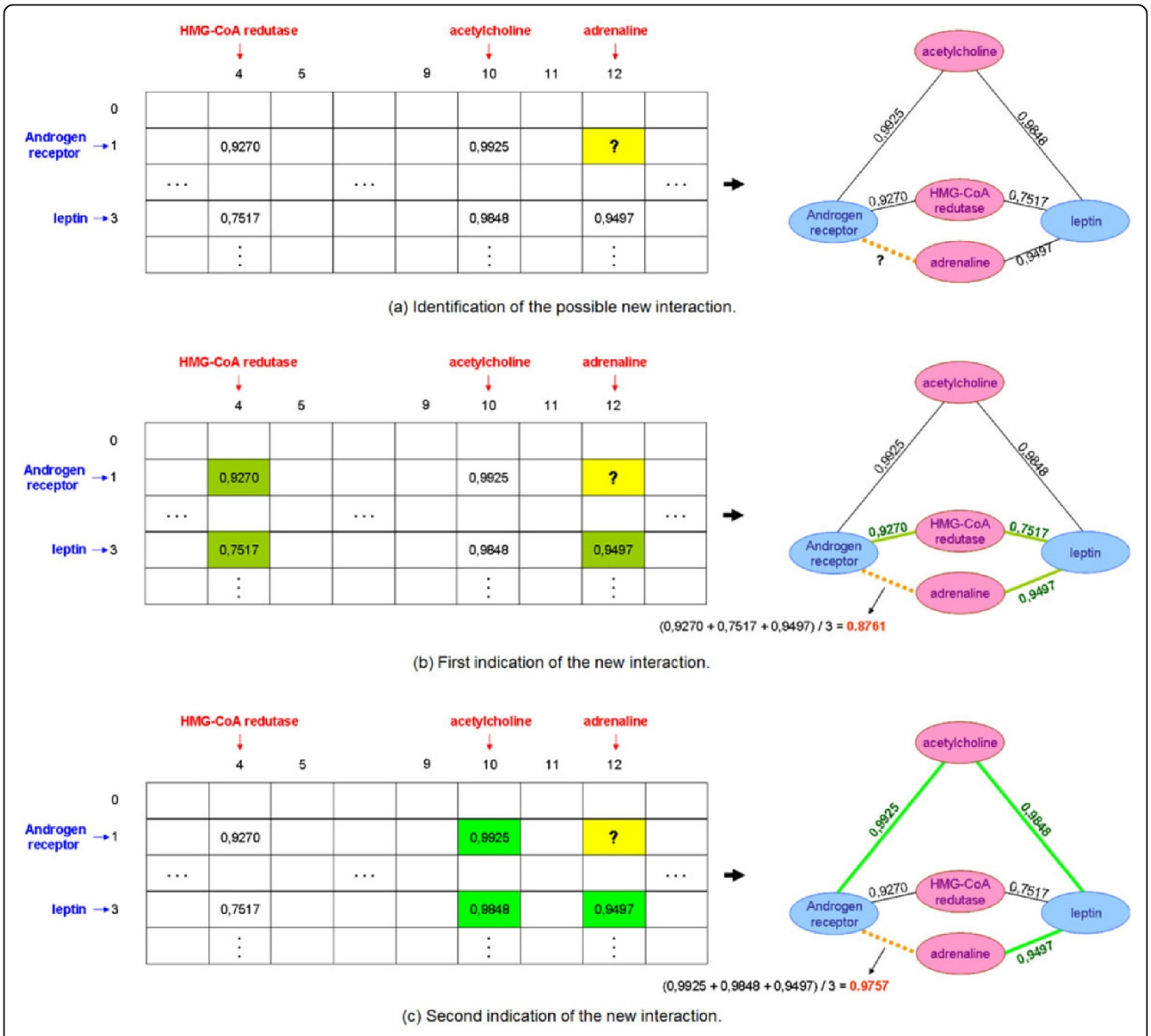


**Figure 6 History of the inference process**. Entities in blue are genes and entities in red are targets. The interaction level of known interactions was determined by the arithmetic average of all similarities returned by the vector space model.

