Construction of human tissue-specific phosphorylation networks with protein expression data

Yin-Ying Wang†
Institute of Systems Biology,
Shanghai University,
Shanghai 200444, China
Email: yingxiao8958@gmail.com

Chenglei Sun†
Institute of Systems Biology,
Shanghai University,
Shanghai 200444, China
Email: lurlin@shu.edu.cn

Luonan Chen
Key Laboratory of Systems Biology,
Shanghai Institutes for Biological Sciences,
Chinese Academy of Sciences,
Shanghai 200031, China
Email: lnchen@sibs.as.cn

Xing-Ming Zhao*
School of Electronics and Information Engineering,
Tongji University,
Shanghai 201804, China
Email: xm_zhao@tongji.edu.cn

***Abstract*—Phosphorylation is a post-translational modification process mediated by kinases through the addition of a covalently bound phosphate group, which plays important roles in a wide range of cellular progresses, such as signaling cascades and development. Over the past years, despite many phosphorylation sites have been determined with mass spectrometry techniques, it is not clear which kinase phosphorylates which proteins. Under the circumstance, we propose a new probabilistic model to identify the substrates phosphorylated by certain kinases. Furthermore, we construct three tissue-specific phosphorylation networks based on protein expression data. Investigating the constructed tissue-specific networks, we find they are functionally consistent with the corresponding tissues, implying the effectiveness and biological significance of our proposed approach.**

I. Introduction

Proteins, the basic functional units in biology, are controlled by various kinds of post-translational modifications (PTMs) [1], among which phosphorylation is one of the most common PTMs. Phosphorylation is mediated by kinases through the addition of a phosphate group and is estimated to affect one-third proteins in eukaryotic cells. Phosphorylation plays pivotal roles in a wide range of cellular processes [2], such as signal transduction and differentiation, and acts as switches in many biological functions. For example, if some kinases are inhibited, protein phosphorylation will be switched off, leading to the abnormal regulation of cell cycle [3]. Furthermore, protein phosphorylation is found to be related to many diseases, including cancer and diabetes.

In general, different kinases modify distinct protein substrates by recognizing specific phosphorylation sites. The interactions between kinases and their corresponding substrates could provide insights into the biological processes in which the phosphorylation is involved. Unfortunately, only few such interactions are known right now. Over the past decade, tens of thousands of phosphorylation sites have been identified with experimental techniques, e.g. mass spectrometry [4]. For example, there are about 42574 serine, threonine and tyrosine phosphorylation sites are reported to be determined for human according to Phospho.ELM database [5]. Despite the large amount of phosphorylation sites available, unfortunately, it is not clear which sites are recognized by which kinases and which proteins interact with which kinases [6]. Therefore, some computational approaches have been developed to predict the interactions between kinases and proteins due to the labor-intensive and time consuming experiments. For example, Hjerrild *et al.* developed a new approach to identify phosphorylation sites based on neural network [7], and Obenauer *et al.* predicted cell signaling interactions based on sequence motifs [8]. More recently, Newman *et al.* successfully constructed a phosphorylation network consists of 230 kinases and 652 substrates based on protein microarray data [9].

Despite the above efforts to predict the interactions between kinases and proteins, most of the kinase-specific phosphorylation are not known. In this work, we present a novel model to predict kinase-specific phosphorylations based on protein-protein interaction and protein expression data. In particular, we assume that a kinase binds to one specific sequence motif within a protein to modify the activity of the protein, where the phosphorylation sites are located in the sequence motifs. Furthermore, we construct three tissue-specific phosphorylation networks for human based on the tissue-specific protein expression data. Investigating the constructed tissue-specific networks, we find they are functionally consistent with the corresponding tissues, implying the effectiveness and biological significance of our proposed approach.

II. Methods and materials

A. Data sources

In this work, all the phosphorylated protein expression data were measured with mass spectrometry for three tissues within three human samples, including cerebrum (CB), prefrontal cortex (PFC) and liver (LV). In total, there are 4250 proteins

* Corresponding author.
† These authors contributed equally to this work.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

