Constructing Signal Transduction Networks Using Multiple Signaling Feature Data

Thanh-Phuong Nguyen[1], Kenji Satou[2], Tu-Bao Ho[1] and Katsuhiko Takabayashi[3]

[1] Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan.
[2] Kanazawa University
Kakuma-machi, Kanazawa 920-1192, Ishikawa, Japan.
[3] Chiba-University
Inohana 1-8-1,Chiba, 260-8677, Japan
{phuong,bao}@jaist.ac.jp, ken@t.kanazawa-u.ac.jp, takaba@ho.chiba-u.ac.jp

Abstract. Signal transduction networks (STN), as *complex biological systems*, are crucial for inter- and intra-cellular signaling. The essence of STN is underlain in some signaling features scattered in various data sources, and the biological components overlapping among STN. The integration of those signaling features presents a challenge. In this paper, we introduce an effective method that combine various signaling data features, and detect out the components overlapping among STN. This work has two main contributions. Many structured data of signaling features, i.e., protein-protein interaction networks, domain-domain interactions, signaling domains, and protein functions, have been extracted and combined to comprehensively construct STN. Those heterogenous data are known to be significant and useful in STN construction. The uncertain components overlapping among STN have already been found, using soft-clustering. We did the experiments with five biological processes in the Reactome database. Those processes were reconstructed with small errors. The experiment results were promising to discover new STN in system biology.

Keywords: signal transduction network, structured data, protein-protein interaction network, signaling features, soft-clustering.

1 Introduction

Signal transduction networks are the primary means by which eukaryotic cells respond to external signals from their environment as well as coordinate complex cellular changes [1]. STN are important in the correct functioning of the cell and producing appropriate outcomes, such as cell division, apoptosis, or differentiation in response to a variety of biological signals. The dominant molecules in STN are proteins, which have several signaling features. To transmit biological signals in cells, those proteins need interactions as signaling channels among them. Therefore, a STN can be considered as a complex protein interactions.

Because of the biologically significant roles of STN in cells, both biologists and bioinformaticians have taken much interest in finding out molecular components and/or the relations among these molecular components in STN. Experimental methods have been effective in generating detailed descriptions of specific linear signaling pathways; however our knowledge of complex signaling networks and their interactions remains incomplete [2]. Recently, an enormous amount of high-throughput protein-protein interaction (PPI) data has been generated [6], [8]. The available PPI data is one of important signaling feature data, and invaluable to study STN. There is a great need for developing computational methods to take advantage of information-rich protein interaction data to study complex signaling mechanisms inside STN.

Constructing STN based on PPI is an area of much ongoing research. Gomez *et al.* modeled STN in terms of domains in upstream and downstream protein interactions, using Markov chain Monte Carlo method [5]. Steffen *et al.* developed a computational method for generating static utilized PPI maps produced from large-scale two-hybrid screens and expression profiles from DNA micro-arrays in STN construction [11]. Liu *et al.* applied a score function that integrated PPI data and micro-array gene expression data for predicting the order of signaling pathway components [8]. Concerning protein modification time-course data, Allen *et al.* applied a method of computational algebra to modeling of signaling networks [1]. Fukuda *et al.* represented the model of signal transduction pathways based on a compound graph structure [4]. One of recent work proposed some cost functions to search for the optimal subnetworks (as STN) from PPI [12].

Although the previous work achieved many results, there are some biological characteristics of STN, which did not take much into account. Firstly, it is known that the deep level underlying the PPI to transmit signals are functional domains, the so-called signaling domains, and their interactions [3], [10]. The data regarding those significant signaling features are structured, complexly relational, and sparse in various data sources. In order to construct STN effectively, those data is needed to be appropriately integrated. Second, STN indeed have many overlapping components, including proteins and their interactions [9]. This work aims to solve those two intricate problems of STN to better construct STN from PPI networks. To this end, we developed an effective computational method to construct STN that (1) integrated multiple signaling features of STN from heterogenous sources, i.e., protein-protein interactions, signaling domains, domain-domain interactions, and protein functions; and (2) detected the overlapping components using soft-clustering. Additionally, in previous work an clustered object was often an individual protein, but our method handled a clustered object as a functional or physical protein interaction (as the signaling means).

We evaluated the performance of the proposed method, using human protein interaction network extracted from the Reactome database. Five complex biological processes in the Reatome database were tested by our method. The experimental results demonstrated that our method could reconstruct those five

processes with small errors and detect nearly the exact number of overlapping components. To the best of our knowledge, this work is the first one that computationally solves the STN problem for *Homo Sapiens*. The preliminary results open a prospect to study other problems related to complex biological systems in *Homo Sapiens*.

The remainder of the paper is organized as follows. In Section 2, we present our proposed method to construct STN using soft-clustering and multiple signaling feature data. The evaluation is given in Section 3. Finally, Section 4 give some concluding remarks.

