Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Theoretical Biology 247 (2007) 608–615
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

The community structure of human cellular signaling network

Yuanbo Diao[a,b], Menglong Li[a,b,*], Zinan Feng[a], Jiajian Yin[a], Yi Pan[a]

[a]College of Chemistry, Sichuan University, Chengdu, Sichuan 610064, China
[b]State Key Laboratory of Biotherapy, Sichuan University, Chengdu, Sichuan, 610064, China

Received 1 August 2006; received in revised form 6 April 2007; accepted 6 April 2007
Available online 12 April 2007

Abstract

Living cell is highly responsive to specific chemicals in its environment, such as hormones and molecules in food or aromas. The reason is ascribed to the existence of widespread and diverse signal transduction pathways, between which crosstalks usually exist, thus constitute a complex signaling network. Evidently, knowledge of topology characteristic of this network could contribute a lot to the understanding of diverse cellular behaviors and life phenomena thus come into being.

In this presentation, signal transduction data is extracted from KEGG to construct a cellular signaling network of *Homo sapiens*, which has 931 nodes and 6798 links in total. Computing the degree distribution, we find it is not a random network, but a scale-free network following a power-law of $P(K)\sim K^{-\gamma}$, with γ approximately equal to 2.2. Among three graph partition algorithms, the Guimera's simulated annealing method is chosen to study the details of topology structure and other properties of this cellular signaling network, as it shows the best performance. To reveal the underlying biological implications, further investigation is conducted on ad hoc community and sketch map of individual community is drawn accordingly. The involved experiment data can be found in the supplementary material.

Keywords: Cellular signaling network; Complex system; Graph Partition Algorithm; Community structure; Crosstalk

1. Introduction

To most biologists of the last century, organism is a complex assembled "machine", following fundamental rules of physics and chemistry. These researchers believe that no matter how complex this machine could be, the mystery of which shall be ultimately unveiled by investigating its every separate part. Evidently this is a philosophy of reduction. Accordingly, last century's life science is a typical experiment discipline adopting the strategy of "divide and conquer". However, along with the accomplishment of Human Genome Project and the advent of Post-Genome era, scientists gradually recognize that life is not an automata but a complex system following rules much different from reductionism (Kitano, 2002).

The characteristic of complex system is nonlinearity, which also means "the whole is not equal to the sum of its parts". The molecular foundation of this nonlinearity is the generic and intricate interactions among all sorts of biomacromolecules, genes and proteins. Moreover, these biomacromolecules never behave or perform their biological functions alone, but have many direct or indirect relations between each other, which could be in physical or chemical manner. It is these relations that bring about various biological networks, such as metabolic network, gene regulation network, and signal transduction network, etc. Ultimately, all activities of life fall back on these networks in their structure and function (Maslov and Sneppen, 2002).

In recent years, signal transduction information has dramatically increased since the wide use of large-scale, high-throughput experiment techniques, such as biochips (or microarrays), bio mass spectrometry (Fenn et al., 1989), yeast two-hybrid system(Vidal and Legrain, 1999), and protein affinity chromatography(Lee, 2004), etc. Now it is possible to investigate cellular signaling network at systematic level. Moreover, knowledge of topology characteristic of cellular signaling network contributes a lot to the study of cellular dynamics, hence the reconstruction of

*Corresponding author. Fax: +86 028 85412356.
E-mail address: liml@scu.edu.cn (M. Li).

0022-5193/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2007.04.007

large scale cellular signaling network would promote our cognition to diverse cellular behaviors and life phenomena thus come into being.

In this paper, the use of signal transduction data to generate a cellular signaling network of *Homo sapiens* is reported. Static geometric analysis proves that this network is not a random network and significant community structure should exist. After a comparison of three graph-partition algorithms, the topologies of the network are studied and biological implications discussed.

2. Materials and methods

2.1. The data set

Kyoto Encyclopedia of Genes and Genomes (KEGG) is a knowledge repository for systematic analysis of gene functions in terms of the networks of genes and molecules (Kanehisa and Goto, 2000). The information in which we are most interested is stored in the PATHWAY database, which contains graphical representations of cellular processes such as metabolic, membrane transportation, signal transduction, cell cycle, etc. The foundation and maintenance of KEGG is sponsored by Japan government, which is free for academic and noncommercial uses.

All the signal transduction data of this study come from the PATHWAY database of KEGG (up to December 2005). Through simple object access protocol (SOAP), we employ the web service provided by KEGG to extract the signal transduction data of *H. sapiens*. By transforming them into neighbor matrixes and then combining these neighbor matrixes through matrix operation, we construct an undirected graph of the cellular signaling network of *H. sapiens*, which contains 931 nodes and 6798 links altogether.

2.2. Graph Traverse Algorithm

Using graph theory to study biology systems can make the problems concerned more intuitive, facilitating illustration and stimulating imagination so as to help reveal the essence of the problem concerned. Graphic approach has been successfully used to study enzyme-catalyzed system (Chou and Forsen, 1980; Chou and Liu, 1981; Chou, 1981, 1983, 1989; Lin and Neet, 1990; Kuzmic and Heath, 1992; Zhou and Deng, 1984), HIV reverse transcriptase inhibition mechanisms (Chou et al., 1994; Althaus et al., 1993a, b, c), protein folding kinetics (Chou, 1990, 1993), and analysis of base frequencies in the anti-sense strands of human protein coding sequences (Zhang and Chou, 1996). Recently, the images of cellular automata were also used to represent biological sequences (Xiao et al., 2005a), predict protein subcellular location (Xiao et al., 2006a), investigate HBV virus gene missense mutation (Xiao et al., 2005b) and HBV viral infections (Xiao et al., 2006b), as well as analyze the fingerprint of SARS coronavirus (Wang et al., 2005).

In the study of a graph, the traverse algorithm is often the first approach to learn its global features, such as the distribution of degree, shortest paths and betweenness, etc. To traverse a graph, two algorithms are frequently used. One is depth-first search algorithm (DFS); the other is breadth-first search algorithm (BFS). In DFS, the deeper is the vertex located, the sooner will it be expanded. Yet in BFS, on the contrary, the deeper vertices could not be reached until all the upper vertices had been traversed and handled.

For complex graphs such as in this study, BFS, using queue (first in first out) as its data infrastructure, is more preferable than DFS thereby, which uses stack (last in first out) instead (Aho and Hopcroft, 1983). The main principle of BFS is as following:

(1) Initialize all vertices as unassigned.
(2) Assign a random vertex r a distance zero to indicate that it is zero steps away from itself. Construct a first-in, first-out queue Q initially containing only node r.
(3) For the vertex u at the head of the Q, follow each attached edge to the vertex v at the other end. If v has not been assigned, assign it a distance $d[u]+1$.
(4) Remove u from the head of the queue. Repeat from step 3 until there are no unassigned nodes remaining.

2.3. Graph Partition Algorithm

Community structure is a common property of many networks, the division of network nodes into groups within which the network connections are dense, but between which they are sparser. The ability to find and analyze such groups can provide invaluable help in understanding and visualizing the structure of networks (Hochbaum, 1996; Subramanian, 2002; Shamira, 2004; Pan, 2005).

2.3.1. Agglomerative method

The traditional method for detecting community structure in networks is hierarchical clustering. One first calculates a distance D_{ij} for every pair i,j of vertices in the network, then takes all the n vertices in the network, with no edges between them, and adds edges between pairs one by one in order of their distance, starting with the pair with the shortest distance and progressing to the longest. As edges are added, the resulting graph shows a nested set of connected subsets of vertices, which are taken to be the communities. Algorithms of this kind are called agglomerative.

Agglomerative methods have their problems, however. One concern is that they are rigid in that once a merge has been done it cannot be undone. Although there are smaller computational costs with this, it can also cause problems if an erroneous merge is done. Another is their tendency to find only the cores of communities and leave out the periphery. In Fig. 1, there are a number of peripheral nodes whose community membership is obvious to the eye, in most cases they have only a single link to a specific

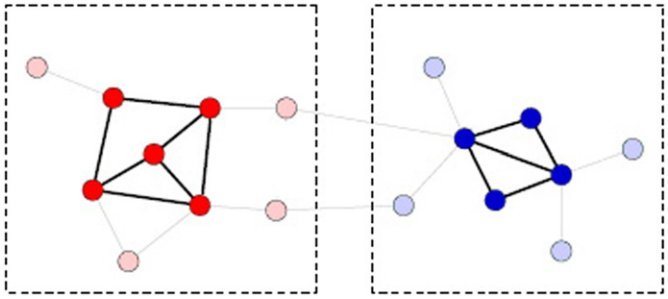

Fig. 1. Agglomerative clustering methods are typically good at discovering the strongly linked cores of communities (bold vertices and edges) but tend to leave out peripheral vertices, even when most of them clearly belong to one community or another.

community, but agglomerative methods often fail to place such nodes correctly.

2.3.2. Divisive algorithm

To sidestep the shortcomings of agglomerative methods, Girvan and Newman (GN) propose an alternative approach to the detection of communities, which could be called divisive algorithm. According to this algorithm, one starts with the whole graph and iteratively cuts the edges, thus dividing the network progressively into smaller and smaller disconnected sub-networks identified as the communities. The crucial point in the divisive algorithm is the selection of the edges to be cut, which has to be those connecting communities and not those within them (Newman, 2001, 2004a, b).

In practice, the selection of the edges to be cut is based on the value of the so-called edge betweenness (Girvan and Newman, 2002; Newman and Girvan, 2004), a generalization of the centrality betweenness introduced by Freeman (1977). For a given graph, the centrality betweenness of edge L is

$$C_B(L) = \sum_{i \neq j} \frac{\sigma_{ij}(L)}{\sigma_{ij}}, \tag{1}$$

where $\boldsymbol{\sigma}_{ij}$ is the total number of shortest paths that go from vertex i to vertex j, and $\boldsymbol{\sigma}_{ij}$ (L) is the number of shortest paths between i and j that go through edge L.

Considering the shortest paths between all pairs of nodes in a network, the betweenness of an edge is the number of these paths running through it. It is clear that, when a graph is made of tightly bound clusters, loosely interconnected, those inter-cluster connections would then have a large betweenness value, through which all shortest paths between nodes in different clusters have to go. By removing iteratively the edges with highest edge-betweenness, the clusters of the graphs are disconnected.

