ISSN : 0974 - 7435 Volume 8 Issue 11

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 8(11), 2013 [1504-1510]

A computational approach for generating models of signal transduction networks

Tong Wang*, Xiaoxia Cao, Tian Xia, Yueping Wu

Institute of Computer and Information, Shanghai Second Polytechnic University, Shanghai, 201209, (CHINA)

E-mail: zjxywt@163.com

ABSTRACT

Signal transduction is important in many different aspects of cellular activity. Many computational methods have been generated in mining signal transduction networks with the increasing of high-throughput genomic and proteomic data. However, more effective means are still needed to understand the complex mechanisms of signaling pathways. In this paper, we have developed a computational approach for generating models of signal transduction networks. Networks are determined entirely by protein-protein interaction data without prior knowledge of any pathway intermediates. This approach should enhance our ability to model signaling networks and to discover new components of known networks. The precision and recall values of our method are comparable with other existing methods. Our method is a more suitable method than existing methods for detecting underlying signaling pathways.

KEYWORDS

Signal transduction networks;
Modeling;
Minimax distance metric algorithm.

INTRODUCTION

Systems biology is an emerging field in biology whose aims are to understand the complex biological systems at the system level[1]. System-level understanding of cell networks requires a lot of principles and methodologies that links the behaviors of molecules to system characteristics and functions. Cell signalling pathways are enormously complex because they include a huge number of different molecules and biochemical interactions. The dynamic behaviors of biological systems are deeply affected by their structural complexity and uncertainty of some kinetics parameters[2]. A world of complicated systems in nature can be described by various networks. One representative network consists of many nodes and the side that connects two nodes. Node is applied to represent different elements in the real system while side is employed to indicate the relationship among elements, usually, there will be a side when two nodes have certain specific relationship. Two nodes connected by side are considered as adjacent. Network model is the most effective model to describe the complicated system.

Signal transduction plays an essential role in cell response to environment changes. Signal transduction is the primary means by which cells coordinate their metabolic, morphologic, and genetic responses to environmental cues such as growth factors, hormones, nutrients, osmolarity, and other chemical and tactile stimuli. This biological process is usually characterized by phosphorylation of some key proteins and generally involves a signal cascade. The signal transduction process often starts from a membrane protein, spans a series of intercellular signaling proteins and then transfers to transcription factors in the nucleus, subsequently raising the expression of downstream genes. Therefore, it is necessary to study how to quantitatively determinate the relation between system behaviors and parameter variations, and how to investigate the interaction of parameters. With the development of molecular biology and high throughput experimental techniques, a large number of data sets have been obtained so that it is possible to study cell signaling transduction networks quantitatively[3,4].

Another source of evidences on the key role of transcription factors in regulating cellular regulatory processes comes from analysis of signal transduction pathways. Multiple signal transduction pathways of a cell transducer extracellular signals from receptors at the cellular membrane to the transcription factors in the nucleus where they regulate the transcription of genes. There are several databases that collect information about signal transduction pathways in different cells. Among them, the TRANSPATH database[5] stores a large body of information on signaling pathways allowing computational search through the graph of signaling reactions. One aim of such searches is to find the key transcription factors that mediate the concerted changes in expression of specific components of the signal transduction network.

Studies demonstrate that many important cellular processes such as cell proliferation, differentiation, cell cycle control and cellular responses to nutrient limiting conditions are involved in different signaling pathways[6,7]. For example, Tang et al[8] showed that the receptor kinase BRI1 and BR-signaling kinases (BSKs) mediate growth regulation related signal transduction in Arabidopsis. The Toll-like receptor (TLR) signaling cascade plays an essential role in recognizing and eliciting responses upon invasion of pathogens[9,10]. Recent highthroughput genomic and proteomic techniques, such as Co-Immunoprecipitation (Co-IP)[11,12], protein chip[13-16] and microarray experiments[17] have generated enormous amounts of data for uncovering signal transduction networks. This abundance of information brings increasing complexity to network analysis, which is a major obstacle to understanding the mechanisms of cell signaling[18].

