Proceedings of the 9th International Symposium on
Dynamics and Control of Process Systems (DYCOPS 2010),
Leuven, Belgium, July 5-7, 2010
Mayuresh Kothare, Moses Tade, Alain Vande Wouwer, Ilse Smets (Eds.)

Target identification in biological systems using network connectivity information from literature mining databases

Ugur Guner*, Jay H. Lee. Omar L. Francone***,Dmitriy Leyfer******

**Georgia Institute of Technology, Atlanta, GA 30332
USA (Tel: 404-388-2149; e-mail: Ugur.Guner@chbe.gatech.edu)*

***Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail:Jay.Lee@chbe.gatech.edu)*

****Pfizer Global Research and Development, Groton, CT 06354 USA (e-mail: Omar.L.Francone@pfizer.com)*
*****Pfizer Global Research and Development, Groton, CT 06354 USA (e-mail: Dmitriy.Leyfer@pfizer.com)*

Abstract: We address the automated drug target identification problem for pharmaceutical research. It is often the case in pharmaceutical industry to bring a new promising target to clinical trials only to find that it has serious safety concerns or lack of efficacy. A gene downstream or upstream in the pathway can be a remedy, however, finding such an alternative target using existing in-silico or bench tools can be extremely labor-intensive. Recently, increasing amounts of information and observations have been compiled from different areas of biological research and deposited on databases. In this work we propose a novel computational method to quantify indirect relationships between the objects of biological research of interest by using existing relationships from text mining databases to automate the search for novel biological targets. We applied our method to analyze 9575 proteins in Ariadne database and create a rank-ordered list of proteins that are most similar to the original query. We also compared our method with the Jaccard similarity index for link prediction performance. Our method outperformed the Jaccard method in predicting the existing links for 9575 proteins in the database.

Keywords: Target identification; Text mining databases; Similarity Score; Link Prediction; Biological networks; Network Connectivity; Bipartite Networks, Information Retrieval, In-silico.

1. INTRODUCTION

Biological processes are the result of interactions involving hundreds of thousands of molecular entities. These interactions form complex networks. To understand diseases and find new drug targets in a systematic way, it is essential to understand the topology of these networks. It is often the case in pharmaceutical industry to bring a new promising target to clinical trials only to find that it has serious safety concerns or lack of efficacy. A gene downstream or upstream in the network might be a solution, however, not all pathways are known, and finding such an alternative target using existing in-silico or bench tools can be extremely labor-intensive. A method that can automatically find implicit relationships between network nodes(proteins, diseases, drugs, compounds etc.) can be invaluable in the search of new target. Increasing amount of information is compiled into biological network format. Text mining is the automated way of collecting the relationships between biological entities through co-occurrences within electronically available records (Wren et al, 2004). It aims at collecting and retrieving useful hidden relations from these resources of information. Therefore, text mining databases represent different sets of pre-compiled information on biological relationships and associations, interactions and facts which have been extracted from the biomedical literature.

In this work we propose a novel computational method of drug target discovery by quantifying indirect relationships between the nodes of biological networks using interactions retrieved from mining databases. This method can also be used to annotate diseases with similar etiology, reposition existing drugs, or discover adverse events for the targets.

This paper is organized as follows. In the next section we will briefly summarize our method for quantifying the similarity between biological objects based on their network connectivity. Section 3 will assess the performance of our method compared to a commonly used existing information retrieval method. Finally, section 4 will present the concluding remarks.

2. METHODS

Our model is based on a computational approach that quantifies the relevance of two biological objects such as genes, proteins, compounds, complexes, drugs, diseases (hereafter referred to simply as “objects” or “entities”) by comparing their common connections against a random network model obtained through the databases. Denoting an

object of interest with 'A', one can identify other objects 'B'. (Fig. 1).