2.3.3. Simulated annealing method

So far, the discussed algorithms define communities operationally as what they find, which is always down to the level of single nodes, independently from the type of graph analyzed. One cannot discriminate between networks that are actually endowed with a community structure and those that are not. As a consequence, in practical applications, one needs additional, nontopological information on the nature of the network to understand whether the identification of a community is reliable. To this problem, a novel geometric measure is introduced (Guimera, 2004, 2005a, b, c, d).

For a given partition of the nodes of a network into communities, the modularity M of this partition is

$$M = \sum_{i=1}^{n} \left[\frac{l_i}{L} - \left(\frac{d_i}{2L} \right)^2 \right], \tag{2}$$

where n is the number of communities, L is the total number of links in the network, l_i is the number of links between nodes in community i, and d_i is the sum of the degrees of the nodes in community i. More precisely, the modularity M estimates the fraction of inward links in a community minus the expectation value of random connections, hence indicates the rationality of the community structure. Therefore, M will equal to zero if nodes are placed at random into communities or if all nodes are in the same cluster.

To find the partition with largest modularity, simulated annealing (Kirkpatrick et al., 1983) is used to obtain the best determination of the community structure of a network by direct maximization of M. As is known, simulated annealing is a stochastic optimization technique that can find 'low cost' configurations without getting trapped in 'high cost' local minima. Here, the cost is $C = -M$. At each temperature, a number of random updates are performed and accepted with probability

$$p = \begin{cases} 1 & \text{if } C_b \leqslant C_a, \\ \exp\left(-\frac{C_a - C_b}{T}\right) & \text{if } C_b > C_a, \end{cases} \tag{3}$$

where C_a is the cost after the update and C_b is the cost before the update; T is computational temperature. When temperature is high, the system can explore configurations of high cost while at low temperature the system only explores low cost regions. Starting at high temperature and slowly decreasing it, the system resides in deep minima, overcoming small cost barriers.

Considering simulated annealing can carry out an exhaustive search and minimize the problem of finding sub-optimal partitions (Chou and Carlacci, 1991), it is not strange that this method exhibits a better performance than that of the standard GN algorithm (Fig. 2). It is noteworthy that, in this method, one does not need to specify *a priori* the number of communities; rather, the number of communities is an outcome of the algorithm.

3. Results and discussion

3.1. Global feature

To learn the global feature of this signaling network, a static geometric analysis is indispensable. Among all the

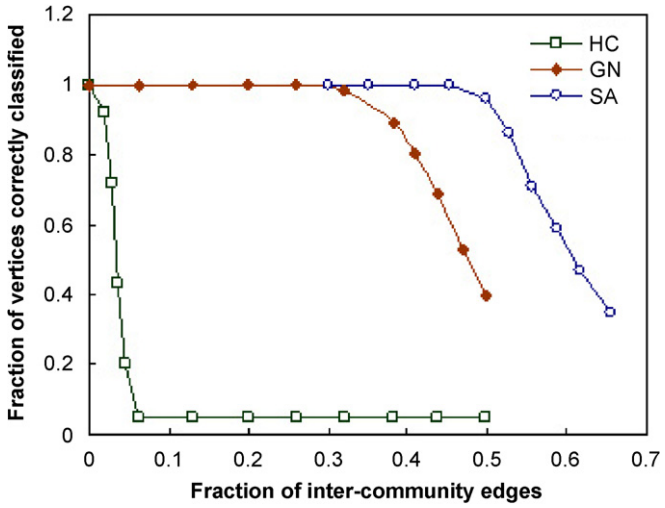

Fig. 2. To test the performance of these methods, random networks with known community structure are built. Each test network comprises 128 nodes divided into 4 communities of 32 nodes, with the average degree of a vertex equal to 16.The x-axis is the fraction of vertices that are classified by the algorithm into their correct communities and the y-axis is the fraction of inter-community edges. Supposing a vertex has 8 inter-community edges, the fraction of inter-community edges of this vertex is 0.5, which also means 50 percent of a vertex's total links is inter-community. HC is the abbreviation of standard hierarchical clustering, GN stands for Girvan–Newman algorithm and SA is for Guimera's simulated annealing method. As it shows, the performance of standard hierarchical clustering (squares) is far inferior to that of GN algorithm (diamonds). Meanwhile, the Guimera's Simulated Annealing method (circles) shows the best performance among the three.

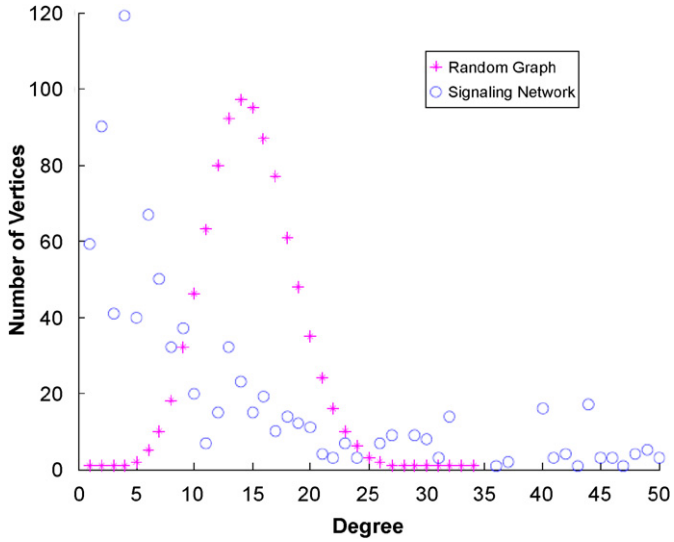

Fig. 3. The circles represent the degree distribution of the signaling network discussed in this study, which has 931 nodes and 6798 edges. The stars represent that of a random network of same size. To avoid random fluctuation, each point is an average over 100 realizations. For example, if there are 4600 nodes with degree equal to 10 in the 100 experiments altogether, the expectation number of nodes with degree equal to 10 is 46.

