Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Prediction of regulation relationship between protein interactions in signaling networks

Wei Liu [*], Hongwei Xie

The College of Mechanical & Electronic Engineering and Automatization, National University of Defense Technology, 410073 Changsha, China

ARTICLE INFO

Article history:
Received 3 September 2013
Available online 1 October 2013

Keywords:
Regulation relationship
Protein interaction
Signaling network
Logistic regression

ABSTRACT

The discovery of regulation relationship of protein interactions is crucial for the mechanism research in signaling network. Bioinformatics methods can be used to accelerate the discovery of regulation relationship between protein interactions, to distinguish the activation relations from inhibition relations. In this paper, we describe a novel method to predict the regulation relations of protein interactions in the signaling network. We detected 4,417 domain pairs that were significantly enriched in the activation or inhibition dataset. Three machine learning methods, logistic regression, support vector machines(SVMs), and naïve bayes, were explored in the classifier models. The prediction power of three different models was evaluated by 5-fold cross-validation and the independent test dataset. The area under the receiver operating characteristic curve for logistic regression, SVM, and naïve bayes models was 0.946, 0.905 and 0.809, respectively. Finally, the logistic regression classifier was applied to the human proteome-wide interaction dataset, and 2,591 interactions were predicted with their regulation relations, with 2,048 in activation and 543 in inhibition. This model based on domains can be used to identify the regulation relations between protein interactions and furthermore reconstruct signaling pathways.

1. Introduction

With the development of high-throughput technologies, large-scale protein–protein interaction (PPI) data for multiple species has been produced, which provided the basis for the investigation of protein function and dynamics [1–6]. An important investigation area is discovering the potential signaling pathways from protein interactions to understand their roles in signal transduction, gene regulation and disease. The typical experimental method to infer the regulation relations between pathway components is perturbing the cells with molecular interventions [7,8]. It needs many experiments to determine their molecular mechanism and regulation relationships, which is expensive, time-consuming and error-prone.

Several groups have made efforts to develop bioinformatics methods to infer signaling pathways [9–14]. For example, Steffen, et al. developed a computational approach to generate static models of signal transduction networks from large-scale two-hybrid screens and expression profiles [9]. Silverbush et al. [10] and Gitter et al. [11] presented several algorithms to discover high-confidence pathways. Shlomi et al. presented a comprehensive framework, Qpath, using homologous pathway queries to identify biologically significant pathways and their functions [12]. We have also proposed two methods to predict the directionality in pairwise proteins, based on the domains and functional annotations [13,14]. These methods can achieve good performance in a part of protein interaction datasets. However, it was still difficult to determine the regulation relationship of protein interactions in the signaling pathways. Giving a pair of interacting proteins, we can predict the direction of signal flow through it using the methods proposed in [13,14], but we cannot distinguish whether its regulation relation is activation or inhibition. Therefore, it is necessary to develop new bioinformatics methods to predict the regulation relations between protein interactions.

In this paper, we introduced a novel method to predict the regulation relationship between protein interactions in the signaling network according to their constituent domains. Firstly, we proposed a measure, Enrichment_ratio, to identify the domain pairs significantly enriched in the activation/inhibition dataset. Then, we trained the classifiers based on three machine learning methods (logistic regression, SVM and naïve bayes) with the activation dataset and the inhibition dataset. Furthermore, we evaluated these methods based on 5-fold cross-validation and the independent test dataset. Finally, we applied the logistic regression method to predict the regulation relations in the human proteome-wide interactions.

[*] Corresponding author.
E-mail addresses: angel_nudt@126.com, liuwei314@nudt.edu.cn (W. Liu).

0006-291X/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.bbrc.2013.09.093

2. Materials and methods

2.1. Extraction of signaling networks in multiple species

As a classical and well-known pathway database, KEGG (Kyoto Encyclopedia of Genes and Genomes) contains manually annotated pathways based on biochemical evidence from the literature, including a large amount of signaling and metabolic pathways [15]. All the signaling networks of human, mouse, rat, fly and yeast were downloaded from KEGG. From these signaling networks, 1,893 protein interactions are extracted with their regulation relationship, including 1,554 in the category of activation and 339 in the inhibition, which are used as the golden standard positive set. In human, rat, mouse, fly and yeast, 76.40% proteins have one or more Pfam domains. Interaction between two proteins typically involves binding between specific domains.