Although these methods have been highly effective in generating detailed descriptions of specific linear signaling pathways, our knowledge of complex signaling networks and their interactions remains incomplete. Recently, new computational methods that capture molecular details from high-throughput genomic data in an automated fashion are desirable and can help direct the established techniques of molecular biology and genetics. Steffen et al[19] developed a static model, NetSearch, to reconstruct the signal transduction network from PPI and gene expression data. Some computational methods, e.g. gene co-expression[20] and semantic similarity of Gene Ontology (GO) annotations[21], indicate that genes with high scored interactions may be involved in the same signaling pathway[22]. However, this information either is limited or has not been incorporated in most databases constructed from experimental data. Though these interactions may not necessarily be direct interactions, using this information may help to improve prediction of signal transduction networks. We define “direct interaction” as a direct physical association between two proteins and “indirect interaction” as no direct physical association between two proteins in the actual state. Two proteins with indirect interaction must function through at least one medial protein.

We present a computational approach for generating static models of signal transduction networks which utilizes protein-interaction maps generated from large-scale two hybrid screens and expression profiles from DNA microarrays. Networks are determined entirely by protein-protein interaction data without prior knowledge of any pathway intermediates. In effect, this is equivalent to extracting subnetworks of the protein interaction dataset whose members have the most correlated expression profiles.

MATERIALS AND METHODS

Dataset

Here, only the PPI dataset was employed. The Yeast

Proteome Database (YPD)[23,24], Saccharomyces Genome Database (SGD)[25] and Database of Interacting Proteins (DIP)[26,27] are the most frequently used PPI databases, but the interaction dataset in those databases is very limited, which may lead to misconnections due to deficient data. In this study, we constructed a PPI dataset from the STRING database (Version 8.3)[28]. The current STRING database contains 6,015 yeast proteins and 245,782 yeast protein interactions. Our database contains both direct and indirect PPIs derived from both computational methods and biological experiments, providing more comprehensive information than previously used.

Scoring system

To score the PPI pairs in the combined database, we used the STRING scoring system[21]. The STRING database infers PPIs through various approaches, including the neighbourhood method, fusion events, co-occurrence, co-expression, experimental methods and text-mining. It integrates all probabilities of those methods and assigns each PPI pair a reasonable score[20]. The original PPI score in STRING database is from 0 to 999, which is subsequently normalized from 0.000 to 0.999 by dividing by 1000. However, not all of the proteins would have a corresponding GO term in the annotation file. So, the protein is represented based on the strategy of hybridizing the gene ontology (GO) database[29] and PseAAC[30]. The GO approach has been used for predicting protein subcellular localization, membrane protein type, and enzyme functional class. Again, not all protein samples can be meaningfully defined in the GO space. To overcome such a problem, an approach was developed by hybridizing the GO space with the PseAAC space. The GO database contains 20,126 numbers. With each of the 20,126 sequences as a vector base, a given protein sample can be defined as a 20,126-D vector according to the following procedures. To compare the protein sequence with each of the 20,126 sequences in the GO database, if “hit” is found, then the ith component of the protein in the 20,126-D space is assigned 1; otherwise, it is assigned 0. The protein sample P in the GO space can be formulated as:

$$\mathbf{P}_{GO} = [A_1 \quad A_2 \quad \cdots \quad A_i \cdots A_{20126}]^T \qquad (1)$$

where T is the transpose operator, and

$$A_i = \begin{cases} 0, & \text{when a hit is found for P in GO database} \\ 1, & \text{otherwise} \end{cases} \qquad (2)$$

On the other hand, according to the concept of PseAAC[30], the protein sample P can be represented by