Our model has predicted 2,930 new interactions considering patent claims issued up to 2004. In 2005, we have 32 patent claims in which these new interactions are mentioned. These 32 patent claims issued in 2005 serve as confirmations for the new interactions predicted in 2004. We have also observed that using the VSM we have ranked up to 69% of these 32 new interactions among the 100 first new interactions of all subnetworks. In other words, 69% of the confirmed interactions would have been identified within the top 100 new indications of all subnetworks. These 32 confirmation patent claims also demonstrate that implicit interactions not easily observed in a textual collection must be recognized as important contributions in the field of literature-based discovery.

We consider the 32 new interactions with confirmations as a significant number mainly when considering the reduction in the number of biotechnological patents filed from 2001 to 2004 [40], and the fact that patents are not filed for the majority of scientific discoveries, being instead published as research papers [41-43]. In fact, we already expected a small number of confirmation patent claims for two main reasons. First due to the fact that these are patents filed in 2005, only one year after the new interactions were indicated by the collection issued up to 2004. A higher number of patents may

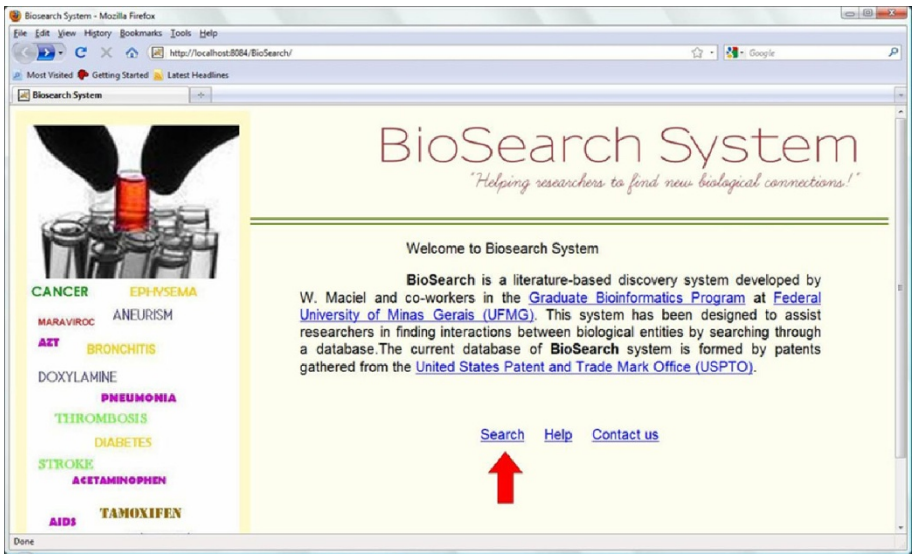


**Figure 7** Biosearch system home page.

have been filed in later years which would provide more indications. Second, filing a patent is the final step of a long sequence of activities related to scientific discovery, and most researchers stop in the scientific article publication phase. As such, a large number of confirmation patent claims should probably never been expected.

Our findings have encouraged us to further investigate biological parameters we have to use in order to improve our representation of biological systems and achieve better results in the inference process and ranking strategy. These parameters have another important function in preventing noise propagation. Interactions poorly established in the network propagate spurious interactions in the inference process. Thus, this study should help impose constraints to the identification of interactions during network construction. The definition of these parameters for several sources and their integration in our model is also an important concern.

In literature-based discovery, simple ranking strategies that promote new interactions based on the raw frequency of known interactions found in textual collections are often used. However, they show, in some situations, implicit interactions that have already been studied but were not documented because they are not


**Figure 8** Biosearch system search page.


**Figure 9** Biosearch system ranking page.

feasible or are unwanted in practice. Our results demonstrate, on the other hand, that ranking strategies based on the VSM are good tools for the identification of significant implicit interactions occurring in texts, mainly those occurring in few documents of a textual collection. This is an important contribution because it is far more difficult to find relevant new interactions from knowledge not frequently co-occurring in a literature than that often observed. However, we must always keep in mind that relevance is a subjective concept. Therefore, biological entity interactions may be considered differently, i.e., with different importance, by different researchers. Then, we must consider strategies in literature-based discovery as complementary tools that help to identify the best new interactions based on the researcher's interests. In this sense, we should even think of systems in which new ranking strategies may be integrated as add-ons.