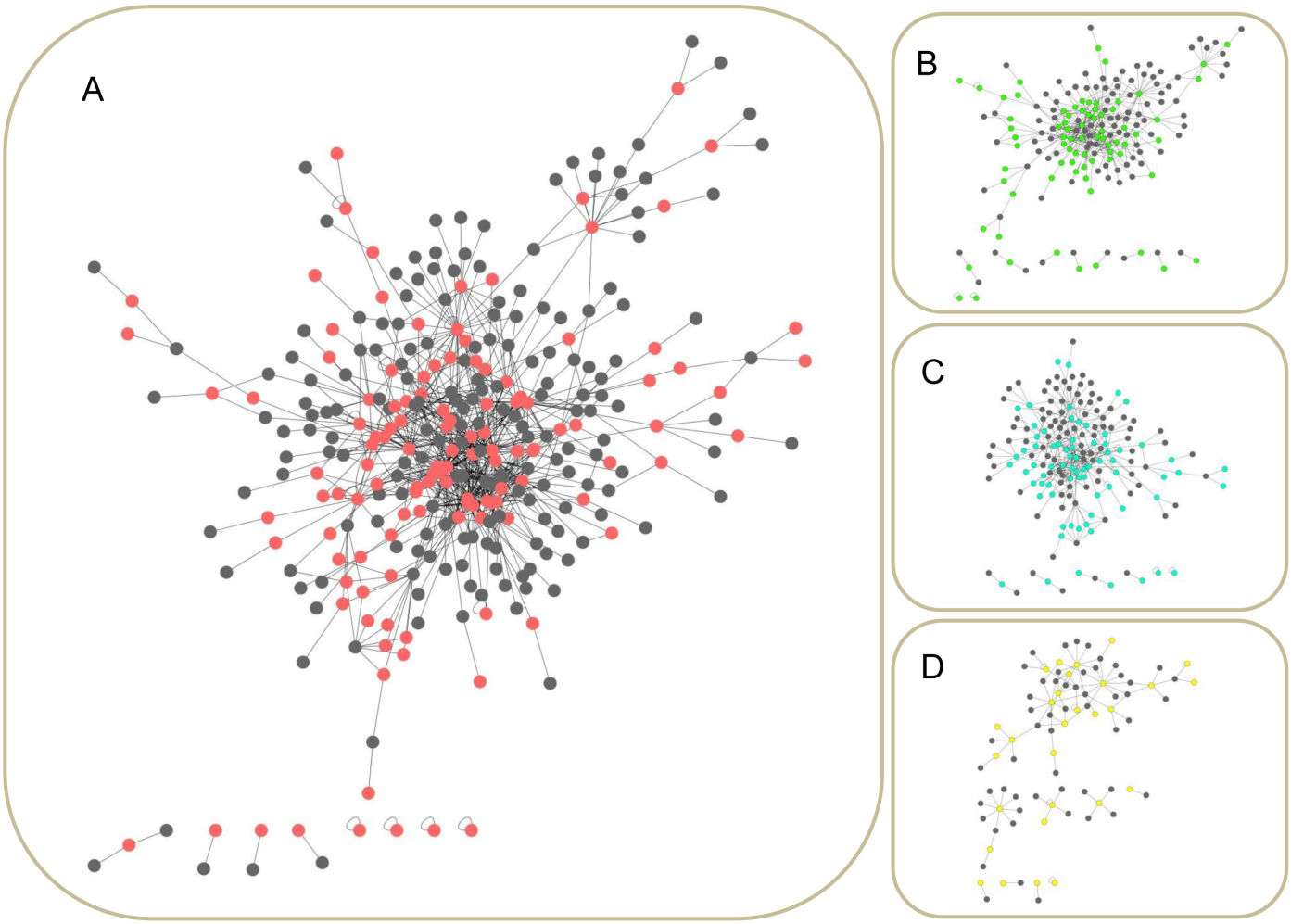

Fig. 1. A. The whole phosphorylation network which contains 711 kinase-substrate interactions in total, where the red nodes represent protein kinases and gray nodes denote protein substrates, respectively. B. The phosphorylation network of Cerebrum, where green nodes represent kinases. C. The phosphorylation network of Prefrontal cortex, where blue nodes denote kinases. D. The phosphorylation network of Liver, where yellow nodes represent kinases.

TABLE I. HUMAN PROTEIN INTERACTIONS FROM DIFFERENT DATABASES.

Database	Interactions	$P_{4250}\_P_{4250}$	$P_{4250}\_K_{234}$
HPRD	38788	7705	2114
BioGRID	40335	7796	1042
IntNetDB	14837	5616	363
STRING	112780	19596	1836

$P_{4250}\_P_{4250}$ denotes the number of interactions among the 4250 expressed proteins. $P_{4250}\_K_{234}$ denotes the number of interactions between the 4250 expressed proteins and the 234 kinases.

that are detected to be expressed, of which 1280 proteins are found to be phosphorylated.

All the kinases were collected from the Human Protein Reference Database (HPRD) [10] and the UniProt database [11]. After mapping to the protein expression data, we kept only those kinases that can be expressed in at least two samples. As a result, we obtained 234 kinases. Furthermore, we scanned all the expressed proteins to see whether they contain any phosphorylation motifs annotated in HPRD. Finally, 124 motifs were found to be contained in at least one phosphorylated protein.