The concept of PseAA (Pseudo Amino Acid) composition was proposed by incorporating the sequence order information completely. According to the PseAA composition discrete model, the protein p can be formulated as

$$P_{PseAA} = [p_1, p_2, ..., p_{20}, p_{20+1}, ..., p_{20+\xi}]^T, \ (\zeta < N) \qquad (3)$$

where the 20+ξ components are given by

$$p_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\xi}\varphi_j}, & (1 \le \text{k} \le 20) \\[2ex] \dfrac{w\varphi_{k-20}}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\xi}\varphi_j}, & (20{+}1 \le \text{k} \le 20{+}\xi) \end{cases} \qquad (4)$$

where w is the weight factor, which was set at 0.05 in Ref. and φ_j is the j th tier correlation factor, which reflects the sequence order correlation between all of the th most contiguous residues. f_k is the occurrence frequencies of 20 amino acids in sequence. Because the length of the shortest protein sequence in the benchmark dataset is N=39, the value allowed forin Eqs. 2 and 3 is 24. Hence, the PseAA is actually corresponding to a 20+24=44-D (Dimensionality) vector.

Data pre-processing

To reduce the false positive rate, two pre-processing steps were carried out to exclude obviously irrelevant proteins but to keep the high correlated proteins as much as possible. Firstly, given several seed proteins, which we assumed to be known components in a signaling pathway, DFS algorithm was realized. DFS algorithm would search for all proteins connected with each seed protein within a certain path length. The common proteins within this scope were kept.

Then, the graph search algorithm Dijkstra is usually employed to calculate the distance between any two proteins. Here, Minimax Distance Metric algorithm is adopted. In order to represent this new dissimilarity metric, we need to introduce a path-based criterion for connectedness firstly. We denote the data set of n points by $X = \{x_1, x_2, \cdots, x_n\}$. The data points can be represented as a fully connected graph with vertices corresponding to the n points. Each edge (x_i, x_j) in the graph is assigned a weight d_{ij} reflecting the original dissimilarity between x_i and x_j. Euclidean distance is usually chosen as original dissimilarity.

A path from one vertex to another vertex through the above fully connected graph is a sequence. There may have many possible paths between this pair of vertices. Let p_{ij} denotes the set of all paths from vertex x_i to vertex x_j through the graph. For example, $p_{ij}^1 = (x_i, x_j)$, $p_{ij}^2 = (x_i, x_5, x_8, x_j)$, $p_{ij}^3 = (x_i, x_4, x_j)$, and so on. We define a single hop is walking from one vertex to another vertex of an edge. And the single hop distance is the weight of the edge. For each path $p_{ij}^k \in p_{ij}$ (where k is an index to enumerate all possible sequences between x_i and x_j), the effective dissimilarity $d_{p_{ij}}^k$ between vertices x_i and x_j (or the corresponding data points and)

is the maximum single hop distance in p_{ij}^k. We define the total dissimilarity M_{ij} between vertices x_i and x_j as the minimum of all effective dissimilarities $d_{p_{ij}}^k$:

$$M_{ij} = \min_k d_{p_{ij}}^k = \min_{p_{ij}^k}\left\{ \max_{(x_i, x_j) \subset p_{ij}^k} d_{ij} \right\}. \qquad (5)$$

From the above Eq. (5), we can easily draw a conclusion that the proposed dissimilarity M_{ij} between the two points is less than original dissimilarity d_{ij} while the two points x_i and x_j lie on the same branch of the manifold (or in same class), and is equal to while on different branches of manifold (or in different class). Now we take an example to show the details of how to choose the appropriate neighborhood in the new neighborhood selection method.