We should also emphasize that our goal is not to ensure a complete coverage of the biological literature, creating an enormous network of known interactions. Instead, we focus in providing a proof of concept to show the VSM applicability to disclose and rank biological entity activities based on implicit connections found in biological literature. Accordingly, we have checked the existence of these implicit connections in patent claims using a small and restricted textual collection just for assessing the model. We are aware that many new interactions inferred by our model have already been reported in scientific papers. However, we have observed that these findings had not received patent protection at the USPTO until 2005 and we have used some of these scientific papers as validation of our results, mainly due to the inexistence of textual collections currently available for validating literature-based discovery systems [21,22,44]. For a production system we should index as much as possible of the current biological literature sources in order to filter prior art. Nevertheless, we have observed that our strategy provides a good tool for tracking scientific advances published in scientific papers but not yet protected under the intellectual property law.

## Conclusions

In this work we have introduced a technique that employs the Vector Space Model (VSM) for the identification of biological entity activities based on a network of biological entity interactions extracted from textual collections. The algebraic framework of the VSM has demonstrated to be a helpful tool in the task of finding known biological entity activities. We have extended the VSM with a transitive closure approach in order to predict new potential biological entity activities. The transitive closure we have used explores the primary and secondary activities of entities in a biological system. In addition, we have imposed a constraint in this transitive closure in order to ensure that interactions established in the network connect entities of distinct categories. This constraint reduces the search space for new interactions, promoting more accurate results. Moreover, we have used the similarity values derived from the VSM to rank the new discovered entity activities.

Our experiments using a collection of USPTO patent claims demonstrate that the biotechnological patent literature has implicit connections that can be explored to

provide further advances in life sciences. Iterating our model according to the years in which the patent claims were issued, new interactions found in a year were confirmed by patent claims not in the collection and issued in more recent years. The experiments also showed that many confirmation patent claims were found for interactions at the top of our ranks of results. For instance, considering the ranking strategy based on the sum of the similarities returned by the VSM we had 69% of the confirmation patents among the first 100 new interactions of all subnetworks. We have also found scientific papers that validate several of the suggested interactions.

For future work we intend to construct networks using other patent fields (e.g. title, abstract and description sections), the whole patent text, and other sources, such as paper abstracts, paper titles and drug labels. We will analyze the contribution of all these pieces of evidence in our inference process when they are considered separately and together. In addition, we intend to explore natural language processing techniques and ontologies in order to improve the identification of entity co-occurrences in the textual collection. Moreover, we also want to conduct our analyses by considering entities co-occurring in one sentence, in a window of sentences, and in a whole paragraph in order to evaluate a phrase-based VSM approach in the context of our model. Then, we will apply proximity criteria for these occurrences in order to ensure the semantic interaction between entities. Furthermore, we will evaluate a set of biological parameters extracted from the literature in order to help with the establishment of interactions in the networks. Finally, we intend to study other possible strategies to rank biological interactions and conduct a trend analysis on how the interaction value evolves when restricting the number of documents in the textual collection.

## Additional material

**Additional file 1: Appendix.** A concise explanation of the Biosearch system interface and the clusters of entity names we have used in our current experiments.

#### Acknowledgements

WDM acknowledges support from CAPES by the grant of the special program BIOMICRO. All authors acknowledge support from FAPEMIG by the grant of publication fees.

This article has been published as part of *BMC Genomics* Volume 12 Supplement 4, 2011: Proceedings of the 6th International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-meeting 2010). The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2164/12?issue=S4

#### Author details

[1]Bioinformatics PhD Program of the Universidade Federal de Minas Gerais, Belo Horizonte, 31270-901, Brazil. [2]Computer Science Department of the Universidade Federal dos Vales do Jequitinhonha e Mucuri, Diamantina, 39100-000, Brazil. [3]Computer Science Department of the Universidade Federal de Minas Gerais, Belo Horizonte, 31270-901, Brazil.

#### Authors' contributions

WDM, ACFC and SVAC conceived the project. ACFC and SVAC directed the project. WDM conceived and designed the model; implemented and tested the algorithms; prepared and tested the data. WDM and MAG conceived, designed and performed the computational experiment. ACFC and SVAC provided support throughout the research process. ACFC and SVAC hosted the BioSearch system in their lab at UFMG. All authors analyzed the data and results; and wrote, read and approved the final paper.

#### Competing interests

The authors declare that they have no competing interests.

Published: 22 December 2011

#### References

1. Silverman RB: **Drug Discovery, Design, and Development.** *The Organic Chemistry of Drug Design and Drug Action.* second edition. Elsevier Academic Press; 2004, 7-120.
2. Salton G, McGill MJ: **Introduction to Modern Information Retrieval.** New York: McGraw-Hill Book Co; 1986.
3. Baeza-Yates RA, Ribeiro-Neto BA: **Modern Information Retrieval.** New York: ACM Press / Addison-Wesley; 1999.
4. Witten IH, Moffat A, Bell TC: **Managing Gigabytes: Compressing and Indexing Documents and Images.** Morgan Kaufmann Publishing;, second 1999.
5. Hristovski D, Friedman C, Rindflesch TC, Peterlin B: **Exploiting Semantic Relations for Literature-Based Discovery.** *American Medical Informatics Association Symposium Proceedings* Washington DC, United States of America; 2006, 349-353.
6. Maciel WD, Faria-Campos AC, Gonçalves MA, Campos SVA: **The BioSearch System.** 2009 [http://luar.dcc.ufmg.br/BioSearch].
7. USPTO: **United States Patent and Trademark Office Home Page.** 2009 [http://www.uspto.gov/].
8. Shinmori A, Okumura M, Marukawa Y, Iwayama M: **Patent Claim Processing for Readability: Structure Analysis and Term Explanation.** *Proceedings of the Workshop on Patent Corpus Processing* Sapporo, Japan; 2003, 56-65.
9. Sastry KSR, Karpova Y, Prokopovich S, Smith AJ, Essau B, Gersappe A, Carson JP, Weber MJ, Register TC, Chen YQ, Penn RB, Kulik G: **Epinephrine Protects Cancer Cells from Apoptosis via Activation of cAMP-dependent Protein Kinase and BAD Phosphorylation.** *Journal of Biological Chemistry* 2007, **282(19)**:14094-14100.
10. Jenssen TK, K J, H E, Laegreid A: **A Literature Network of Human Genes for High-Throughput Analysis of Gene Expression.** *Nature Genetics* 2001, **28**:21-28.
11. Bruza P, Weeber M: **Literature-Based Discovery.** Springer; 2008.
12. Swanson DR: **Fish-oil, Raynaud's Syndrome, and Undiscovered Public Knowledge.** *Perspectives in Biology and Medicine* 1986, **30**:7-18.
13. Swanson DR: **Medical Literature as a Potential Source of New Knowledge.** *Bulletin of the Medical Library Association* 1990, **78**:29-37.
14. Smalheiser NR, Swanson DR: **Using Arrowsmith: a Computer-Assisted Approach to Formulating and Assessing Scientific Hypotheses.** *Computer Methods and Programs in Biomedicine* 1998, **57**:149-153.
15. Weeber M, Klein H, de Jong-van den Berg LTW, Vos R: **Using Concepts in Literature-Based Discovery: Simulating Swanson's Raynaud-Fish Oil and Migraine-Magnesium Discoveries.** *Journal of the American Society for Information Science and Technology* 2001, **52(7)**:548-557.
16. Swanson DR, Smalheiser NR, Torvik VL: **Ranking Indirect Connections in Literature-Based Discovery: The Role of Medical Subject Headings.** *Journal of the American Society for Information Science and Technology* 2006, **57(11)**:1427-1439.
17. Hristovski D, Peterlin B, Mitchell JA, Humphrey SM: **Using Literature-Based Discovery to Identify Disease Candidates Genes.** *International Journal of Medical Informatics* 2005, **74**:289-298.
18. Hristovski D, Stare J, Peterlin B, Dzeroski S: **Supporting Discovery in Medicine by Association Rule Mining in Medline and UMLS.** *Studies in Health Technology and Informatics* 2001, **84**:1344-1348.

19. Hristovski D, Peterlin B, Mitchell JA, Humphrey SM: **Improving Literature Based Discovery Support by Genetic Knowledge Integration.** *Studies in Health Technology and Informatics* 2003, **95**:68-73.
20. Wren JD, Bekeredjian R, Stewart JA, Shohet RV, Garner HR: **Knowledge Discovery by Automated Identification and Ranking of Implicit Relationships.** *Bioinformatics* 2004, **20(3)**:389-398.
21. Kostoff RN: **Literature-Related Discovery (LRD): Introduction and background.** *Technological Forecasting and Social Change* 2008, **75**:165-185.
22. Kostoff RN: **Where is the Discovery in Literature-Based Discovery?** *Literature-Based Discovery* Springer; 2008, 57-72.
23. Kostoff RN, Briggs MB, Solka JL, Rushenberg RL: **Literature-Related Discovery (LRD): Methodology.** *Technological Forecasting and Social Change* 2008, **75**:186-202.
24. Kostoff RN, Block JA, Stump JA, Johnson D: **Literature-Related Discovery (LRD): Potential Treatments for Raynaud's Phenomenon.** *Technological Forecasting and Social Change* 2008, **75**:203-214.
25. Kostoff RN: **Literature-Related Discovery (LRD): Potential Treatments for Cataracts.** *Technological Forecasting and Social Change* 2008, **75**:215-225.
26. Kostoff RN, Briggs MB: **Literature-Related Discovery (LRD): Potential Treatments for Parkinson's Disease.** *Technological Forecasting and Social Change* 2008, **75**:226-238.
27. Kostoff RN, Briggs MB, Lyons TJ: **Literature-Related Discovery (LRD): Potential Treatments for Multiple Sclerosis.** *Technological Forecasting and Social Change* 2008, **75**:239-255.
28. Kostoff RN, Solka JL, Rushenberg RL, Wyatt JA: **Literature-Related Discovery (LRD): Water Purification.** *Technological Forecasting and Social Change* 2008, **75**:256-275.
29. Kostoff RN, Block JA, Solka JL, Briggs MB, Rushenberg RL, Stump JA, Johnson D, Lyons TJ, Wyatt JR: **Literature-Related Discovery (LRD): Lessons Learned, and Future Research Directions.** *Technological Forecasting and Social Change* 2008, **75**:276-299.
30. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet J, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub1 TR: **The Connectivity Map: using gene-expression signatures to connect small molecules, genes and disease.** *Science* 2006, **313**:1929-1935.
31. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L: **Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network.** *Science* 2001, **292**:929-934.
32. Campillos M, Kuhn M, Gavin A, Jensen LJ, Boork P: **Drug Target Identification Using Side-Effect Similarity.** *Science* 2008, **321**:263-266.
33. Lechter MA, Clifford EC, Famiglio RB, Joenk RJ: **Successful Patents and Patenting for Engineers and Scientists.** The Institute of Electrical and Electronics Engineers, Inc., New York; 1990.
34. Tseng Y, Lin C, Lin Y: **Text mining techniques for patent analysis.** *Information Processing and Management: an International Journal* 2007, **43**:1216-1247[http://portal.acm.org/citation.cfm?id=1241109.1241327].
35. Trippe AJ: **Patinformatics: tasks to tools.** *World Patent Information* 2003, **25**:211-221.
36. Mukherjea S, Bamba B: **BioPatentMiner: An Information Retrieval System for BioMedical Patents.** *Proceedings of 30th Very Large Database (VLDB) Conference* Toronto, Ontario, Canada; 2004, 1066-1077.
37. Larkey LS: **A Patent Search and Classification System.** *Proceedings of the Fourth ACM conference on Digital libraries* Berkeley, California, United States; 1999, 179-187.
38. Fall CJ, Törcsvári A, Benzineb K, Karetka G: **Automated Categorization in the International Patent Classification.** *Proceedings of the ACM SIGIR Forum 37 (1)* Toronto, Canada; 2003, 10-25.
39. Holme P: **Model Validation of Simple-Graph Representations of Metabolism.** *J R Soc Interface* 2009, **6**:1027-1034.
40. Horn CEV, Lipsey CE: **Biotechnology Innovation Repost 2004 - Benchmarks.** Finnegan, Henderson, Farabow, Garrett and Dunner, LLP; 2004.
41. Looy BV, Magerman T, Debackere K: **Developing technology in the vicinity of science: An examination of the relationship between science intensity (of patents) and technological productivity within the field of biotechnology.** *Scientometrics* 2007, **70(2)**:441-458.
42. Narin F, Olivastro D: **Linkage between patents and papers: An interim EPO/US comparison.** *Scientometrics* 1998, **41(1-2)**:51-59.
43. Glänzel W, Meyer M: **Patents cited in the scientific literature: An exploratory study of "reverse" citation relations.** *Scientometrics* 2003, **58(2)**:415-428.
44. Smalheiser NR, Torvik VI: **The Place of Literature-Based Discovery in Contemporary Scientific Practice.** *Literature-Based Discovery* Springer; 2008, 13-22.
45. KarolinskaInstitutet: **Karolinska Institutet Alphabetical List of Diseases.** 2009 [http://www.mic.stacken.kth.se/Diseases/Alphalist.html].
46. MayoClinic: **Mayo Clinic Alphabetical List of Diseases and Conditions.** 2009 [http://www.mayoclinic.com/].
47. TTD: **Therapeutic Target Database Home Page.** 2009 [http://xin.cz3.nus.edu.sg/group/cjttd/ttd.asp].
48. DrugBank: **Drug Bank Home Page.** 2009 [Http://www.drugbank.ca/].
49. MedlinePlus: **Health Topics.** 2009 [http://www.nlm.nih.gov/medlineplus/healthtopics.html].
50. Drugscom: **Drug Information and Side Effects Online.** 2009 [http://www.drugs.com/].
51. PatientUK: **Patient UK Home Page.** 2009 [http://www.patient.co.uk/].
52. KEGG: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** 2009 [http://www.genome.jp/kegg/].
53. HGNC: **HUGO Gene Nomenclature Committee.** 2009 [http://www.genenames.org/].
54. NCBI: **NCBI Entrez Gene.** 2009 [http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene].
55. TheFreeDictionary: **The Free Dictionary Home Page.** 2009 [http://www.thefreedictionary.com/].
56. Kaufman KR, Marin H, Menza M: **Trazodone and ejaculatory inhibition.** *Journal of Sex and Marital Therapy* 2007, **33(3)**:225-230.
57. Szupera Z: **The role of the antiepileptic drugs at the development of the sexual dysfunctions in male epileptic patients.** *Ideggyogyaszati Szemle* 2007, **60(1-2)**:4-13.
58. Otis JS, Ashikhmin YI, Brown LA, Guidot DM: **Effect of HIV-1-related protein expression on cardiac and skeletal muscles from transgenic rats.** *AIDS Res Ther* 2008, **5(8)**:1-9.
59. Ruff KR, Puetter A, Levy LS: **Growth regulation of simian and human AIDS-related non-Hodgkin's lymphoma cell lines by TGF-beta1 and IL-6.** *BMC Cancer* 2007, **7(35)**:1-13.
60. Carceles MD, Aleixandre F, Fuente T, López-Vidal J, Laorden ML: **Effects of rolipram, pimobendan and zaprinast on ischaemia-induced dysrhythmias and on ventricular cyclic nucleotide content in the anaesthetized rat.** *European Journal of Anaesthesiology* 2003, **20(3)**:205-211.
61. Periyasamy S, Warrier M, Tillekeratne MP, Shou W, Sanchez ER: **The immunophilin ligands cyclosporin A and FK506 suppress prostate cancer cell growth by androgen receptor-dependent and -independent mechanisms.** *Endocrinology* 2007, **148(10)**:4716-4726.
62. Ranganathan S, Harmison GG, Meyertholen K, Pennuto M, Burnett BG, Fischbeck KH: **Mitochondrial abnormalities in spinal and bulbar muscular atrophy.** *Human Molecular Genetics* 2009, **18**:27-42.

doi:10.1186/1471-2164-12-S4-S1
**Cite this article as:** Maciel *et al.*: Can the vector space model be used to identify biological entity activities? *BMC Genomics* 2011 **12**(Suppl 4):S1.