Since human protein-protein interaction data is not complete, we integrated interactions from different databases, including HPRD [10], BioGRID [12], IntNetDB [13], and STRING [14]. The detailed information about the interactions from distinct data sources can be found in TABLE I. The union of all the interactions was used for further analysis. In particular, we only considered physical interactions among proteins that were found to be expressed in our datasets.

B. Prediction of kinase-specific interactions

We assumed that a kinase interacts with its substrates by binding to certain specific sequence motifs within the substrates. In other words, the kinase-protein interactions are accomplished with the kinase-motif interactions. Furthermore, we assumed that to phosphorylate a protein, one kinase has to have physical interaction with the protein. With the motif

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

TABLE II. TISSUE-SPECIFIC INTERACTIONS.

Tissues	Interactions	*P-value* of significance		
		CB	PFC	LV
CB	PRKAR2A_CSK	**0.0350**	0.1092	0.5715
	MAPK3_MAP2K1	**0.0000**	0.2339	0.6667
	STAT3_PTK2B	**0.0283**	0.2601	0.2123
	STAT3_CSK	**0.0000**	0.6667	0.2123
	NCK2_INSR	**0.0000**	0.3333	0.6667
	JAK1_INSR	**0.0000**	0.3333	0.3333
PFC	PRKAR2A_PRKACB	0.3918	**0.0452**	0.9048
	SPTBN1_PRKACA	0.8823	**0.0470**	0.6309
	STAT3_MAPK3	0.3333	**0.0000**	0.8790
LV	SDPR_PRKCA	0.4668	0.2601	**0.0000**
	YWHAQ_FRAP1	0.2731	0.5922	**0.0298**
	ARHGEF12_ROCK2	0.6667	0.4544	**0.0000**
	NUMB_PRKCA	0.1999	0.4065	**0.0000**

All proteins in the table are expressed in the three tissues, and the interactions with *P-value* below 0.05 in one tissue while above 0.1 in the other two tissues will be considered as tissue-specific interactions, where the *P-value* was calculated as the significance of Pearson's correlation coefficients.

composition of proteins, protein expression profiles and protein interactome, the possibility of a motif M_i interacting with a kinase K_j can be described as follows.

$$Prob(M_i, K_j) = \frac{\sum_{P_n \in P, M_i \in P_n} S_{jn} \cdot I_{jn}}{\sum_{P_m \in N, M_i \in P_m} S_{jm} \cdot I_{jm}} \quad (1)$$

where $Prob(M_i, K_j)$ is the probability that motif M_i interacts with kinase K_j, P denotes the phosphorylated protein set containing motif M_i while N represents all the proteins containing motif M_i, S_{jn} denotes the correlation coefficient between the kinase K_j and protein P_n calculated based on their expression profiles, and I_{jn} is an indicator function defines as follows.

$$I_{jn} = \begin{cases} 1, \ if \ protein \ j \ and \ n \ interact \\ 0, \ otherwise \end{cases} \quad (2)$$

After obtaining the probability of kinase-motif interactions, we can predict the probability that a kinase interacts with a protein as follows.

$$Prob(P_n, K_j) = 1 - \prod_{M_i \in P_n} (1 - I_{jn} \cdot Prob(M_i, K_j)) \quad (3)$$

where $Prob(P_n, K_j)$ represents the probability that protein P_n containing motif M_i interacts with kinase K_j, and $Prob(M_i, K_j)$ is the probability of motif M_i interacting with kinase K_j. We can set a threshold to determine whether a kinase interacts with a protein, where those kinase-protein pairs with probabilities above the threshold were treated as interacting pairs.

III. RESULTS AND DISCUSSION

A. Prediction of kinase-specific interactions

With the assumption that kinases phosphorylate their substrate proteins through the interactions between kinases and certain motifs within proteins, we obtained the probabilities of 102 motifs interacting with 137 kinases with the model described in Eq (1). With these kinase-motif interactions, we calculated the probabilities of kinases interacting with proteins based on Eq (3). With a threshold of 0.6, we obtained 711 kinase-substrate interactions between 126 kinases and 221 proteins. We further constructed a phosphorylation network based on the interactions as shown in Fig.1 A, where one edge was laid if there exists an interaction between a kinase and a protein. The network was visualized with the Cytoscape software [15]. The phosphorylation network constructed here can provide insights into the biological processes in which phosphorylation is involved, such as signal processing, development and disease. In addition, the kinase-motif interactions identified can help one understand the phosphorylation processes.