5000 points randomly sampled on the Swissroll data is given in Figure 1. There are five points on this Swissroll manifold. It can be observed from the figure that x_1, x_2, x_4 and x_5 are on one branch and x_3 is on another one. The Euclidean distance d_{13} is smaller than d_{12} seen from Figure 1. According to original neighborhood selection method based on Euclidean distance, x_3 but not x_2 may be chosen as the neighbor of x_1. Though, the Euclidean distance between andmay be deceptively small in the three-dimensional Swissroll space, their distance on an intrinsic two-dimensional manifold is large (the intrinsic dimension of Swissroll is two). So the Euclidean distance may not accurately reflect their intrinsic dissimilarity. This problem can be remedied by using minimax distance metric. In Figure 1, supposed that p_{12} is one path from x_1 to x_2 through the fully connected graph (a consecutive path on the same branch of the manifold or in the same classification) and each single hop distance between one point and its nearest point on the path is sufficiently small (i.e., the single hop distance d_{45} between x_4 and x_5). However, for another path p_{13} (a path from one branch of the manifold or classification to another), assuming the single hop distance d_{13} between x_1 and x_3 is big. Namely,there is at least one “great leap” from x_1 to x_3. Under the new minimax distance metric, the

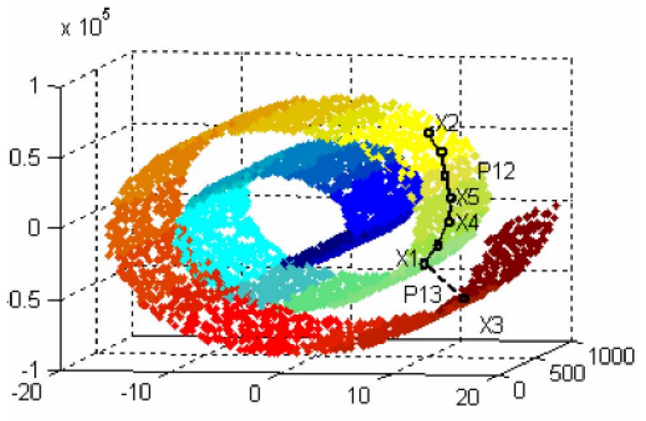

Figure 1 : A sample of applying minimax distance metric for neighborhood selection

dissimilarity M_{12} (x_1 and x_2 on the same branch of the manifold) is smaller than the original distance d_{12} for its small single hop distance on the path p_{12} . On the other hand, the dissimilarity M_{13} (x_1 and x_3 on the different branch of manifold) is still equal to d_{13} . As a result, it makes it possible that the dissimilarity M_{12} is smaller than M_{13} and x_2 is the neighbor of of x_1 instead of x_3 .

Given the initial node, this algorithm will find the minimax distance between this node and any other node in the graph. Hence, it is often used in solving routing related problems. Consequently, the distances between each protein and any seed protein are obtained. Generally speaking, if the distance of a node to the “actual network” is maximal, the node represents the steepest descent direction. However, the “actual network” is unknown, we simply use the distance of the node to any of the given nodes (i.e. seed proteins) as the distance to the “actual network”. Hence, the node corresponding to the maximal distance will be selected as a candidate steepest descent node to remove from the network. While, if removing the candidate steepest descent node leads to a disconnection between given nodes, the node corresponding to the next maximal distance is selected as a candidate steepest descent node, and so forth.

To date, the network with the highest score has been obtained. For comprehensive consideration, we extended this restriction to the top N highest scored network.

RESULTS AND DISCUSSION

The filamentous growth pathway regulates cellular response to nutrient limiting conditions. For this pathway, there are many common proteins with other MAPK pathways. So, we randomly selected three or four seed proteins. Different parameters were also tested. After five independent experiments, we obtained an average of 85% recall and 32% precision (TABLE 1). TABLE 1 shows the performance of our method in detecting the filamentous growth pathway compared with that of NetSearch. Our method clearly shows both higher recall and precision than the other method. In addition, our method seems to predict fewer edges between the proteins in the predicted signal transduction networks comparing with other methods. Hence, even though the membrane receptor and transcription factor are not known, we still know where the signal is from and to among those proteins, since most proteins have only one link to the preceding and succeeding element in the predicted network. In fact, if we require the order between the proteins to be more intuitive, fewer edges should be kept in the predicted network. We achieved this goal by maximizing the average weight of the network while keeping most of the reliable interactions.

TABLE 1 : Performance comparison between different methods in precision and recall for filamentous growth pathway

Method	Precision(%)	Recall(%)
Our method	32	85
Netsearch	29	63

CONCLUSIONS

Generally, some potential proteins involved in a signaling pathway stimulated by environmental factors are easily available through various reliable means, such as manual literature curation and biological experiments. But in most situations, not all or none of these proteins are membrane receptors or transcription factors. Moreover, the proteins we obtained may be more than just two proteins. The proteins were represented by hybridizing the GO (gene ontology) approach with the PseAAC (pseudo amino acid composition) approach. Therefore, our method is more suitable for actual biological application compared with existing methods such as NetSearch. Nevertheless, although those methods utilize a more reliable dataset, the data is limited. However, using computationally predicted interactions may make up for the deficiency of experiment data, which is also one of our original aims.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (Grant No. 61301249 and

Grant No.61272036), Innovation Program of Shanghai Municipal Education Commission (Grant No.12YZ175) and The Soft Engineering of Key Subjects Construction in Shanghai Second Polytechnic University (Grant No. XXKZD1301).

REFERENCES

[1] H.Kitano; Systems Biology: Toward System-level Understanding of Biology Systems, In Foundations of Systems Biology, MIT Press, Cambridge, Massachusetts, **(2001)**.

[2] U.S.Bhalla, R.Iyengar; Emergent properties of networks of biological signaling pathways, Science, **283**, 381-387 **(1999)**.

[3] S.R.Neves, R.Iyengar; Modeling of signaling networks, BioEssays, **24(12)**, 1110-1117 **(2002)**.

[4] A.Asthagiri, D.Lauffenburger; Bioengineering models of cell signaling, Annu.Rev.Biomed.Eng., **2**, 31-53 **(2000)**.

[5] M.Krull, S.Pistor, N.Voss, A.Kel, I.Reuter, D.Kronenberg, H.Michael, K.Schwarzer, A.Potapov, C.Choi, O.Kel-Margoulis, E.Wingender; TRANSPATH: An Information Resource for Storing and Visualizing Signaling Pathways and their Pathological Aberratins. Nucleic Acids Res., **34**, D546-D551 **(2006)**.

[6] T.Hunter; Signaling–2000 and beyond. Cell, **100**, 113-127 **(2000)**.

[7] A.Takahashi, N.Ohtani, E.Hara; Irreversibility of cellular senescence: dual roles of p16INK4a/Rb-pathway in cell cycle control. Cell Div, **2**, 10 **(2007)**.

[8] W.Tang, T.W.Kim, J.A.Oses-Prieto, Y.Sun, Z.Deng, S.Zhu, R.Wang, A.L.Burlingame, Z.Y.Wang; BSKs mediate signal transduction from the receptor kinase BRI1 in Arabidopsis. Science; **321**, 557-560 **(2008)**.

[9] T.Lang, A.Mansell; The negative regulation of Toll-like receptor and associated pathways. Immunol Cell Biol, **85**, 425-434 **(2007)**.

[10] T.Ito, T.Chiba, R.Ozawa, M.Yoshida, M.Hattori, Y.Sakaki; A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA, **98**, 4569-4574 **(2001)**.

[11] R.B.Free, L.A.Hazelwood, D.R.Sibley; Identifying novel protein-protein interactions using co-immunoprecipitation and mass spectroscopy. Curr Protoc Neurosci, Chapter 5, Unit 5, 28 **(2009)**.

[12] J.S.Bridgeman, M.Blaylock, R.E.Hawkins, D.E.Gilham; Development of a flow cytometric co-immunoprecipitation technique for the study of multiple protein-protein interactions and its application to T-cell receptor analysis. Cytometry A, **77**, 338-346 **(2010)**.

[13] H.Zhu, M.Bilgin, R.Bangham, D.Hall, A.Casamayor, P.Bertone, N.Lan, R.Jansen, S.Bidlingmaier, T.Houfek et al; Global analysis of protein activities using proteome chips. Science, **293**, 2101-2105 **(2001)**.

[14] M.G.Smith, G.Jona, J.Ptacek, G.Devgan, H.Zhu, X.Zhu, M.Snyder; Global analysis of protein function using protein microarrays. Mech Ageing Dev, **126**, 171-175 **(2005)**.

[15] J.Fasolo, M.Snyder; Protein microarrays. Methods Mol Biol, **548**, 209-222 **(2009)**.

[16] L.A.Kung, M.Snyder; Proteome chips for whole-organism assays. Nat Rev Mol Cell Biol, **7**, 617-622 **(2006)**.

[17] M.Gilchrist, V.Thorsson, B.Li, A.G.Rust, M.Korb, J.C.Roach, K.Kennedy, T.Hai, H.Bolouri, A.Aderem; Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4. Nature, **441**, 173-178 **(2006)**.

[18] Kai Wang1, Fuyan Hu2, Kejia Xu2, Hua Cheng1, Meng Jiang1, Ruili Feng1, Jing Li1 and Tieqiao Wen1*; CASCADE_SCAN: mining signal transduction network from high-throughput data based on steepest descent method, BMC Bioinformatics, **12(1)**, 164 **(2011)**.

[19] M.Steffen, A.Petti, J.Aach, P.D'Haeseleer, G.Church; Automated modeling of signal transduction networks. BMC Bioinformatics, **3**, 34 **(2002)**.

[20] C.Von Mering, M.Huynen, D.Jaeggi, S.Schmidt, P.Bork, B.Snel; STRING: a database of predicted functional associations between proteins. Nucleic Acids Res, **31**, 258-261 **(2003)**.

[21] S.Jain, G.D.Bader; An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. BMC Bioinformatics, **11**, 562 **(2010)**.

[22] Y.Liu, H.Zhao; A computational approach for ordering signal transduction pathway components from genomics and proteomics Data. BMC Bioinformatics, **5**, 158 **(2004)**.

[23] W.E.Payne, J.I.Garrels; Yeast Protein database (YPD): a database for the complete proteome of Saccharomyces cerevisiae. Nucleic Acids Res, **25**, 57-62 **(1997)**.

[24] P.E.Hodges, W.E.Payne, J.I.Garrels; The Yeast

Protein Database (YPD): a curated proteome database for Saccharomyces cerevisiae. Nucleic Acids Res., **26**, 68-72 **(1998)**.

[25] K.R.Christie, S.Weng, R.Balakrishnan, M.C.Costanzo, K.Dolinski; S.S.Dwight, S.R.Engel, B.Feierbach, D.G.Fisk, J.E.Hirschman et al; Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. Nucleic Acids Res., **32**, D311-314 **(2004)**.

[26] I.Xenarios, L.Salwinski, X.J.Duan, P.Higney, S.M.Kim, D.Eisenberg; DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res, **30**, 303-305 **(2002)**.

[27] L.Salwinski, C.S.Miller, A.J.Smith, F.K.Pettit, Bowie JU, Eisenberg D: The Database of Interacting Proteins: 2004 update. Nucleic Acids Res., **32**, D449-451 **(2004)**.

[28] L.J.Jensen, M.Kuhn , M.Stark, S.Chaffron, C.Creevey, J.Muller, T.Doerks, P.Julien, A.Roth, M.Simonovic et al; STRING 8–a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res., **37**, D412-416. **(2009)**

[29] M.Ashburner, C.A.Ball, J.A.Blake, D.Botstein, H.Butler, J.M.Cherry, A.P.Davis, K.Dolinski, S.S.Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L.Issel-Tarver, A.Kasarskis, S.Lewis, J.C.Matese, J.E.Richardson, M.Ringwald, G.M.Rubin, G.Sherlock; Gene ontology: tool for the unification of biology, Nat.Genet; **25**, 25–29 **(2000)**.

[30] K.C.Chou; Prediction of protein cellular attributes using pseudo amino acid composition, Proteins, **43**, 246–255 (Erratum: 44 (2001) 60) **(2001)**.