2 Method

The proposed method has two main tasks. The first one is to extract and preprocess signaling feature data from various data sources. Those relational data in heterogenous types were then weighted and normalized by some proposed functions. The second is to combine extracted data and cluster protein-protein interactions into STN using soft-clustering. Two subsections, 2.1 and 2.2, describe two mentioned tasks in succession.

2.1 Extracting signaling feature data from multiple data sources

STNs have a two-level signaling machinery. The first level of complexity in cellular signaling derives from the large number of molecules and multiple types of interactions between them. The second level of complexity of signaling biochemistry is apparent from the fact that signaling proteins often contain multiple functional domains, thus enabling each to interact with numerous downstream targets [3]. Based on those facts of STN, we extracted a lot of signaling feature data as follows.

1. Protein-protein interactions (PPI): the upper level consists of the components as interfaces to transmit signals. PPI data were extracted from the Reactome database[4].
2. Domain-domain interactions (DDI): the deep level consists of interactions of protein domains, which are the basic elements in PPI. DDI data were extracted from the iPfam database[5].
3. Signaling domain-domain interactions: the deeper functional level consists of signaling domains (specific functional domains) that act as key factors to transduce signals inside STN. Signaling DDI data were extracted from SMART database[6] and referred in [10].

Protein function were also extracted from Uniprot database[7] as keywords tagged to proteins).

[4] www.reactome.org/
[5] www.sanger.ac.uk/Software/Pfam/iPfam/
[6] smart.embl-heidelberg.de/
[7] www.uniprot.org/

The raw data in different databases are stored in different types, e.g., the numerical type for number of PPI, interaction generality, number of signaling DDI or categorical type for protein functions. Those data have the complex relations. For example, one protein may have many PPI and one PPI may have many DDI. Interacting partners of one DDI may be a signaling domain or not. To exploit those relations, after extracting data from multi-data sources, we weighted and normalized by some proposed weight functions. Table 1 shows the functions and their corresponding explanations.

Table 1. Weight functions for the extracted signaling features.

Weight functions	Notations and explanation
$w_{ppi}(p_{ij}) = \frac{g_{ij}^2}{n_i * n_j}$	g_{ij}: Interaction generality, the number of proteins that interact with just two interacting partners, p_i and p_j. n_i: The number of protein-protein interactions of the protein p_i.
$w_{Sddi}(p_{ij}) = \frac{n_{Sddi}+1}{n_{ddi}+1}$	n_{Sddi}: The number of signaling domain-domain interactions shared between two interacting proteins. n_{Sddi}: The number of domain-domain interactions shared between two interacting proteins.
$w_{func}(p_{ij}) = \frac{k_{ij}^2}{k_i * k_j}$	k_{ij}: The number of sharing keywords k_{ij} of two interacting partners, p_i and p_j. k_i: The number of keywords of the protein p_i.

- PPI weight function (w_{ppi}): The topological relation of proteins in a PPI network was extracted in terms of the numbers of interactions of each partner and the interaction generality.
- Signaling DDI weight function (w_{Sddi}): The relation between a PPI and their DDI was exploited in terms the numbers of DDI and the numbers of signaling DDI, which mediate the PPI.
- Keyword weight function (w_{func}): The relation of a PPI and protein functions was taken into account in terms of the keywords tagged in each partner and the keywords shared between them.

2.2 Combining signaling feature data to construct STN using soft-clustering

After weighting signaling features, it is necessary to combine them in a unified computational scheme to take full advantage of those data. We integrated these data and represented them in forms of feature vectors. Every PPI has its own feature vector, which has three elements corresponding to three features, v_{ij} = $\{w_{ppi}, w_{Sddi}, w_{func}\}$. Subsequently, we employed a soft-clustering algorithm to cluster the PPI based on their features vectors. Soft-clustering can construct STN and detect the overlapping components that can not be found by traditional

hard-clustering. Note that we used the Mfuzz software package [7] to implement fuzzy c-means (FCM) clustering algorithm in our experiments. Fuzzy c-means (FCM) clustering algorithm is one of popular soft-clustering algorithms.

Figure 1 summarizes the key steps of our method that does (1) extracting and weighting signaling features and (2) integrating the extracted features and cluster PPI into STN, using a soft-clustering algorithm. Given a large protein-protein interaction network $\mathfrak{N}$, the outputs are STN, which are considered as the subgraphs of edges (as protein interactions) and nodes (as proteins). Step 1 is to obtain the binary interactions from the protein-protein interaction network $\mathfrak{N}$. Steps 2 to 5 are to carry out the first task, extracting and then weighing signaling feature data by the functions shown in Table 1. These steps are done for all binary PPI to exploit the relations between PPI and signaling features. Step 6 is to perform the second task, combining weighted feature data, representing them in forms of feature vectors $v_{ij} = \{w_{ppi}, w_{Sddi}, w_{func}\}$. Step 7 is to soft-clustering PPI with their feature vectors into STN $\mathcal{S}$. Finally, STN $\mathcal{S}$ are returned in Step 8.

Figure 1 The proposed method to construct STN from PPI networks using soft-clustering and multiple signaling feature data.

Input:
Protein-protein network $\mathfrak{N}$.
Set of signaling features $\mathcal{F} \subset \{f_{ppi}, f_{Sddi}, f_{func}\}$.

Output:
Set of signal transduction networks $\mathcal{S}$.

1: Extract binary interactions $\{p_{ij}\}$ from the protein-protein network $\mathfrak{N}$. $\mathcal{P} := \{p_{ij}\}$.
2: For each interaction $p_{ij} \subset \mathcal{P}$
3: Extract and formalize data for the feature PPI f_{ppi}
4: Extract and formalize data for the feature signaling DDI f_{Sddi}
5: Extract and formalize data for the feature function f_{func}
6: Combine and represent the all features in the feature vectors $v_{ij} = \{f_{ppi}, f_{Sddi}, f_{func}\}$.
7: Apply a soft-clustering algorithm with the set of feature vectors $\{v_{ij}\}$ to cluster interactions p_{ij} into signal transduction networks $\mathcal{S}$.
8: **return** $\mathcal{S}$.

3 Evaluation

To evaluate the performance of the method, we considered a complex PPI network to detect STN out of other biological processes. The tested PPI network does contain not only signaling processes, but also other biological processes functioned inside the PPI network. The mixture of the diverse processes in a PPI network is popular in cells. The experimental results are needed to reflect this complicated phenomena. Namely, the signaling processes are reconstructed with small error and the overlapping components are detected out. We extracted five heterogeneous processes from the Reactome database and the results demonstrated that our method effectively constructed STN from the PPI network with their overlapping components.

3.1 Experiments

The Reactome database consists of 68 *Homo sapiens* biological processes of 2,461 proteins. There are 6,188 protein interactions, and 6,162 interactions participating in biological processes. 636 proteins partakes in at least 2 different processes, 400 proteins in at least 3 processes, 119 proteins in 5 processes. Therefore, we can see that there exists a lot of proteins and their interactions overlapping among biological processes.

In our experiments, we extracted a group of five biological processes and two of them are signaling processes. Table 2 shows some information related to those five processes. In total, this group consists of 145 distinct interactions of 140 distinct proteins. There are many interactions and proteins overlapping among the processes.

Table 2. Five tested biological processes and some related information.

Reactome annotation	Description	#Proteins	#Interactions
REACT_1069	Post-translational protein modification	40	23
REACT_1892	Elongation arrest and recovery	31	68
REACT_498	Signaling by Insulin receptor	39	44
REACT_769	Pausing and recovery of elongation	31	68
REACT_9417	Signaling by EGFR	40	25

The proteins partaking in five processes were extracted and looked for their interactions in the Reactome interactions set. We strictly extracted only the interactions that have both interacting partners joining in processes because the method considers the proteins but more importantly their interactions. The extracted interactions and their signaling features were then input in the soft-clustering algorithm.

In this paper, we applied the Mfuzz software package to run fuzzy c-means (FCM) clustering algorithm. It is based on the iterative optimization of an objective function to minimize the variation of objects within clusters [7]. As a result, fuzzy c-means produces gradual membership values μ_{ij} of an interaction i between 0 and 1 indicating the degree of membership of this interaction for cluster j. This strongly contrasts with hard-clustering, e.g., the commonly used k-means clustering that generates only membership values μ_{ij} of either 0 or 1. Mfuzz is constructed as an R package implementing soft clustering tools. The additional package Mfuzzgui provides a convenient TclTk-based graphical user interface.

Concerning the parameters of Mfuzz, the number of clusters was 5 (because we are considering 5 processes) and the so-called fuzzification parameter μ_{ij} was chosen 0.035 (because the testing data is not noisy).

3.2 Results and Discussion

Actually, two processes REACT_1892 and REACT_498 share the same set of proteins and the same interactions as well. Also, two signaling processes, REACT_9417 and REACT_498 have 16 common interactions. Nevertheless, the

process *'post-translational protein modification'* is separated from the other processes. The complex STN were effectively constructed and the overlaps among STN were detected.

We set a threshold ε as 0.1. The threshold ε means that if the membership of an interaction i with a cluster j $\mu_{ij} \geq \varepsilon$ (0.1), this interaction highly correlates with the cluster j and it will be clustered to cluster j. Five clusters were produced and then matched with 5 processes. The results are shown in Table 3.

Table 3 shows that we can construct signal transduction networks with the small error and can detect the nearly exact number of overlapping interactions. The combination of signaling feature data distinguished signaling processes from other biological processes, and soft-clustering detected the overlapping components among them. When we checked the overlapping interactions among the clusters, there were exact 16 interactions that are shared in two signaling processes *'signaling by Insulin receptor' and 'signaling by EGFR'*. Also, the same interaction set of the process 'elongation arrest recovery' and the process '*pausing and recovery of elongation'* are found in their clusters. In fact, REACT_1069 does not overlap other processes but the results return three overlapping interactions, i.e., one with REACT_1892 and REACT_769 and two with REACT_498 and REACT_9417.

Table 3. Clustered results for five tested biological processes.

Process	True positive[1]	False negative[2]	False positive[3]	#Overlap_Int[4]
REACT_1069	0.565	0.174	0.435	3/0
REACT_1892	1.000	0.103	0.000	70/68
REACT_498	0.818	0.068	0.182	17/16
REACT_769	1.000	0.103	0.000	70/68
REACT_9417	0.960	0.120	0.040	17/16

1 True positive: the number of true interactions clustered/the number of interactions of the fact process.
2 False negative: the number of interactions missed in fact processes/the number of interactions of the fact process.
3 False positive : the number of false interactions clustered/the number of interactions of the fact process.
4 #Overlap_Int: the number of overlapping interactions among the clusters/the number of overlapping interactions among the fact processes.

We analyzed interaction (P00734, P00734) shared among REACT_1069, REACT_498 and REACT_9417. Protein P00734 (Prothrombin) functions in blood homeostasis, inflammation and wound healing and joins in biological process as cell surface receptor linked signal transduction (have GO term GO:0007166). In the Reactome database, interaction(P00734, P00734) does not happen in the processes REACT_498 and REACT_9417, however according to the function of P00734, it probably partakes in one or two signaling processes REACT_498 and REACT_9417.

Although, the experiment carried out a case study of five biological processes; the proposed method is flexible to be applied to the larger scale of human in-

teraction networks. In the intricate relations of many biological processes, the proposed method can well construct signal transduction networks.

4 Conclusion

In this paper, we have presented a soft-clustering method to construct signal transduction networks from protein-protein networks. Many structured data of signaling features were extracted, integrated, using soft-clustering. The experimental results demonstrated that our proposed method could construct STN effectively. The overlapping components among STN were well detected. In future work, we would like to further investigate signaling features of proteins and protein interactions. The experiments with various parameters and other soft-clustering algorithms (not only FCM algorithm in Mfuzz) should be tested. We will consider some other methods in relational learning and statistical learning to improve the work. It is also promising to discover the novel signal transduction networks from large interaction networks.

References

1. E.E. Allen, J.S. Fetrow, L. W. Daniel, S.J. Thomas, and D.J. John. Algebraic dependency models of protein signal transduction networks from time-series data. *Journal of Theoretical Biology*, 238(2):317–330, 2006.
2. Anand R. Asthagiri and Douglas A. Lauffenburger. Bioengineering models of cell signaling. *Annual Review of Biomedical Engineering*, 2(1):31–53, 2000.
3. Narat J. Eungdamrong and Ravi Iyenga. Modeling cell signaling networks. *Biology of the Cell*, 96(5):355–362, 2004.
4. K. Fukuda and T. Takagi. Knowledge representation of signal transduction pathways. *Bioinformatics*, 17(9):829–837, 2001.
5. S. M. Gomez, S. Lo, and A. Rzhetsky. Probabilistic Prediction of Unknown Metabolic and Signal-Transduction Networks. *Genetics*, 159(3):1291–1298, 2001.
6. T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. In *Proc. Natl. Acad. Sci. USA 98*, pages 4569–4574, 2001.
7. L. Kumar and M.E. Futschik. Mfuzz: A software package for soft clustering of microarray data. *Bioinformation*, 2(1):5–7, 2007.
8. Y. Liu and H. Zhao. A computational approach for ordering signal transduction pathway components from genomics and proteomics data. *BMC Bioinformatics*, 5(158), 2004.
9. Susana R. Neves and Ravi Iyengar. Modeling Signaling Networks. *Sci. STKE*, 2005(281):tw157–, 2005.
10. T. Pawson, M. Raina, and N. Nash. Interaction domains: from simple binding events to complex cellular behavior. *FEBS Letters*, 513(1):2–10, 2002.
11. M. Steffen, A. Petti, J. Aach, P. D'haeseleer, and G. Church. Automated modelling of signal transduction networks. *BMC Bioinformatics*, 3(34), 2002.
12. X.M. Zhao, R.S. Wang, L. Chen, and K. Aihara. Automatic modeling of signal pathways from protein-protein interaction networks. In *The Sixth Asia Pacific Bioinformatics Conference*, pages 287–296, 2008.