The above phosphorylation network was constructed based on all the protein expression data, and serve as background network for future analysis. Considering the protein expression and phosphorylation data in three tissues, we constructed three phosphorylation networks respectively for the three tissues based on their corresponding protein expression data. Finally, we obtained 350 interactions between 124 proteins and 76 kinases for the cerebrum, 321 interactions between 115 proteins and 73 kinases for the prefrontal cortex, and 94 interactions between 61 proteins and 29 kinases for the liver. The detailed networks of the three tissues are shown in Fig.1 B-D, from which we can see the phosphorylation networks of the three tissues are different, implying the specificity of different tissues.

Furthermore, we investigated the three phosphorylation networks to see whether there are tissue-specific interactions. We assumed that the correlation coefficient between a kinase and its substrate will be high if this interaction indeed exists in a certain tissue, and the correlation coefficients were therefore regarded as the interaction strength for the interacting protein pair. In particular, we regarded those kinase-protein pairs as tissue-specific interactions if their correlation coefficients are significantly higher in one tissue than in the other two tissues. Specially, those interactions with *p-value* below 0.05 in one tissue but above 0.1 in the other two tissues were treated as tissue-specific interactions. Table II summarizes the tissue-specific interactions we identified for the three tissues, where only the interactions that appear in all three tissues were shown.

To further evaluate the tissue specificity of the three phosphorylation networks, we utilized Network Ontology Analysis (NOA) [16] to perform functional enrichment analysis of the three networks. As shown in Table III, those functions enriched in the networks are functionally consistent with the functions of the three tissues, indicating the tissue-specificity characterized by the networks. For example, the phosphorylation network constructed for liver is enriched in metabolic processes and response to toxin. The enrichment of cerebrum tissue-specific network is related to the developmental process, negative regulation of cell communication and receptor signaling protein tyrosine kinase activity while the prefrontal cortex is associated with aging and protein serine/threonine kinase activity. Surprisingly, the subnetworks and functional enrichment of CB and PFC is more similar which can be explain by that similar tissue will be associated with similar functions. It can strongly indicates the reliability of the tissue specificity subnetworks and may further prompt that the functional data can be used to identify substrates with similar functions.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

TABLE III. THE ENRICHMENT OF THE BIOLOGICAL FUNCTION IN THREE TISSUE-SPECIFIC PHOSPHORYLATION NETWORKS.

Tissues	GO Terms	*p-value*	Definition
CB	GO:0032502	0.0000	developmental process
CB	GO:0050896	0.0000	response to stimulus
CB	GO:0010648	0.0000	negative regulation of cell communication
CB	GO:0035467	0.0000	negative regulation of signaling pathway
CB	GO:0042221	0.0000	response to chemical stimulus
CB	GO:0004871	0.0000	signal transducer activity
CB	GO:0060089	0.0000	molecular transducer activity
CB	GO:0004716	0.0091	receptor signaling protein tyrosine kinase activity
CB	GO:0004715	0.0157	non-membrane spanning protein tyrosine kinase activity
PFC	GO:0010648	0.0000	negative regulation of cell communication
PFC	GO:0035467	0.0000	negative regulation of signaling pathway
PFC	GO:0048583	0.0000	regulation of response to stimulus
PFC	GO:0007568	0.0000	aging
PFC	GO:0010646	0.0001	regulation of cell communication
PFC	GO:0048519	0.0002	negative regulation of biological process
PFC	GO:0009892	0.0002	negative regulation of metabolic process
PFC	GO:0004674	0.0380	protein serine/threonine kinase activity
LV	GO:0051171	0.0058	regulation of nitrogen compound metabolic process
LV	GO:0005057	0.0074	receptor signaling protein activity
LV	GO:0019219	0.0076	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
LV	GO:0048545	0.0116	response to steroid hormone stimulus
LV	GO:0031323	0.0123	regulation of cellular metabolic process
LV	GO:0060255	0.0132	regulation of macromolecule metabolic process
LV	GO:0032502	0.0225	developmental process
LV	GO:0009636	0.0334	response to toxin
LV	GO:0050790	0.0334	regulation of catalytic activity
LV	GO:0019899	0.0366	enzyme binding

IV. CONCLUDING REMARKS

Phosphorylation is a reversible post-translational modification process that plays key roles in many biological process. Identification of protein substrates phosphorylated by kinases is the key to understanding the phosphorylation. In this work, a novel approach was presented to predict the proteins phosphorylated by kinases with the assumption that kinase-protein interactions are accomplished by kinase-motif interactions. We constructed three tissue-specific phosphorylation networks based on protein expression data and protein-protein interactions, where functional enrichment analysis of the networks indicate the three tissue-specific phosphorylation networks are functionally consistent with the corresponding tissues. The phosphorylation networks constructed here can provide insights into phosphorylation and other important biological processes.