static geometric quantities and measures, the degree distribution is the most representative. As Fig. 3 shows, it can hardly be explained within the framework of random graph theory why nodes with degree larger than 35 should exist. In those 100 experiments, the node with its degree equal to 34 appeared only once. Meanwhile, According to the random graph model defined by Erdos and Renyi (Bollobás, 1985), the peak value should coincide with the average degree

$$\bar{d} = \frac{1}{n}\sum_{v\in V} \deg(v) = \frac{2m}{n}, \tag{4}$$

where $n = |V|$ and $m = |L|$. By computation, it should be 14. However, it actually moves left to about 4, thus exhibiting a far more different degree distribution from that of random network. Contrastively, in Fig. 4 it can be observed that the degree distribution of the signaling network following a power-law in the form of $P(K) \sim K^{-\gamma}$, with γ approximately equal to 2.2, which suggests that this network is a scale-free network. However, there still exist two discrepancies between the Barabasi–Albert (BA) model and what is observed in this study, which calls for more attention.

First, it can be noticed that nodes with only one link are actually fewer than nodes with 2, 3 or 4 links and the most common nodes are those having 4 neighbors. This is much different from the computer-generated scale-free networks,

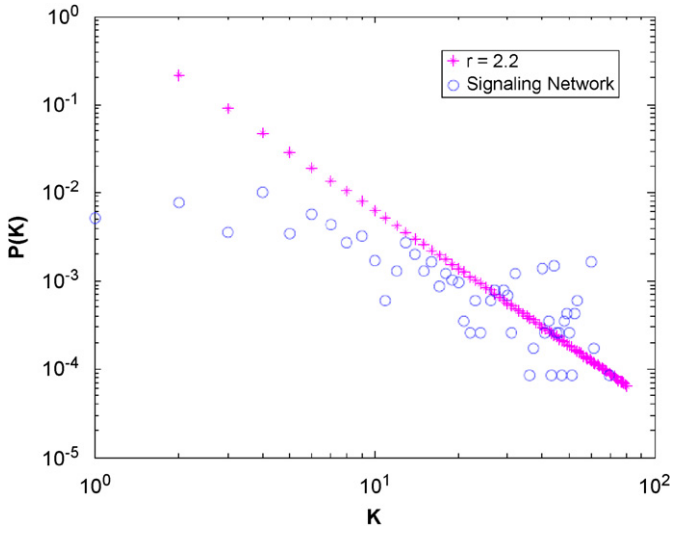

Fig. 4. The circles represent the degree distribution of the signaling network discussed in this study. The stars show a trend-line whose γ equals to 2.2. To improve the quality of the plot, the coordinate is switched from linear scale to logarithmic scale, although not essential to reveal the power-law in the degree distribution.

in which the most common nodes are terminals. The reason of this difference may lie in that the surface area of living cells is finite and the type of terminals or membrane proteins is also limited. The reception, procession and emission of signals are frequently performed in the form of coding (Buck, 2000; Buck and Axel, 1991), not in simple linear form such like receptor A only receives signal a, then transmits to the next node B, till reaches the information processing center, where decisions are made. The practical

situation is usually that a group of similar or identical receptors could receive and emit different signals, according to the different location, number and excitation time of activated receptors. This modulating procedure conduces to promote the efficiency, resilience and robustness of cellular signaling network, compared with simple linear transmission mechanisms.

Second, according to the BA model, the exponent γ should be around 3.0. Our different exponent suggests that the signaling network discussed in this study should not have been formed through linear growth and linear preferential attachment as those in classical BA model (Albert et al., 1999; Albert and Barabasi, 2002; Barabasi and Albert, 1999; Strogatz, 2001). Besides the difference in evolution mechanism, another possible reason for this small exponent could be the general existence of decentralized decision-making in organisms, which dramatically decreases the quantity of hub proteins that have enormous neighbors. To explain this, we construct another 100 scale-free networks comprising 931 nodes and 6798 edges, following the same evolution mechanism defined in classical BA model. In these networks, the hub node with most neighbors has 106 links to other nodes. If we construct enough more scale-free networks of this kind, a star network would theoretically emerge, in which all terminal nodes link to the central hub to have its degree equal to 930. At that time, computing the degree distribution of all these computer-generated networks, we would find the exponent γ to be exactly 3. However, in organisms, this situation will never occur.

Acting as information process center, those 'hub' nodes having much more links than normal often are cores of complexes that may behave as a whole (Smith and Scott, 2002). However, the capacity or information processing capability of these hubs has an upper-limit, which prevents them to hold too many direct links. Meanwhile, this is also the major reason for the general existence of decentralized decision-making in organisms, which could significantly shorten the distance of information transmission and endow it with timely disposal, hence promoting efficiency.

3.2. Condition test

Since this network is proven not to be a random network, it is of significance to investigate the possible existent community structure, which is performed by the Guimera method as it shows the best performance in detecting community structure among the three algorithms. In the condition test of this method, we build a number of random networks with known community structure to explore the scope of its applicability and validity. Each computer-generated network may include two, three or four communities and each community comprises 32 nodes with 16 neighbors on average, which means the average degree of each node is 16. As is shown in Fig. 5, we have defined different boundary for different types of networks.

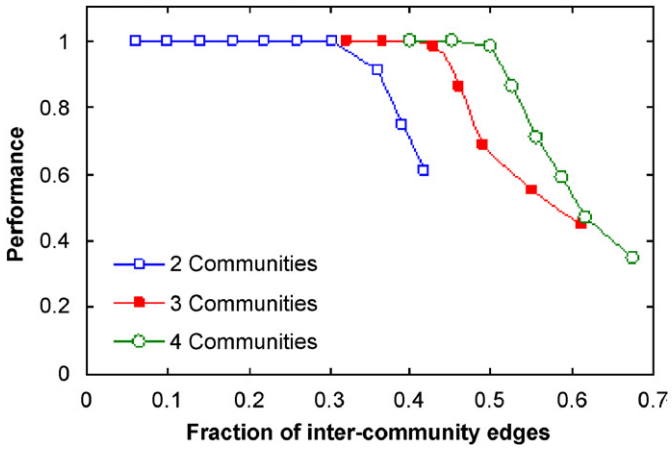

Fig. 5. The squares represent the analysis result of 2-community networks, using the Guimera's simulated annealing method; the solid squares represent that of 3-community networks and the circles for 4-community networks. Note that these curves have different end-points, which corresponds to their different boundary condition described in the text. For 2-community, 3-community and 4-community networks, the end point corresponds to 7, 10 and 11 inter-community links, respectively. To avoid random fluctuation, each point is an average over 100 different computer-generated networks.

For those comprising 2 communities, the boundary is 8 inter-community links per vertex, which means each vertex has as many inter-community links as its intra-community links and no significant community structure should exist in this network then. For those 3-community networks, the boundary is extended to 10 inter-community links per vertex, with 5 links to the other two foreign communities each (on average). We say a specific vertex belongs to a community because it has more intra-community links in this community than in any other one. Therefore, in a 3-community network, a vertex could never belong to a community with its intra-community links less than 6 then.

For those 4-community networks, the boundary is further extended to 12 inter-community links per vertex, with 4 links to the other three foreign communities each, the reason of which is the same as above.

It can be expected that along with the increase of inter-community links, the border of community fades and identification accuracy drops (Fig. 5). Moreover, if the proportion of inter-community links to total links is related to the corresponding community identification accuracy, it can be observed that, for 2-community, 3-community and 4-community networks, the last point with 100% identification accuracy has a proportion of 0.303, 0.366 and 0.437, respectively. It seems that along with the increase of community number, the curves become steeper. From the current result, it might be anticipated that the more communities the incoming network has in nature, the more reliable should the result be, as far as this method is concerned.

Likewise, we compute that proportion of our signaling network, which is about 0.083. Since this method illustrates better performance along with the increase of the number

of natural community, it can be expected that any network having more than 2 communities should be properly handled under such little proportion.

3.3. Community structure

Before using the Guimera method to detect community structure, BFS is applied to investigate the connectivity of this signaling network. It is found that, among the 931 proteins extracted from KEGG, 10 proteins form three connected components and the other 921 form a giant component, which is the subject of the Guimera method accordingly. Result shows that it has 14 communities in total (Fig. 6A). Comprising too much nodes, Fig. 6A is not clear enough for direct observation. By shrinking all nodes in the same community into one node, Fig. 6A is transformed to Fig. 6B. With the size of nodes proportionate to the number of its constituent proteins and the thickness of edges proportionate to the number of inter-community links, it has only 14 nodes and 32 edges, which clearly reveals the relationships between those communities. To learn the detail of the inward-relationships of those communities, we conduct further investigation on the constitution of individual community.

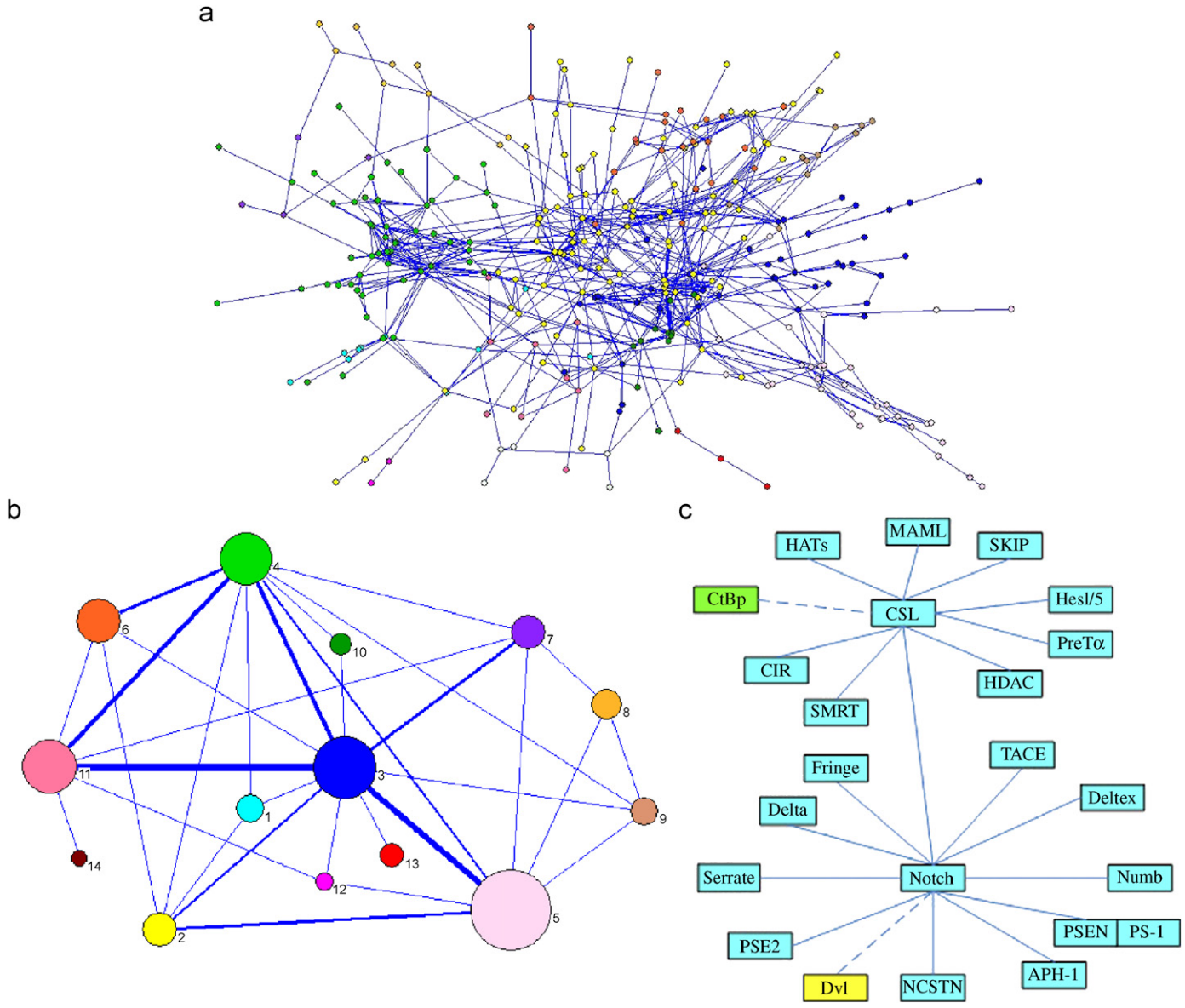

Fig. 6. (A) The Guimera's simulated annealing method divides the giant component of the signaling network into 14 partitions, which has 921 nodes and 6791 links altogether and is shown in different color. Network visualization was done using the Pajek program (Batagelj and Mrvar, 1998). (B) For easy observation, every community is shrunk to one node, with its size proportionate to the number of its constituent proteins and edge thickness proportionate to the number of inter-community links. For convenient reference, these communities are numbered from 1 to 14. (C) Details of the No.1 community. Note that there are 2 proteins that do not belong to community 1.One is *CtBp*, which belongs to community 4 because it has more neighbors in community 4 than in community 1; the other is *Dvl*, which belongs to community 2 and is labeled in the same color as community 2 then.

The chosen community comprises 40 nodes and 139 edges. Among these edges, 118 are between members of this community, and the other 21 lead to other communities. For detailed discussion, all the 40 proteins are listed in the following: 10683 28514 54567 182 3714 3955 4242 4851 4853 4854 4855 6868 55851 23385 8650 9253 113878 1840 151636 196403 23220 5663 5664 51107 83464 55534 84441 9794 2648 22938 343295 11317 3516 9612 3065 3066 3280 388585 171558 9541.

According to KEGG, every ID number corresponds to only one protein. However, homologs commonly exist among these proteins. For instance, 11317 and 3516 are homologs of protein *CSL*; 4851, 4853, 4854 and 4855 are homologs of protein *Notch*, etc. By shrinking homologs, we draw a sketch map of this community (Fig. 6C). Not surprisingly, it can be easily recognized as an intact signaling pathway named Notch, which is found in most animal cells and plays important roles in the fate decision of various types of tissue and cells, including epidermis, nerve, blood, muscle, etc.

Besides capable of finding reliable community, it also helps to recognize crosstalks between signaling pathways. Classified to community 2, *Dvl* belongs to Wnt signaling pathway, having direct interaction with 8 proteins in that community, including *Fzd*, *Stbm*, *Idax*, *Axam*, *Nkd*, etc. From Fig. 6C, it can be easily observed that there is a crosstalk between Wnt and Notch signaling pathway. In practice, lots of crosstalks such alike can be found in Fig. 6A, from which many potential regulation techniques should derive.

Nevertheless, in the process of shrinking, two exceptions occur. While 5663 represents protein *PS-1* and 5664 represents protein *PSEN*, *PS-1* and *PSEN* are homologs. To avoid unnecessary confusion, 5663 and 5664 are shrunk to one node named *PSEN/PS-1*. The other exception happens to the protein *Hats*, which has three homologs named 1387, 2033 and 2648, respectively. Although all the three homologs have relations to protein *CSL*, they follow different patterns. While 2648 having only one neighbor (*CSL*), the other two have much more neighbors besides *CSL*, mainly belong to community 3, which make them apportioned to community 3 accordingly. As a result, the three homologs are shrunk to two *Hats* with one into community 1 and the other into community 3.

However, it should be reminded that, although having provided a good visualization, the shrinking process actually undermines our study and impedes the follow-on simulations, besides the above-mentioned ambiguous *Hats*. To apprehend this, one should think about homologs in serial and parallel processing of signal transduction. Suppose there is a relation like A—B—C, A has three homologs and B has two. In fact, it is a network and could no longer be described with linear models for its topology structure and dynamic property. Moreover, it also might explain the inefficacy of traditional simplified linear representations in modeling real biological networks, to some extent.

4. Conclusions

In 2000, theoretical physicist Stephen Hawking said that the next century would be a century of complexity, who pointed out the major challenge with which theoretical scientists should confront in the 21st century would be handling complex system. That is to say, it is needed to develop and establish a set of theoretical system much different from the past simple system, which becomes more urgent along with the intercross between theoretical science and life science.

Scientists have employed reduction theory for 400 years, which claims research should be conducted through disassemble. One should first remove the research object from its environment, then separate it into parts and explain the whole with isolated parts. Notwithstanding this traditional technique has achieved successive triumphs on colony, individual, organ, cell and molecule level, lots of secrets lie in the emergence, which is the characteristic of the whole. These conspicuous new features derive from the interaction of many simple units that constitute a complex system, which cannot be predicted in advance. For instance, single molecule has no temperature or pressure. Only until lots of molecules assemble, temperature and pressure make sense. Reduction theory cannot interpret these phenomena because the emergence disappears once the wholeness is break. Recently, researchers begin to give more attention to the wholeness of their research objects, especially system dynamics and complex situation thus comes into being (Bhalla and Iyengar, 1999; Jeong et al., 2000; Gavin, 2002; Ravasz, 2002; Wuchty and Stadler, 2003; Rives and Galitski, 2003; Chou and Cai, 2006; Chou et al., 2006).

As the framework of complex networks provides a remarkable tool to describe complex systems of interacting entities, this study explores possible application of it in bioinformatics domain. Firstly, constructs a cellular signaling network of *H. sapiens* from KEGG and conducts regular static geometric analysis on this network; then, compares three community detection algorithms and studies the topology of the signaling network accordingly; finally, discusses the biological implication of community structure and its potential influence in basic research and drug discovery.

Acknowledgments

Thanks to Roger Guimera for stimulating discussions and anonymous reviewers for their helpful comments. This work was supported by the State Key Laboratory of Chemo/Biosensing and Chemometrics, Hunan University.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jtbi.2007.04.007.

References

Aho, A.V., Hopcroft, J.E., 1983. Data Structures and Algorithms, first ed. Addison-Wesley, New York, pp. 198–253.

Albert, R., Barabasi, A.L., 2002. Statistical mechanics of complex networks. Rev. Mod. Phys. 74, 47–97.
Albert, R., Jeong, H., Barabasi, A.L., 1999. Diameter of the world-wide web. Nature 401, 130–131.
Althaus, I.W., Gonzales, A.J., Diebel, M.R., Chou, K.C., 1993a. Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. J. Biol. Chem. 268, 6119–6124.
Althaus, I.W., Gonzales, A.J., Diebel, M.R., Chou, K.C., 1993b. The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. J. Biol. Chem. 268, 14875–14880.
Althaus, I.W., Gonzales, A.J., Diebel, M.R., Chou, K.C., 1993c. Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry 32, 6548–6554.
Barabasi, A.L., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509–512.
Batagelj, V., Mrvar, A., 1998. Pajek: program for large network analysis. Connections 21, 47–57.
Bhalla, U.S., Iyengar, R., 1999. Emergent properties of networks of biological signaling pathways. Science 283, 381–387.
Bollobás, B., 1985. Random Graphs, second ed. Academic Press, New York, pp. 34–78.
Buck, L.B., 2000. The molecular architecture of odor and pheromone sensing in mammals. Cell 100, 611–618.
Buck, L.B., Axel, R., 1991. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell 65, 175–187.
Chou, K.C., 1981. Two new schematic rules for rate laws of enzyme-catalyzed reactions. J. Theor. Biol. 89, 581–592.
Chou, K.C., 1983. Advances in graphical methods of enzyme kinetics. Biophys. Chem. 17, 51–55.
Chou, K.C., 1989. Graphical rules in steady and non-steady enzyme kinetics. J. Biol. Chem. 264, 12074–12079.
Chou, K.C., 1990. Review: applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. Biophys. Chem. 35, 1–24.
Chou, K.C., 1993. Graphic rule for non-steady-state enzyme kinetics and protein folding kinetics. J. Math. Chem. 12, 97–108.
Chou, K.C., Cai, Y.D., 2006. Predicting protein-protein interactions from sequences in a hybridization space. J. Proteome Res. 5, 316–322.
Chou, K.C., Carlacci, L., 1991. Simulated annealing approach to the study of protein structures. Protein Eng. Des. Sel. 4, 661–667.
Chou, K.C., Forsen, S., 1980. Graphical rules for enzyme-catalyzed rate laws. Biochem. J. 187, 829–835.
Chou, K.C., Liu, W.M., 1981. Graphical rules for non-steady state enzyme kinetics. J. Theor. Biol. 91, 637–654.
Chou, K.C., Kezdy, F.J., Reusser, F., 1994. Review: steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. Anal. Biochem. 221, 217–230.
Chou, K.C., Cai, Y.D., Zhong, W.Z., 2006. Predicting networking couples for metabolic pathways of Arabidopsis. EXCLI J. 5, 55–65.
Fenn, J.B., Mann, M., Meng, C.K., Whitehouse, C.M., 1989. Electrospray ionization for mass spectrometry of large biomolecules. Science 246, 64–71.
Freeman, L., 1977. A set of measures of centrality based on betweenness. Sociometry 40, 35–41.
Gavin, A.C., 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415, 141–147.
Girvan, M., Newman, M.E.J., 2002. Community structure in social and biological networks. Proc. Natl Acad. Sci. 99, 7821–7826.
Guimera, R., 2004. Modularity from fluctuations in random graphs and complex networks. Phys. Rev. E 70, 025101.
Guimera, R., 2005a. Functional cartography of complex metabolic networks. Nature 433, 895–900.
Guimera, R., 2005b. Team assembly mechanisms determine collaboration network structure and team performance. Science 308, 697–702.
Guimera, R., 2005c. The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles. Proc. Natl Acad. Sci. 102, 7794–7799.
Guimera, R., 2005d. Cartography of complex networks: modules and universal roles. J. Stat. Mech., P02001.
Hochbaum, D.S., 1996. Approximation Algorithms for NP-Hard Problems, first ed. PWS Publishing, Boston.
Jeong, H., Tombor, B., Albert, R., 2000. The large-scale organization of metabolic networks. Nature 407, 651–654.
Kanehisa, M., Goto, S., 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27–30.
Kirkpatrick, S., Gelat, J.C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. Science 220, 671–680.
Kitano, H., 2002. Systems biology: a brief overview. Science 295, 1662–1664.
Kuzmic, P., Heath, T.D., 1992. Mixtures of tight-binding enzyme inhibitors. Kinetic analysis by a recursive rate equation. Anal. Biochem. 200, 68–73.
Lee, W.C., 2004. Applications of affinity chromatography in proteomics. Anal. Biochem. 324, 1–10.
Lin, S.X., Neet, K.E., 1990. Demonstration of a slow conformational change in liver glucokinase by fluorescence spectroscopy. J. Biol. Chem. 265, 9670–9675.
Maslov, S., Sneppen, K., 2002. Specificity and stability in topology of proteins networks. Science 296, 910–913.
Newman, M., 2001. Clustering and preferential attachment in growing networks. Phys. Rev. E 64, 025102.
Newman, M., 2004a. Detecting community structure in networks. Eur. Phys. J. B 38, 321–330.
Newman, M., 2004b. Fast algorithm for detecting community structure in networks. Phys. Rev. E 69, 066133.
Newman, M., Girvan, M., 2004. Finding and evaluating community structure in networks. Phys. Rev. E 69, 026113.
Pan, J.J., 2005. Path partition for graphs with special blocks. Discrete Appl. Math. 145, 429–436.
Ravasz, E., 2002. Hierarchical organization of modularity in metabolic networks. Science 297, 1551–1555.
Rives, A.W., Galitski, T., 2003. Modular organization of cellular networks. Proc. Natl Acad. Sci. 100, 1128–1133.
Shamira, R., 2004. Cluster graph modification problems. Discrete Appl. Math. 144, 173–182.
Smith, F.D.I., Scott, J.D., 2002. Signaling complexes: junctions on the intracellular information super highway. Curr. Biol. 12, 32–40.
Strogatz, S.H., 2001. Exploring complex networks. Nature 410, 268.
Subramanian, C.R., 2002. General partitioning on random graphs. J. Algorithm. 42, 153–172.
Vidal, M., Legrain, P., 1999. Yeast forward and reverse 'n'-hybrid systems. Nucleic Acids Res. 27, 919–929.
Wang, M., Yao, J., Huang, Z., Chou, K.C., 2005. A new nucleotide-composition based fingerprint of SARS-CoV with visualization analysis. Med. Chem. 1, 39–47.
Wuchty, S., Stadler, P.F., 2003. Centers of complex networks. J. Theor. Biol. 223, 45–53.
Xiao, X., Shao, S., Ding, Y., Chou, K.C., 2005a. Using cellular automata to generate image representation for biological sequences. Amino Acids 28, 29–35.
Xiao, X., Shao, S., Ding, Y., Chou, K.C., 2005b. An application of gene comparative image for predicting the effect on replication ratio by HBV virus gene missense mutation. J. Theor. Biol. 235, 555–565.
Xiao, X., Shao, S., Ding, Y., Chou, K.C., 2006a. Using cellular automata images and pseudo amino acid composition to predict protein sub-cellular location. Amino Acids 30, 49–54.
Xiao, X., Shao, S., Chou, K.C., 2006b. A probability cellular automaton model for hepatitis B viral infections. Biochem. Biophys. Res. Commun. 342, 605–610.
Zhang, C.T., Chou, K.C., 1996. An analysis of base frequencies in the anti-sense strands corresponding to the 180 human protein coding sequences. Amino Acids 10, 253–262.
Zhou, G.P., Deng, M.H., 1984. An extension of Chou's graphical rules for deriving enzyme kinetic equations to system involving parallel reaction pathways. Biochem. J. 222, 169–176.