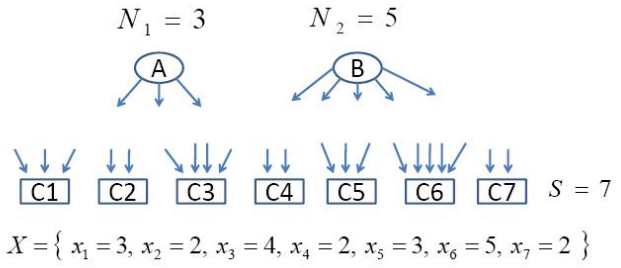

Fig. 1. Random bipartite network model for entities A and B

The text mining database information can be represented as a directed bi-partite network. In graph theory a bipartite graph is graph whose vertices can be divided into two sets. This is a directed graph where the relations between the nodes are represented as arrows with originating from a source node and ending in a sink node. Out-degree of a source node in a directed graph is the number of edges (arrows) originating from the node and in-degree of a sink node is the number of arrows ending in a sink node. In other words an out-degree is the number of distinct objects that a source node (first set object) is effecting and in-degree is the number of distinct objects a sink node (second set object) is being effected by a source node(first set object). Figure 1 is a directed bi-partite graph model where the random network model is sought using the parameters of the network. It consists of two sets of nodes. The first set nodes are the source nodes A and B and the second set nodes are the sink nodes $C_1,...,C_7$. Each node in both sets refers to certain biological object. The parameters are the out-degrees of the pair of the entities A and B and in-degrees of objects in the second set along with the number of entities in this set. Out-degrees of A and B are represented with N_1 and N_2, whereas in-degrees are denoted as $\{x_1, x_2, .., x_S\}$. S is the total number of entities in the second set of the bi-partite graph. Let us denote the parameter set that we obtain from the database with, $\theta = \{N_1, N_2, x_1, x_2, .., x_S, S\}$.

Random graph is a method to model the possible ways for A and B to connect to the objects of the second set. This allows us to quantify the randomness of A and B having common downstream objects. We then compare the observed common downstream connections against this random graph model to quantify the similarity between A, and B.

Let us define the two different events on this bi-partite graph. First event is the number of common entities that A and B are connected and second event is the identity of these common entities. The joint probability of these two events can be represented with following expression;

$$P\left(M = \{ m_1, m_2, ..., m_k \}, i = k \mid \theta \right) \quad (1)$$

,where M is the list of the identities of the common downstream entities, and i is the number of common entities. Using the definition of joint probability distribution, one can write the following equation;

$$\begin{aligned} &P\left(M = \{ m_1, m_2, ..., m_k \}, i = k \mid \theta \right) \\ &\quad = P\left(i = k \mid \theta \right) . P\left(M = \{ m_1, ..., m_k \} \mid i = k, \theta \right) \end{aligned} \quad (2)$$

In this equation $P\ (i = k \mid \theta)$ is the probability of first event given the parameters, and $P\left(M = \{ m_1, ..., m_k \} \mid i = k, \theta \right)$ is the conditional probability of second event conditional on the first event given the parameters.
The first term in this equation, $P\left(i = k \mid \theta \right)$ can be derived as a function of the parameters; N_1, N_2 and S. In figure 2, a random configuration of bipartite graph for A and B is shown. Connections that are common for the pair are represented with solid, whereas node specific connections are displayed by dashed lines.

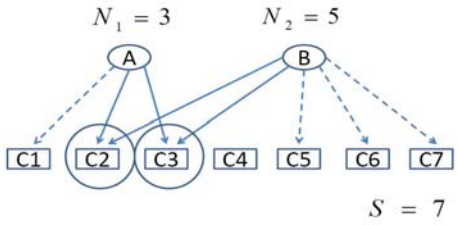

Figure 2. Example of A and B having common second set entities, C_2, C_3.

In order to derive the probabilistic distribution for the number of common objects that A and B share, we start with enumeration of different possibilities. The number of combinations of the N_1 connections that A can make with S different second set objects is calculated by the following;

$$C(S, N_1) = \frac{S!}{N_1!(S - N_1)!} \quad (3)$$

One can obtain similar equation for B;

$$C(S, N_2) = \frac{S!}{N_2!(S - N_2)!} \quad (4)$$

Let L denotes the total number of combinations of A and B connections to any N_1 and N_2 second set objects. L can be calculated as the multiplication of the combinations of both cases;

$$L = C(S, N_1) . C(S, N_2) = \frac{S!}{N_1!(S - N_1)!} \frac{S!}{N_2!(S - N_2)!} \quad (5)$$

The number of combinations for A and B having k common downstream objects (second set objects or sink nodes) can be represented as follows;

$$C(S,k)=\frac{S!}{k!\cdot(S-k)!} \tag{6}$$

Once k connections of A and B are fixed, they have (N_1-k) and (N_2-k) connections remaining respectively. The number of objects available in the second set is reduced to $(S-k)$. The number of ways that the remaining connections of A could be chosen out of $(S-k)$ entities can be calculated as follows;

$$C\left((S-k),(N_1-k)\right)=\frac{(S-k)!}{(N_1-k)!\cdot(S-N_1)!} \tag{7}$$

This will fix the number of all N_1 connections of A and there will be $(S-N_1)$ objects left for (N_2-k) remaining connections of B. The number of combinations for remaining connections of B for the remaining objects is represented as follows;

$$C\left((S-N_1),(N_2-k)\right)=\frac{(S-N_1)!}{(N_2-k)!\cdot(S-N_1-N_2+k)!} \tag{8}$$

The overall number of combinations that A and B are connected to k common objects will be the denoted with D .It can be written as;

$$\begin{aligned} D &= C(S,k)\,C\left((S-k),(N_1-k)\right).C\left((S-N_1),(N_2-k)\right) \\ &= \frac{S!}{k!\cdot(S-k)!}\,\frac{(S-k)!}{(N_1-k)!\cdot(S-N_1)!}\,\frac{(S-N_1)!}{(N_2-k)!\cdot(S-N_1-N_2+k)!} \end{aligned} \tag{9}$$

The probability that A and B are connected to k common objects is the ratio of the total number of combinations of A and B are connected to k objects in common to the total number of combinations that the pair is connected to objects in any possible way. The probability is written as;

$$\begin{aligned} &P(i=k\mid\theta) \\ &=\frac{D}{L}=\frac{C(S,k).C\left((S-k),(N_1-k)\right).C\left((S-N_1),(N_2-k)\right)}{C(S,N_1)C(S,N_2)} \\ &=\frac{\frac{S!}{k!(S-k)!}\,\frac{(S-k)!}{(N_1-k)!\cdot(S-N_1)!}\,\frac{(S-N_1)!}{(N_2-k)!\cdot(S-N_1-N_2+k)!}}{\frac{S!}{N_1!(S-N_1)!}\,\frac{S!}{N_2!(S-N_2)!}} \end{aligned} \tag{10}$$

After cancelations, we obtain;

$$\begin{aligned} &P(i=k\mid\theta) \\ &=\frac{N_1!\cdot N_2!\,.\,(S-N_1)!\,.\,(S-N_2)!}{S!\,.\,k!\,.\,(N_1-k)!\,.\,(N_2-k)!\cdot(S-N_1-N_2+k)!} \end{aligned} \tag{11}$$

This expression can be approximated by a Poisson distribution.

$$P\left(i=k\mid\theta\right)=\frac{1}{\alpha}\frac{\lambda^{i}e^{-\lambda}}{i!}$$

$$\lambda=\frac{N_1N_2}{S} \qquad \alpha=\sum_{i=0}^{\min\left(N_1,N_2\right)}\frac{\lambda^{i}e^{-\lambda}}{i!} \tag{12}$$

,where λ is a function of N_1, N_2 and S while α is the normalization factor. It normalizes the cumulative distribution to one at $i=\min\left(N_1,N_2\right)$ as the probability is not defined beyond this point. This approximation allows us to obtain a compact representation for the probability term. It is less computationally intensive. The aim is to derive a compact similarity score function between two objects that makes sense intuitively starting from a formal probabilistic framework.

To check the validity of the approximation we calculated sum of absolute deviation of the equation (11) from the Poisson approximation for all possible values of $i=\{0,1,...,\min(N_1,N_2)\}$ at different values of N_1 and N_2 . This corresponds to the deviation of cumulative distributions for Poisson and equation (11). We defined the percentage deviation as follows;

$$\begin{aligned} &E(N_1,N_2,S) \\ &=100*\sum_{i=1}^{\min\left(N_1,N_2\right)}\left|\left(\frac{1}{\alpha}\frac{\lambda^{i}e^{-\lambda}}{i!}\right)-\left(\frac{N_1!\cdot N_2!\,.\,(S-N_1)!\,.\,(S-N_2)!}{S!\,.\,i!\,.\,(N_1-i)!\,.\,(N_2-i)!\cdot(S-N_1-N_2+i)!}\right)\right| \end{aligned} \tag{13}$$

In figure (3), we illustrated the calculation the value of E on an example. The absolute deviation of Poisson approximation corresponds to the sum of lengths of the dotted lines.

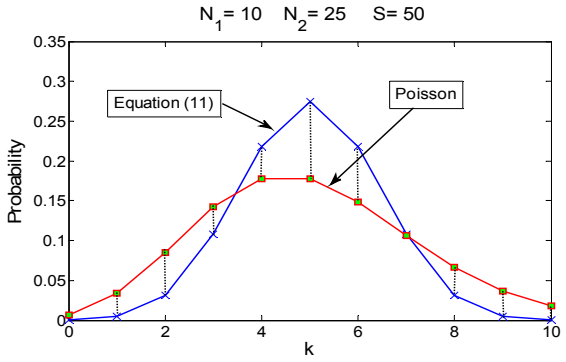

Figure 3. Deviation of Poisson distribution for $N_1=10$, $N_2=25$, $S=50$

In figure (4), the curves for various E values are shown at different values of N_1 and N_2. The area under each curve shows the region for the values of N_1 and N_2 where Poisson approximation exceeds the given percentage deviation. For example the deviation of Poisson approximation is less than 10% when one of the objects has four or more connections ($N_1\geq 4$) and the other object connected to less than 34% of all second layer objects ($N_2\leq 0.34\times S$). N_1 and N_2 can be used interchangeably

and the area under the curves remain same for different values of S . This figure shows that Poisson distribution is a reasonably good approximation for a large span of N_1 and N_2 values.

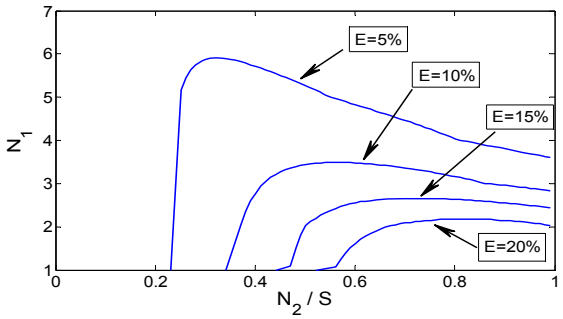

Figure 4. Deviation of Poisson approximation for different N_1 and N_2 values.

One can also derive the conditional probability term on the right hand side of equation (2). In figure (5), a possible connection pattern is shown for illustration purposes. k is the number of shared entities between A and B (in this example there are two common entities) , M shows the list of common entities and X is the set of in-degree values for these commonly shared objects.

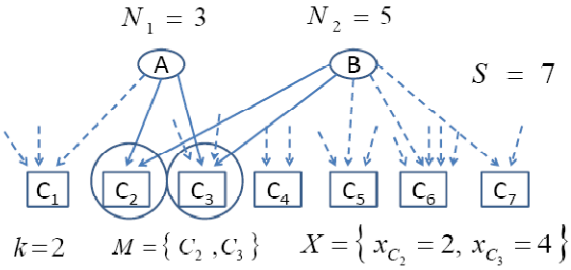

Figure 5. Example of A and B having common second set entities, C_2 , C_3 with their in-degrees.

Let us consider the general case for k common objects. The number of possible ways for A and B to be both connected to a particular second set object (sink node) with a given in-degree of x_i is equal the number of 2-combinations of x_i . In other words, it is the number of combinations that two objects (A and B) can be connected to a particular object that is known to have x_i objects connected to it. It can be calculated as;

$$c_i = C(x_i, 2) = \frac{x_i(x_i - 1)}{2} \tag{14}$$

In this equation the number of combinations where A and B are both connected i^{th} object is denoted by c_i .

The number of possible connections of A and B to any k objects with known in-degrees in the downstream is written as follows;

$$Z = \sum_{i=1}^{S} c_i \sum_{j=1}^{S} c_j \ldots \sum_{z=1}^{S} c_z \tag{15}$$

In this equation there are k embedded summation terms corresponding to k common objects. Each common object can be chosen out of S different objects. For large S , this summation term would be difficult to calculate. Therefore we introduce the following approximation.

$$c_1 = c_2 = \ldots = c_S = \hat{c} \tag{16}$$

Here we assume that all c_i terms are equal to an average $\hat{c}$ term. If (16) is plugged into expression (15), we obtain the following approximation,

$$Z \approx \sum_{i=1}^{S} \hat{c} \sum_{j=1}^{S} \hat{c} \ldots \sum_{z=1}^{S} \hat{c} = S^k \hat{c}^k \tag{17}$$

One can represent the number possible ways that A and B are connected to k particular objects as follows;

$$T = \prod_{i=1}^{k} c_i \tag{18}$$

The probability that A - B pair are connected to k particular objects is calculated as the ratio of the number of combinations that this pair is connected to k particular objects to the number of different ways that they are connected to any k objects.

$$P(M = \{ m_1, m_2, \ldots, m_k \} | , i = k, \theta) = \frac{T}{Z} = \frac{\prod_{i=1}^{k} c_i}{S^k \hat{c}^k} \tag{19}$$

Plugging expression (19) and (12) into expression (2), we obtain

$$P(M = \{ m_1, m_2, \ldots, m_k \}, i = k \mid \theta) = \frac{1}{\alpha} \frac{\lambda^k e^{-\lambda}}{k!} . \frac{\prod_{i=1}^{k} c_i}{S^k \hat{c}^k} \tag{20}$$

This equation gives us the probability of two entities having k common downstream objects from the set M . It is derived based on a random bi-partite network model using the parameter set, θ. The similarity between the pair of entities; A and B is assumed to be based on the statistical significance of their common connections according to the probability of occurrence in a random network model. To quantify the significance of an observed connectivity structure of the pair that has common downstream entities, we defined the following score function;

$$Score = -\log\left(P\left(M = \{ m_1, m_2, ..., m_k \}, i = k \mid \theta \right)\right)$$
$$= -\log\left(\frac{1}{\alpha} \frac{\lambda^k e^{-\lambda}}{k!} . \frac{\prod_{i=1}^{k} c_i}{S^k \hat{c}^k} \right) \quad (21)$$

Hence, the lower the probability of occurrence for a random model is, the more significant the event is and therefore the higher the score. One can write score in an open form as follows;

$$Score = \log(\alpha) + \log(k!) - k\log(\lambda) + \lambda + k\log(S.\hat{c}) - \sum_{i=1}^{k} \log(c_i) \quad (22)$$

One can use Sterling approximation for the term, $\log(k!)$;

$$\log(k!) \approx k\log(k) - k \quad (23)$$

Using (23), expression for λ in (12) and rearranging the terms, expression (22) can be rewritten as follows;

$$Score = \left(\log(\alpha) + \lambda - k\right) + \sum_{i=1}^{k} \log\left(\frac{S}{N_1} \frac{S}{N_2} \frac{\hat{c}}{c_i} k \right) \quad (24)$$

This equation can be further simplified by the following assumption;

$$\left(\log(\alpha) + \lambda - k\right) << \sum_{i=1}^{k} \log\left(\frac{S}{N_1} \frac{S}{N_2} \frac{\hat{c}}{c_i} k \right) \quad (25)$$

Finally, one can obtain the following expression;

$$Score = \sum_{i=1}^{k} \log\left[\frac{S}{N_1} \frac{S}{N_2} \frac{\hat{c}}{c_i} k \right]$$
$$= k.\log\left(S^2 \hat{c}\right) + \sum_{i=1}^{k} \log\left[\frac{k}{c_i N_1 N_2} \right] \quad (26)$$

This function gives us the similarity score between A and B based on the network structure and properties. In this expression, $\log\left(S^2 \hat{c}\right)$ is a network domain-dependent constant. A network domain can be defined as part of the network with all biological interactions of a certain type. Examples of such domains can be transcriptional regulation, protein binding, protein modification and any other biological function that connects one biological entity to another. Each domain might have different number of second layer entities (S) and connectivity structure (c).

One can see that the similarity score is directly proportional to number of common downstream objects, k. This is an expected result as one expects two entities to be similar when they have more common downstream effects. Score is also inversely proportional to both N_1 and N_2. This can be interpreted as the more connected the species are the more likely they have common downstream effect by chance. Finally, the score is inversely proportional to in-degree of the common objects connected to the pair. This is the result of the fact that the pair will more likely to have common downstream entities that have high in-degree by chance. Hence, this commonality gives relatively lower significance for the similarity.

3. ASSESSING THE PERFORMANCE

We applied our algorithm on Ariadne database. Ariadne is a Systems Biology software that consists of computational methods to generate databases from the literature. Ariadne database represent different sets of biological relationships which have been extracted from the biomedical literature (Novichkova et al., 2003). A rank list for each protein among 9575 proteins was created according to our similarity score in expression (26). We decided to evaluate the link prediction performance of our method for this set of proteins. Link prediction in networks is the problem of inferring missing links from an observed network connections. In other words, in a number of domains one constructs a network of interactions based on observable data and then tries to infer additional links that, while not directly visible, are likely to exist. Link prediction offers a very natural basis for evaluation as it allows one to assess the capability of a method to infer meaningful inferences from the observed network data (Nowell and Kleinberg, 2004). Here, we slightly modified the link prediction problem and measured the capability of our method to infer existing links (rather than missing links as they can only be validated through biological experiments) using the observed information from the database. In our approach we derived a similarity score function (eq.(26)) starting from the probabilistic model using only the parameters of the network. Therefore, we don't use any information of existence of any particular link in the network. Our similarity score functions quantifies the similarity between two objects and we assume that higher similarity between a pair of objects can imply existence of a network link between them as similar objects tend to regulate each other or take parts in same processes.

We compare our method with the Jaccard similarity index that has been a commonly used metric in information retrieval (Salton and McGill,1983). Jaccard score can be described with the following equation;

$$Score\left(A, B\right) = \frac{\left|\Gamma(A) \cap \Gamma(B)\right|}{\left|\Gamma(A) \cup \Gamma(B)\right|} \quad (27)$$

,where $\Gamma(A), \Gamma(B)$ are the sets of entities that objects A and B are connected respectively and $|.|$ is the cardinality of the set . For the network example in figure 5, these sets can be written as;

$$\Gamma(A) = \{C_1, C_2, C_3\},\ \Gamma(A) = \{C_2, C_3, C_5, C_6, C_7\} \quad (28)$$

We listed the proteins that are connected to each protein from the database as in figure 6. In this figure, P_i represents i^{th} protein and $P_{i,J}$ is the j^{th} protein connected to it. M_i is the total number of proteins connected. The following expression represents the list for each protein.

$$L(P_i) = \{P_{i,1}, ..., P_{i,Mi}\} \quad (29)$$

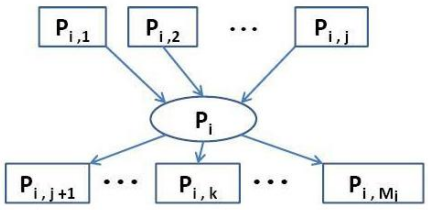

Figure 6. Representation of a protein and the entities that are connected to it.

We applied our method and Jaccard similarity score to each protein and first M_i proteins are collected from each rank list.

$$L^{(J)}(P_i) = \{P_{i.1}^{(J)}, ..., P_{i,Mi}^{(J)}\}, \quad L^{(*)}(P_i) = \{P_{i.1}^{(*)}, ..., P_{i,Mi}^{(*)}\} \quad (30)$$

In expression (30), $L^{(J)}(P_i)$ and $L^{(*)}(P_i)$ denote the top first M_i proteins from the rank list of similarity scores for i^{th} protein. Each rank list is created with a descending order of similarity scores. These lists were then compared to $L(P_i)$ for each protein and number of elements that are matching was counted. $T_i^{(J)}$ and $T_i^{(*)}$ represent the number of matching list elements for Jaccard and our scoring scheme respectively. In other words, we are comparing the list of proteins connected to a particular protein i (shows the existent links of i^{th} protein and it is denoted by $L(P_i)$) to the list of top similar proteins of i^{th} protein according to a particular similarity scoring framework. As higher similarity between a pair of proteins implies presence of an actual network link, we expect the top similar entities of a particular entity to be actually linked to the entity. Therefore, higher match between actual links and top similar entities shows a better prediction performance.

We defined total number of matching proteins for both methods as follows;

$$T^{(J)} = \sum_{i=1}^{R} T_i^{(J)} \qquad T^{(*)} = \sum_{i=1}^{R} T_i^{(*)} \quad (31)$$

In this expression R denotes the total number of proteins (R = 9575). One can also count the total number of connections in the networks as;

$$M = \sum_{i=1}^{R} M_i^{(*)} \quad (32)$$

There are $M = 268{,}706$ total connections in this network. We calculated total number of matching list elements for all proteins as follows; $T^{(*)} = 119{,}116$, $T^{(J)} = 89{,}436$. This shows that our method outperforms the Jaccard method for prediction of existing links in the network. Furthermore, we defined a relative prediction performance measure and plotted it for all proteins. This measure defined as follows;

$$Performance = \frac{T_i^{(*)} - T_i^{(J)}}{M_i} \quad (33)$$

In figure 7, one can see that for most of the proteins (5658 out of 9575 proteins) our method predicts more existing connections than the Jaccard. Only for 314 proteins the Jaccard score has better performance.

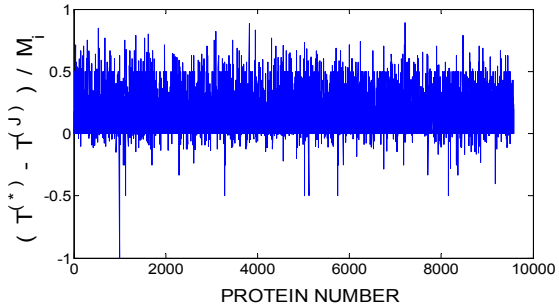

Figure 7. Relative link prediction performance of our method against Jaccard as in expression (33) for 9575 proteins.

4. CONCLUSIONS

The contribution of this work can be summarized in two ways. First, our method is a novel computational algorithm to quantify indirect relationships between the objects of biological research of interest by using existing relationships from text mining databases to automate the search for novel drug targets. This method can also be used for different purposes such as; annotating diseases with similar etiology, reposition of existing drugs, or discovering adverse events for the targets. Secondly, in a case study involving 9575 proteins in the Ariadne database, our method outperformed the Jaccard method for the prediction of existing links for all proteins. This illustrates its prediction capability for biological networks.

REFERENCES

Wren J. D., A., Bekeredjian, R., Stewart, J.A., Shohet, R.V., and Garner, H.R. (2004). *Knowledge Discovery by Automated Identification and Ranking of Implicit Relationships*. Bioinformatics, 20(3), 389-398.

Novichkova S., Egorov S., and Daraselia, N (2003). *MedScan a natural Language Processing engine for Medline abstracts,* Bioinformatics, 19 (13), 1699-1706.

Liben-Nowell, D., and Kleinberg, J. (2003). *Link Prediction Problem for Social Networks,* Proceedings of the Twelfth Annual ACM International Conference on Information and Knowledge Management.

Salton, G., and McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.