ACKNOWLEDGMENT

This work was partly supported by the National Natural Science Foundation of China (91130032, 61103075), Innovation Program of Shanghai Municipal Education Commission (13ZZ072), and Shanghai Pujiang Program (13PJD032)..

REFERENCES

[1] B. T. Seet, I. Dikic, M. M. Zhou and T. Pawson, Reading protein modifications with interaction domains, *Nature Reviews Molecular Cell Biology*, vol.7, no.7, pp. 473-483, Jul 2006.

[2] S. Zolnierowicz and M. Bollen, Protein phosphorylation and protein phosphatases. De panne, belgium, september 19-24, 1999, *The EMBO journal*, vol.19, no.4, pp. 483-488, Feb 15 2000.

[3] M. Mann, S. E. Ong, M. Gronborg, H. Steen, O. N. Jensen and A. Pandey, Analysis of protein phosphorylation using mass spectrometry: Deciphering the phosphoproteome, *Trends in biotechnology*, vol.20, no.6, pp. 261-268, Jun 2002.

[4] R. Aebersold and M. Mann, Mass spectrometry-based proteomics, *Nature*, vol.422, no.6928, pp. 198-207, Mar 13 2003.

[5] H. Dinkel, C. Chica, A. Via, C. M. Gould, L. J. Jensen, T. J. Gibson and F. Diella, Phospho.Elm: A database of phosphorylation sites–update 2011, *Nucleic acids researc*h, vol.39, no.Database issue, pp. D261-267, Jan 2011. 2011;39:D261-7.

[6] G. Manning, D. B. Whyte, R. Martinez, T. Hunter and S. Sudarsanam, The protein kinase complement of the human genome, *Science*, vol.298, no.5600, pp. 1912-1934, Dec 6 2002.

[7] M. Hjerrild, A. Stensballe, O. N. Jensen, S. Gammeltoft and T. E. Rasmussen, Protein kinase a phosphorylates serine 267 in the homeodomain of engrailed-2 leading to decreased DNA binding, *FEBS letters*, vol.568, no.1-3, pp. 55-59, Jun 18 2004.

[8] J. C. Obenauer, L. C. Cantley and M. B. Yaffe, Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs, *Nucleic acids research*, vol.31, no.13, pp. 3635-3641, Jul 1 2003.

[9] R. H. Newman, J. Hu, H. S. Rho, Z. Xie, C. Woodard, J. Neiswinger, C. Cooper, M. Shirley, H. M. Clark, S. Hu, W. Hwang, J. S. Jeong, G. Wu, J. Lin, X. Gao, Q. Ni, R. Goel, S. Xia, H. Ji, K. N. Dalby, M. J. Birnbaum, P. A. Cole, S. Knapp, A. G. Ryazanov, D. J. Zack, S.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

Blackshaw, T. Pawson, A. C. Gingras, S. Desiderio, A. Pandey, B. E. Turk, J. Zhang, H. Zhu and J. Qian, Construction of human activity-based phosphorylation networks, *Molecular systems biology*, vol.9, pp. 655 2013.

[10] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady and A. Pandey, Human protein reference database–2009 update, *Nucleic acids research*, vol.37, no.Database issue, pp. D767-772, Jan 2009.

[11] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi and L. S. Yeh, Uniprot: The universal protein knowledgebase, *Nucleic acids research*, vol.32, no.Database issue, pp. D115-119, Jan 1 2004.

[12] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz and M. Tyers, Biogrid: A general repository for interaction datasets, *Nucleic acids research*, vol.34, no.Database issue, pp. D535-539, Jan 1 2006.

[13] K. Xia, D. Dong and J. D. Han, Intnetdb v1.0: An integrated protein-protein interaction network database generated by a probabilistic model, *BMC bioinformatics*, vol.7, pp. 508 2006.

[14] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering and L. J. Jensen, String v9.1: Protein-protein interaction networks, with increased coverage and integration, *Nucleic acids researc*h, vol.41, no.Database issue, pp. D808-815, Jan 2013.

[15] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, Cytoscape: A software environment for integrated models of biomolecular interaction networks, *Genome research*, vol.13, no.11, pp. 2498-2504, Nov 2003.

[16] J. Wang, Q. Huang, Z. P. Liu, Y. Wang, L. Y. Wu, L. Chen and X. S. Zhang, Noa: A novel network ontology analysis method, *Nucleic acids research*, vol.39, no.13, pp. e87, Jul 2011.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE