ALIVE: A Multi-relational Link Prediction Environment for the Healthcare Domain

Reid A. Johnson, Yang Yang, Everaldo Aguiar,
Andrew Rider, and Nitesh V. Chawla

Department of Computer Science and Engineering
University of Notre Dame, Notre Dame, IN 46556
{rjohns15,yyang1,eaguiar,arider1,nchawla}@nd.edu

Abstract. An underlying assumption of biomedical informatics is that decisions can be more informed when professionals are assisted by analytical systems. For this purpose, we propose ALIVE, a multi-relational link prediction and visualization environment for the healthcare domain. ALIVE combines novel link prediction methods with a simple user interface and intuitive visualization of data to enhance the decision-making process for healthcare professionals. It also includes a novel link prediction algorithm, MRPF, which outperforms many comparable algorithms on multiple networks in the biomedical domain. ALIVE is one of the first attempts to provide an analytical and visual framework for healthcare analytics, promoting collaboration and sharing of data through ease of use and potential extensibility. We encourage the development of similar tools, which can assist in facilitating successful sharing, collaboration, and a vibrant online community.

Keywords: Link Prediction, healthcare analytics, multi-relational networks.

1 Motivation

An idea that has taken root as the "fundamental theorem" of biomedical informatics is that a person working in partnership with an information resource is better than that same person unassisted. For this theorem to hold, however, the information resource must offer something that the person does not already have. As the people who interact with these resources often possess a high degree of knowledge relating to their domain of expertise, it can be challenging to offer people a resource that they find truly useful and informative.

Link prediction in complex networks has attracted attention from computer scientists and biologists for its ability to provide useful information. However, while most existing link prediction studies are designed for homogeneous networks, where only one type of object exists in the network [1, 11–13, 15], most networks are in reality heterogeneous and multi-relational [4, 9], and attribute values of objects are often difficult to fully obtain. Therefore, the use of topological features between objects in a heterogeneous network is critical to predicting

T. Washio and J. Luo (Eds.): PAKDD 2012 Workshops, LNAI 7769, pp. 36–46, 2013.

links in a holistic way. In multi-relational homogeneous networks, topological features have different values in different dimensions (relations), while in multi-relational heterogeneous networks the situation becomes more complicated, as the linkage types are different.

By applying link prediction to these types of networks, one can explore unknown or potential links between diseases, genes, and drugs, with findings that can lead to improved biological knowledge and clinical standards, and which can ultimately benefit the quality of healthcare. We propose an approach that uses cutting-edge link prediction algorithms to supply the accuracy needed to provide useful information and a visual environment that can assist healthcare professionals in making observations that can lead to innovation in the healthcare domain.

2 Proposed Environment

Healthcare professionals need data that is correct and informative, both of which can be challenging tasks. To address these challenges, we have developed a virtual platform called "ALIVE" (A Link Information and Visualization Environment). ALIVE is an online link-prediction environment oriented towards healthcare professionals and aimed at benefiting users from a variety of domains and with differing degrees of expertise. The environment takes advantage of the availability of health information records, facilitating data and knowledge management, network analysis, and visual analytics to pursue pioneering inter-disciplinary integration and providing tailored information. ALIVE also encompasses a novel link prediction algorithm that can improve the analysis of healthcare data. In our development of ALIVE, we have focused on developing a tool that will charter a path from data to knowledge to insight, ultimately supporting its users in making more informed healthcare decisions. Such tools can assist many efforts: an epidemiologist might learn of a potential relation between two diseases from a hunch, or a pharmaceutical researcher may discover that a particular drug is unexpectedly effective against a virulent disease for which it was not originally intended.

We foresee the potential for ALIVE to truly fulfill the concept encapsulated by the fundamental theorem of biomedical informatics. The environment has the potential to grow into a full-fledged virtual organization, serving as a data and knowledge warehouse and fostering expert collaboration. As an online platform, ALIVE has the potential to be a powerful information resource combined with a collaborative effort of informed people, which can undoubtedly achieve a vision far greater than the sum of its parts.

3 Background

The problem of predicting unknown links between diseases and genes continues to attract active interest from biological scientists, as it has proven useful in assisting research and make their work more efficient.

Despite a significant and continuous increase in medical research spending, the annual number of new drugs approved and new drug targets identified has remained almost constant for the past 20-25 years, with about twenty new drugs and about five new targets per year. At this rate it will take more than 300 years to double the number of available drugs [7, 14]. However, there are several ways to address these burdens. Promising areas of drug design include: wide-range screens of existing drugs, seeking novel applications, combination therapy, (the combined use of several drugs or short DNA oligomers) and the development of multi-target drugs [2, 3, 6, 8, 16].

Currently, interactions between diseases, genes, and drugs are studied separately; researchers usually only use interaction networks of diseases and gene (Fig. 1a) to predict disease and gene interactions, or only employ interactions between drug and gene (Fig. 1b) to predict drug-target interaction. We propose to combine these two kinds of networks together to improve understanding and analysis in the medical domain. We believe that the multi-relational network approach will allow us to improve predictions made relating to drug-target interaction.

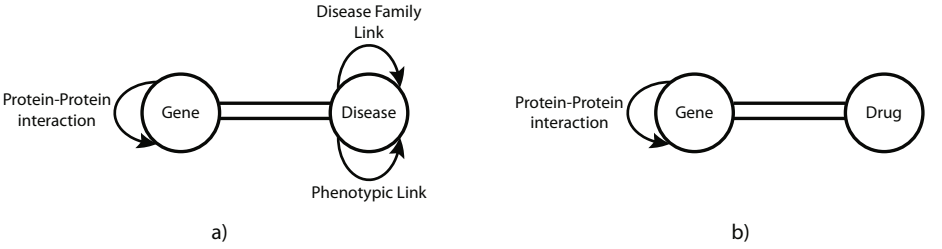

Fig. 1. A visual depiction of the types of interactions that exist in disease, gene, and drug networks. a) The links that may be present in disease-gene interactions. b) The links that may be present in disease-drug interactions.

4 Multi-relational Heterogeneous Networks

A network consists of nodes, representing some concept such as disease, and edges, representing relationships between these nodes. These relationships can encode a variety of information, such as whether diseases tend to occur in the same patient, whether they can be treated by the same drug, and whether they have the same underlying genetic causes. A typical network approach considers only a single type of edge. In contrast, a multi-relational network allows all edge types to exist in the network simultaneously, even overlapping each other. Overlapping edges contain additional information not available in the typical network approach: they may give additional support for two nodes or diseases being linked, or in combination they may specify a particular kind of relationship that was previously not understood.

An additional layer of information and complication is added in heterogenous networks when nodes can represent multiple concepts. For example, different nodes could represent either diseases or genes. An edge between the two types of nodes might represent the confidence with which a gene is related to a disease.

4.1 Link Prediction in Multi-relational Networks

Multi-relational link prediction is a new field of research in data mining. Few attempts have been made to solve this problem due to both difficulty in obtaining real data and the complications inherent in multi-relational networks.

The multi-relational link prediction problem can be described as follows: Given a multi-relational network $G = (V, E_1, E_2, \ldots, E_k)$, predict whether there is a link of type $i = (1, 2, \ldots, k)$ between pair of nodes u and v. To solve the problem of multiple edge types, one needs to know the relationships between each pair of nodes in the network. We define a parameter $\sigma(E_1, E_2)$ to represent the influence between two kinds of edges/relations in the network. $\sigma(E_1, E_2)$ is an asymmetric value, which means $\sigma(E_1, E_2) \neq \sigma(E_2, E_1)$. In other words, relation A and relation B may influence each other with differing degree. For instance, location-based data about people could greatly assist the prediction of their friendship relations, while friendships may not support the prediction of location to the same degree.

Our work builds on two previous approaches to link prediction, the Katz method and PropFlow. The Katz method is a variation on shortest path distance, directly summing over all the paths that exist between a pair of vertices. Specifically,

$$Katz(x, y) = \sum_{l=1}^{\beta_l} paths(x, y) \times l \tag{1}$$

where l is the path length and x and y are a pair of vertices, and β_l is a tuning parameter [10]. In effect, the method penalizes the contribution of longer paths in the similarity computation by exponentially reducing the contribution of a path by a factor of β_l.

PropFlow is an unsupervised path-based link predictor that models the link prediction score as being propagated radially outward from the source [13]. The algorithm uses a breadth-first search approach to propagate the probability that a restricted random walk starting at v_i ends at v_j in l steps or fewer using link weights as transition probabilities, where each score s_{ij} can serve as an estimation of the likelihood of new links. Formally, this likelihood score between nodes u and v is computed as

$$flow(u, v) = score(u) \times \frac{w(u, v)}{d(u)} \times \beta^{h-1} \tag{2}$$

where w is weight, d is degree, and h is the shortest number of hops from u to v.

5 MRPF Algorithm

In our experiments, we find that if we combine the original PropFlow method with the Katz method, we can achieve a higher area under the Receiver Operating Characteristic (AUROC) score than by using either alone. We alter PropFlow by penalizing scores by β so that the similarity between nodes u and v not only depends upon the weights of the shortest path between them, but also upon the number of hops in the path.

However, the PropFlow method as it is currently formulated cannot be directly applied to multi-relational networks; it is designed to work exclusively on single-relational networks, such as single mode homogeneous or bipartite networks. Therefore, we have developed a method to generalize PropFlow features to work on multi-relational networks, which we term multi-relational PropFlow (MRPF). The heuristics of MRPF are as follow:

1. For any two kinds of edges E_1 and E_2 in the network, the influence between E_1 and E_2 can be expressed by the correlation coefficient between their corresponding networks, denoted $\sigma(E_1, E_2)$ as previously described.
2. For node s and its neighbor t, the influence that flows from s to t in edge type E_i is as described in equation 3.

We find that using $p(E_1|E_2)$ in place of the correlation coefficient can achieve a better AUROC score for all of the datasets we have tested. Accordingly, we modify the calculation of flow from 2 to the following:

$$\begin{aligned} flow(s,t,i) = score(s) \times \beta^{h-1} \times \tfrac{w(s,t,i)}{d(s,k)} + \\ \beta^{h-1} \times p(i) \times (1-p(i)) \times \tfrac{\sum_{k \neq i}^{K} p(i|k) \times \frac{w(s,t,k)}{d(s,k)}}{K-1} \end{aligned} \tag{3}$$

where w is weight, d is degree, K is the number of edge types incident to source node s, and β is a tuning parameter. Generally we set $\beta = 0.05$.

Like PropFlow, our algorithm employs a breadth-first search to propagate information through the whole network, with the addition that we must compute each propagation K (number of edge types) times through an edge rather than only once. Therefore, the complexity of our algorithm is $K \cdot O(|V| \cdot |E|)$, which provides us the means of executing the algorithm in real-time for most practical datasets.

Figure 2 shows a conceptual overview of the MRPF algorithm. In the example, flow is propagated to successive nodes in relation to the degree of correlation. Starting from the source node with a score of 1, all neighboring nodes are given a weighted share of the score. The scores continue to flow outward, summing together for nodes that are reached by several paths.

6 Data

We acquired the data from one of our previous studies [5]. The disease networks were constructed based on the disease-gene associations from OMIM, Swiss-Prot,

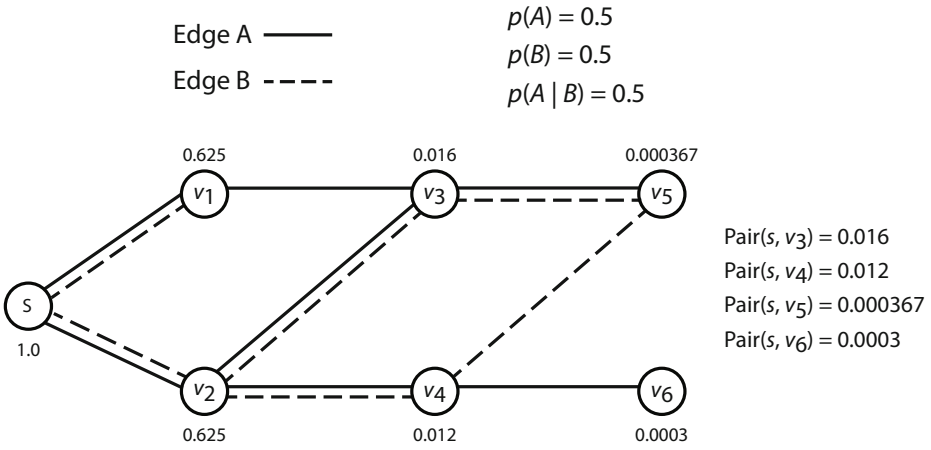

Fig. 2. A conceptual overview of our MRPF alogrithm. Flow propagates outward from the source node S.

and HPRD. The diseases are classified by Disease Ontology (DO) codes and the gene names are based on the HUGO Gene Nomenclature. We constructed a gene-disease network from this data using diseases and nodes, and establishing a link between nodes if the diseases share significantly more gene associations than randomly expected based on generality of the diseases.

Disease co-morbidity was calculated from patient medical diagnoses collected from a regional health system. Each data record is a single visit represented by an anonymized patient ID and a primary diagnosis code, as defined by the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). For consistency with the first dataset, the ICD-9-CM codes have been converted to Disease Ontology codes based on mappings provided within the DO coding. The mapping is many to many, so a single ICD-9-CM code often translates to a list of DO codes, and a DO code may apply to multiple ICD-9-CM codes as well. We constructed a phenotypic disease network from patient data, where the nodes are diseases and links indicated disease co-morbidity, where co-morbidity can be broadly defined as co-occurrence in the same patients significantly more than chance.

7 Evaluation

For all experiments, we use a 10-fold cross-validation stratified edge holdout scheme. We use holdout evaluation because longitudinal data was either not available or not relevant for disease, gene, and drug networks. Link prediction is evaluated for each edge type x separately on all eligible node pairs (s, t).

We evaluate each link predication algorithm using the the receiver operating characteristic curve (ROC). The ROC curve presents achievable true positive rates with respect to all possible false positive rates by varying the decision

threshold on probability scores. ROC curves can provide information about the operating range of link predictors, with the area under the ROC curve (AUROC) providing a measure of the performance over all predictive thresholds.

We discuss an example to calculate the AUROC. Assume a simple graph with five nodes, seven existing links, and three non-existent links ((1, 2), (1, 4), and (3, 4)). To test the algorithm's accuracy, we select some existing links as probe links. We may, for instance, pick (1, 3) and (4, 5) as probe links, which are presented by dashed lines in the right plot. This means that any algorithm being evaluated may only make use of the information contained in the training graph without (1, 3) and (4, 5).

Let us assume that the scores assigned by an algorithm to non-observed links are $s_{12} = 0.4, s_{13} = 0.5, s_{14} = 0.6, s_{34} = 0.5$, and $s_{45} = 0.6$. Then to calculate AUROC, we need to compare the scores of a probe link and a nonexistent link. There are in total six pairs: $s_{13} > s_{12}, s_{13} < s_{14}, s_{13} = s_{34}, s_{45} > s_{12}, s_{45} = s_{14}$ and $s_{45} > s_{34}$. Hence, the AUROC value equals 0.67.

8 Results

We applied MRPF and other link prediction methods to three biological networks, including a disease-gene network, a disease-disease-phenotype network, and a protein-protein interaction (PPI) network [5]. Table 1 shows results in terms of area under the ROC curve for several link prediction methods. The selected link predictors are among those most frequently used in the task of link prediction; included in these predictors is the latest method proposed by [4].

MRPF outperforms all of the methods on the disease network and PPI network and performs nearly as well as the best method in the phenotypic network. It is worth noting that while MRPF is outperformed on the phenotypic network by several methods, its performance is a significant improvement over that obtained by PropFlow; MRPF also demonstrates incremental improvements on the other networks. These results indictate that incorporating information on relation types into the flow algorithm can significantly improve performance.

Table 1. AUROC statistics for link prediction algorithms used on genetic, phenotypic, and protein-protein interaction networks. The highest values for each type of network are in bold.

Disease	PA	PF	JC	CN	AA	MRLP	MRPF
Genetic	0.903	0.951	0.957	0.951	0.956	0.974	**0.975**
Phenotypic	**0.943**	0.762	0.771	0.909	0.911	0.938	0.901
PPI	0.827	0.888	0.786	0.788	0.789	0.808	**0.890**

Note: PA = Preferential attachment, PF = PropFlow, JC = Jaccard's coefficient, CN = Common neighbors, AA = Adamic/Adar, MRLP = Multi-relational link predictor proposed by [4], and MRPF = Multi-relational PropFlow, proposed herein.

9 Interface Implementation

One of the goals of ALIVE is to facilitate analysis of medical data by healthcare professionals. Therefore we aspire to have an interface that allows non-computer scientists to use cutting-edge network analysis tools. The interface design is a key part of ALIVE that attempts to present analysis options and the results in an intuitive way.

Though the tools necessary to facilitate web-based visualization are not yet as advanced as those for application use, there are several libraries that provide the level of interaction that our project requires. We elected to use D3, a JavaScript library that allows one to bind arbitrary data to a Document Object Model (DOM), and to then apply data-driven transformations to the document. We use D3 to provide an interactive visualization of the relevant networks. Figure 3 provides a high-level illustration of how this interface is organized with relation to the MRLP framework.

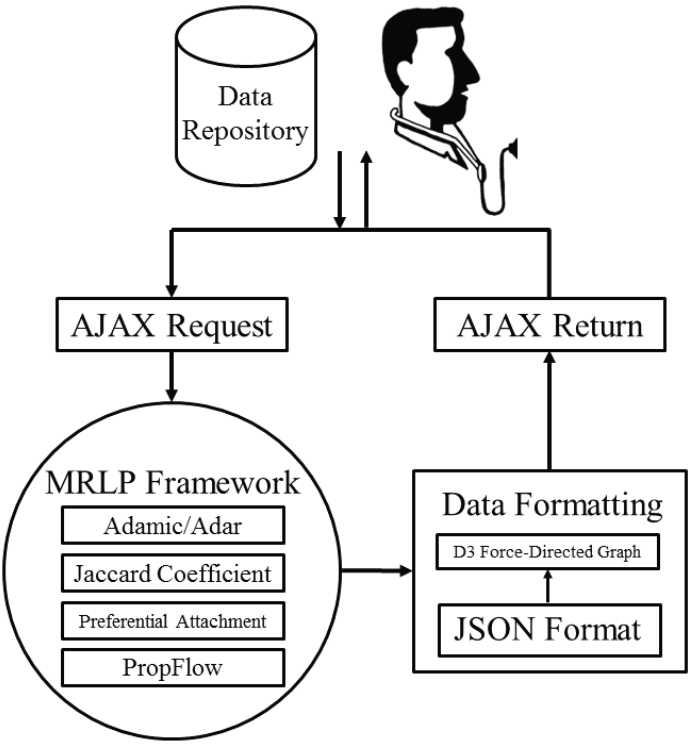

Fig. 3. A high-level component overview

ALIVE seeks to combine cutting-edge network analytics with healthcare data for use by healthcare professionals. Therefore, we designed the user interface

with a clean and simple layout, consisting of two basic parts. First there is the main tab, where most of the functionality is available, and to which users are first taken upon navigating to our web-service. There, they are be able to choose between one of the several (previously uploaded) datasets. Our front-end also allows users to select which predictor or predictors they wish to run on the selected data. In order to expand the applicability of our tool, users are also permitted to upload their own datasets and run a variety of algorithms on them.

9.1 Network Visualization

Networks are a natural fit to our goal of providing an informative view of healthcare data. Networks contain emergent properties that are held in their structure and made up of the way in which nodes are related to each other. A healthcare professional may be able to notice emergent patterns in a network that can be the basis for a hypothesis. ALIVE provides missing links in a network and allows healthcare professionals to apply their domain knowledge to a more complete picture of the data, and it does so with a dynamic, interactive visualization of the network generated by user-supplied data.

This work involved several components. As the visualization feature utilizes the functionality of a JavaScript package called D3, the output of the functions that compute the network attributes-which are written in Java and provide output as comma-separated files-needed to be converted to the JavaScript Object Notation (JSON), a file format compatible with JavaScript. This conversion was accomplished via the implementation of an additional Java class and corresponding methods. The D3 package was then leveraged to create a dynamic HTML page with the JSON file as input.

The current implementation is interactive and allows for nodes to be grouped by color, link weights to be designated by line stroke width, and tag information to be displayed for each node. Conceptually, the interface allows users to upload data, which can be evaluated by the MRLP framework with relation to a current data repository. The resulting scores computed by the link prediction algorithms are output visually.

9.2 Extensibility

As developed, our framework also allows for enormous extensibility. With minor additions, users would not only be able to interact with the output via the visualization, but could also have the option of exporting that particular output and saving it for later reference. For networks containing a large number of nodes, we could adjust the graph crop factor (the cutoff for edge inclusion) and let the user zoom in and out for a more versatile visualization. Moreover, as a web-based tool, ALIVE could be expanded into a general repository of healthcare information, allowing health professionals to submit and share data.

10 Conclusions

ALIVE has a great deal of potential as a useful tool in healthcare analytics. Not only have we contributed to the science of link prediction, but we have also provided the basis for an accessible, web-based tool, that has potential to be the nucleosing agent for a healthcare data warehouse. If expanded, ALIVE could ultimately foster a vibrant online community of healthcare professionals, providing the tools necessary to facilitate successful collaboration.

References

1. Al Hasan, M., Chaoji, V., Salem, S., Zaki, M.: Link prediction using supervised learning. In: Workshop on Link Discovery: Issues, Approaches and Apps, Citeseer (2005)
2. Borisy, A.A., Elliott, P.J., Hurst, N.W., Lee, M.S., Lehár, J., Price, E.R., Serbedzija, G., Zimmermann, G.R., Foley, M.A., Stockwell, B.R., et al.: Systematic discovery of multicomponent therapeutics. Proceedings of the National Academy of Sciences 100(13), 7977 (2003)
3. Csermely, P., Agoston, V., Pongor, S.: The efficiency of multi-target drugs: the network approach might help drug design. Trends in Pharmacological Sciences 26(4), 178–182 (2005)
4. Davis, D., Lichtenwalter, R., Chawla, N.V.: Multi-relational link prediction in heterogeneous information networks. In: 2011 International Conference on Advances in Social Networks Analysis and Mining, pp. 281–288. IEEE (2011)
5. Davis, D.A., Chawla, N.V.: Exploring and exploiting disease interactions from multi-relational gene and phenotype networks. PloS One 6(7), e22670 (2011)
6. Diacon, A.H., Pym, A., Grobusch, M., Patientia, R., Rustomjee, R., Page-Shipp, L., Pistorius, C., Krause, R., Bogoshi, M., Churchyard, G., et al.: The diarylquinoline tmc207 for multidrug-resistant tuberculosis. New England Journal of Medicine 360(23), 2397–2405 (2009)
7. DiMasi, J.A., Hansen, R.W., Grabowski, H.G.: The price of innovation: new estimates of drug development costs. Journal of Health Economics 22(2), 151–185 (2003)
8. Fitter, S., James, R.: Deconvolution of a complex target using dna aptamers. Journal of Biological Chemistry 280(40), 34193 (2005)
9. Han, J.: Mining Heterogeneous Information Networks by Exploring the Power of Links. In: Gama, J., Costa, V.S., Jorge, A.M., Brazdil, P.B. (eds.) DS 2009. LNCS, vol. 5808, pp. 13–30. Springer, Heidelberg (2009)
10. Katz, L.: A new status index derived from sociometric analysis. Psychometrika 18(1), 39–43 (1953)
11. Leroy, V., Cambazoglu, B.B., Bonchi, F.: Cold start link prediction. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 393–402. ACM (2010)
12. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology 58(7), 1019–1031 (2007)
13. Lichtenwalter, R.N., Lussier, J.T., Chawla, N.V.: New perspectives and methods in link prediction. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 243–252. ACM (2010)

14. Ma'ayan, A., Jenkins, S.L., Goldfarb, J., Iyengar, R.: Network analysis of fda approved drugs and their targets. Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine 74(1), 27–32 (2007)
15. Wang, C., Satuluri, V., Parthasarathy, S.: Local probabilistic models for link prediction. In: Seventh IEEE International Conference on Data Mining, ICDM 2007, pp. 322–331. IEEE (2007)
16. Wong, P.K., Yu, F., Shahangian, A., Cheng, G., Sun, R., Ho, C.M.: Closed-loop control of cellular functions using combinatory drugs guided by a stochastic search algorithm. Proceedings of the National Academy of Sciences 105(13), 5105 (2008)