2.2. Transforming human signaling pathways to protein interactions

By transforming protein interactions in signaling pathways into binary model, we transformed the pathways in 7 databases, including PID, BioCarta, Reactome, NetPath, INOH, SPIKE and KEGG and established the human protein interaction dataset with known regulation relations. This dataset include 6,791 protein interactions (Additional file 1), with 5,261 in activation and 1,530 in inhibition. Abandoning the interactions recorded in the golden standard positive set, the rest can be used as the independent test dataset to evaluate the performance of our classifier.

2.3. The computation of Enrichment_ratio and P-value of domain pairs

To investigate the enrichment extent of a domain pair appearing in the activation dataset or inhibition dataset, compared to the whole protein interaction dataset, we proposed a novel measure Enrichment_ratio. It is defined as:

$$\text{Enrichment\_ratio} = \frac{\frac{m}{M}}{\frac{n}{N}} \tag{1}$$

where N is the number of protein interactions in the whole standard dataset, M is the number of protein interactions in the activation/inhibition dataset. For a specific pair of domains, n is defined as the number of protein interactions containing this pair in the whole standard dataset, and m is the number of protein interactions containing this pair in the activation/inhibition dataset. For a given pair of domains, we can calculate two Enrichment_ratio values, one of which represents the enrichment extent in the activation dataset and the other represents the enrichment extent in the inhibition dataset. If the Enrichment_ratio in the activation dataset is larger than a certain cutoff, such as 1, then this pair is relatively enriched in this dataset. If the Enrichment_ratio is smaller than 1, it is relatively lacked. The enriched domain pairs in the inhibition dataset can be extracted by the similar method.

Furthermore, to investigate whether two domains always appear in pairs to introduce protein interaction or they only appear accidentally, we made a hypothesis testing. We used the hypergeometric cumulative distribution to analyze the enrichment significance of domain pairs appearing in the activation/inhibition dataset. For a specific pair of domains, its *P*-value is defined as:

$$P - value = \sum_{m'=m}^{n} \frac{\binom{M}{m'}\binom{N-M}{n-m'}}{\binom{N}{n}} \quad (\text{Enrichment\_ratio} \geq 1) \tag{2}$$

$$P - value = \sum_{m'=0}^{m} \frac{\binom{M}{m'}\binom{N-M}{n-m'}}{\binom{N}{n}} \quad (\text{Enrichment\_ratio} < 1) \tag{3}$$

Through setting the cutoff the *P*-value, for instance 0.05, we can discover the domain pairs significantly enriched in the activation/inhibition dataset. If the Enrichment_ratio > 1 and *P*-value < 0.05, this pair of domains is regarded as significantly enriched in the activation/inhibition dataset.

2.4. Machine learning

Three machine-learning algorithms were investigated: logistic regression, and support vector machine (SVM) based on PolyKernel, and naïve bayes, all of which have been widely used for pattern classification and regression problems. For a specific domain pair selected by the enrichment analysis, if it appears in the protein interactions of the training dataset, the corresponding feature is set as its Enrichment_ratio, otherwise this feature is set as zero. The WEKA package [16] was used to build classifiers that could distinguish the activation relations from inhibition, using selected features.

We can evaluate the performance of three classifiers using 5-fold cross-validation. During the test process, 20% of the interactions in the positive and negative datasets were singled out in turn to become the test sample, and the remaining interactions were used as the training set to predict the class of the interactions in the test sample. The performance was measured by the analysis of receiver operating characteristic (ROC) curves. A ROC curve can show the efficacy of one test by presenting both sensitivity and specificity for different cutoff points [17]. Sensitivity and specificity can measure the ability of a test to identify true positive and false positives in a data set. These two indexes can be calculated as Sensitivity = TP/T and Specificity = 1 − (FP/F) where TP and FP are the number of identified true and false positives, respectively, whereas T and F are the total number of positives and negatives in a test. The area under the ROC curve (AUC) provided the metric for the overall performance of the classifier. The closer the AUC of a test was to 1.0, the higher the overall efficacy of the test.

3. Results

3.1. Extraction of domain pairs enriched in activation/inhibition dataset

Domains are elements of proteins in a sense of structure and function. Most proteins interact with each other through their domains. Therefore, it is crucial and useful to understand PPIs based on the domains [18]. In Fig. 1, we gave an example to demonstrate the domain pairs contained in the protein interactions. Protein A contains three domains D_1, D_2 and D_3, and Protein B contains two domains E_1 and E_2. In principle, the domains contained in Protein A and the domains contained in Protein B can compose 6 domain pairs. In fact, only few domain pairs will be significantly enriched in protein interactions of the activation or inhibition dataset. These domain pairs significantly enriched in the activation or inhibition dataset may suggest the regulation relations between protein interactions, and can be used as the valid features to build classifiers in order to distinguish the activation relations from the inhibition relations.

According to the constituent domains of interacting proteins, we computed the Enrichment_ratios and P-values of domain pairs in the activation/inhibition dataset (see Section 2). We extracted 7,805 pairs of domains significantly enriched with their Enrichment_ratio > 1 and P-value < 0.05, in which 5,796 pairs are enriched in the activation dataset and 2,009 pairs enriched in the

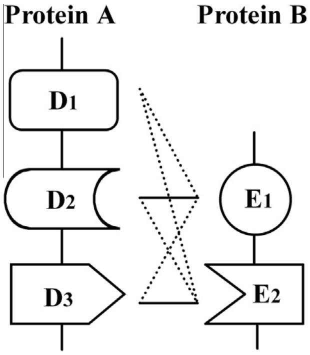

Fig. 1. The domain pairs contained in protein interactions.

inhibition dataset. Since some of them only appear in the activation/inhibition dataset once or two times, we can filter the domain pairs according to their appearing number. When the number of domain pairs appearing in the activation/inhibition dataset restricted with larger than 2, we obtained 2,949 pairs of domains significantly enriched, including 2,091 pairs in the activation dataset and 858 pairs in the inhibition dataset (Additional file 2). These pairs of domains can be used as features to distinguish the regulation relationship between protein interactions. In Table 1, we gave a partial list of domain pairs enriched in the activation/inhibition dataset. The relations of some domain pairs well match known knowledge about signal transduction. For example, the regulation relationship between proteins containing domain RasGEF and Ras are always activation, with the Enrichment_ratio = 1.218 and *P*-value < 0.05.

3.2. Model evaluation based on 5-fold cross-validation and the independent test dataset

Classifiers based on the three methods (logistic regression, SVM and naïve bayes) were trained with the activation dataset and the inhibition dataset. When we restricted the number of pair appearing in the activation/inhibition dataset larger than 2, the activation Enrichment_ratio of domain pairs larger than 1.2 and the inhibition Enrichment_ratio of domain pairs larger than 1.5, the number of features decreased to 817. These selected features were used as input to the classifiers to distinguish between the activation and inhibition relationship in the test datasets.

Based on 5-fold cross-validation, we evaluated the performance of three machine learning classifiers. The ROCs for the cross-validation results were shown in Fig. 2. The AUC of the methods based on logistic regression, SVM and naïve bayes is 0.946, 0.905 and 0.809, respectively. The logistic regression model has the largest AUC, suggesting that it has a relatively high ability to predict the regulation relations of protein interactions in the signaling network.

Furthermore, we applied the model based on logistic regression to the protein interactions distilled from the human signaling networks (see Section 2). As a result, our classifier can distinguish most of the protein interactions between activation and inhibition, with accuracy 73.24%. Our method also achieved good performance in multiple signaling pathways of human (see Table 2). Consequently, it shows that the model based on domains can achieve satisfactory results in the complex signaling pathways of human. This model based on domains can be used to identify the regulation relations between protein interactions and furthermore reconstruct signaling pathways.

3.3. Application to the human proteome-wide interaction dataset

We collected human protein interactions from HPRD, DIP, MINT, BIND database and the previous resources [19,20]. After processing, we obtained 45,238 non-redundant interactions which have corresponding Entrez Gene ID and not reported in protein complex. These interactions composed the human proteome-wide interaction dataset, most of regulation relations of which were unknown. As an application, we used the method based on logistic regression to comprehensively predict the regulation relations of

Table 1
The partial list of domain pairs enriched in the activation/inhibition dataset.

Regulation relationship	Domain A	Pfam name	Domain B	Pfam name	Enrichment_ratio	*P*-value
Activation	PF00048	IL8	PF00001	7tm_1	1.218	0
Activation	PF05296	TAS2R	PF00048	IL8	1.218	0
Activation	PF01748	Serpentine_recp	PF00048	IL8	1.218	0
Activation	PF00229	TNF	PF00020	TNFR_c6	1.218	0
Activation	PF01461	7tm_4	PF00048	IL8	1.218	0
Activation	PF07654	C1-set	PF07686	V-set	1.218	0
Activation	PF01030	Recep_L_domain	PF00008	EGF	1.218	0
Activation	PF00001	7tm_1	PF00025	Arf	1.218	0
Activation	PF00503	G-alpha	PF00001	7tm_1	1.218	0
Activation	PF00069	Pkinase	PF07974	EGF_2	1.218	0
Activation	PF07714	Pkinase_Tyr	PF07974	EGF_2	1.218	0
Activation	PF07654	C1-set	PF00047	ig	1.218	0
Activation	PF01030	Recep_L_domain	PF07974	EGF_2	1.218	0
Activation	PF00617	RasGEF	PF00071	Ras	1.218	0
Activation	PF00617	RasGEF	PF01926	MMR_HSR1	1.218	0
Inhibition	PF00051	Kringle	PF00079	Serpin	5.584	0
Inhibition	PF03761	DUF316	PF00079	Serpin	5.584	0
Inhibition	PF00653	BIR	PF00656	Peptidase_C14	5.584	0
Inhibition	PF03029	ATP_bind_1	PF00616	RasGAP	5.584	0
Inhibition	PF00019	TGF_beta	PF00019	TGF_beta	5.584	0
Inhibition	PF00688	TGFb_propeptide	PF00019	TGF_beta	5.584	0
Inhibition	PF00688	TGFb_propeptide	PF00688	TGFb_propeptide	5.584	0
Inhibition	PF00097	zf-C3HC4	PF00017	SH2	5.584	0
Inhibition	PF00056	Ldh_1_N	PF00071	Ras	5.584	0
Inhibition	PF02262	Cbl_N	PF00017	SH2	5.584	0
Inhibition	PF02761	Cbl_N2	PF00017	SH2	5.584	0
Inhibition	PF00024	PAN_1	PF00079	Serpin	5.584	0
Inhibition	PF00056	Ldh_1_N	PF08477	Miro	5.584	0
Inhibition	PF00056	Ldh_1_N	PF00009	GTP_EFTU	5.584	0
Inhibition	PF00018	SH3_1	PF00097	zf-C3HC4	5.584	0

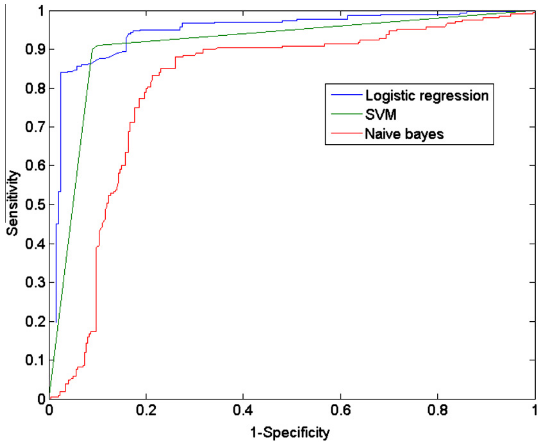

Fig. 2. ROCs of the classifier based on logistic regression, SVM and naïve bayes.

Table 2
The prediction results of the method in human classical signaling pathways.

Signaling pathways	Accuracy (%)
MAPK signaling pathway	100
GnRH signaling pathway	100
B cell receptor signaling pathway	100
mTOR signaling pathway	75
VEGF signaling pathway	100
Cell cycle	90
Jak-STAT signaling pathway	100
ErbB signaling pathway	100
Apoptosis	66.67
Wnt signaling pathway	96.97
Fc epsilon RI signaling pathway	100
Adipocytokine signaling pathway	100
Insulin signaling pathway	100
T cell receptor signaling pathway	100
TGF-beta signaling pathway	86.96

protein interactions in this dataset. Totally, 2,591 protein interactions are predicted with their regulation relations, with 2,048 in activation and 543 in inhibition. Then we compared the predicted results with known signaling pathway databases. 235 predicted interactions were present in known signaling databases, with accuracy 76.17%. The rest of interactions were firstly predicted with regulation relations, which should be valuable resources to unveil potential signaling pathways (Additional file 3).

As a result, we established the first predicted human interaction network marked with regulation relations, including 1,544 proteins and 2,591 interactions (Fig. 3A). From it, many potential signaling pathways can be distilled. For example, two typical clusters were given in Fig. 3B and C. These inferred pathways expand the size of known signaling pathways and suggest new biological mechanisms that are not currently present in signaling networks.

4. Discussion and conclusions

Regulation relationship is one of the most important features of the protein interactions in signaling networks. The determination of the regulation relations of protein interactions is crucial to reveal potential signaling pathways and construct signaling network. Reconstruction of signaling networks from protein interactions might be applied to understanding signaling transduction process, complex drug actions, and dysfunctional signaling in diseased cells [21].

In this paper, we proposed three methods based on logistic regression, SVM and naïve bayes to infer the regulation relations through protein interactions based on their corresponding domains. Through the evaluation based on 5-fold cross-validation and independent test dataset, the method based on logistic regression was proved to have higher sensitivity and specificity. Finally, this method was applied to predict regulation relations in human proteome-wide interactions and provided a global regulation relation annotation of the protein interaction network. As a conclusion, this method can provide the comprehensive understanding for the signaling network, and suggest new biological mechanisms that are not currently present in signaling networks.

The limit of this method lies in that it cannot be applied to the interacting proteins, if they do not contain domains. With more biological data sources and more types of evidences integrated into the classifier, this method will achieve better performance and reconstruct larger sale signaling network.

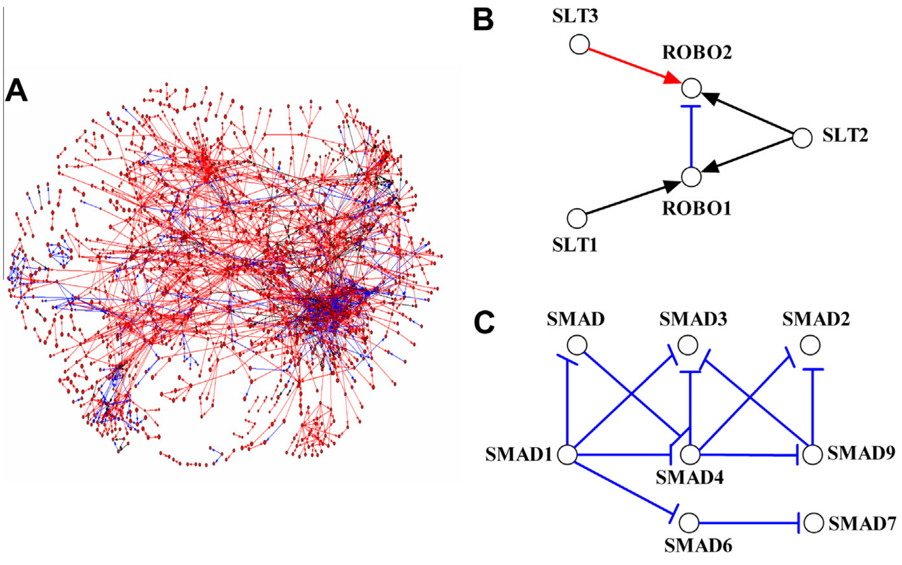

Fig. 3. The human protein interaction network marked with predicted regulation relations and its two typical clusters. (A) The whole interaction network marked with predicted regulation relations. According to the predicted regulation relations, the protein interactions are marked with different colors. The interactions predicted with activation relations are marked with red lines, interactions predicted with inhibition marked with blue lines, and those with known regulation relations in signaling databases marked with black. (B) One cluster with predicted relations and known relations. (C) Another cluster whose relations were all predicted as inhibition.

Acknowledgments

We thank Drs. Jiyang Zhang, Tengjiao Wang and Changming Xu for their excellent advice and assistance as well as all the members in the Bioinformatics Laboratory, College of Mechanical & Electronic Engineering and Automatization, National University of Defense Technology for helpful discussions. This work was supported by the National Natural Science Foundation of China (Grant Nos. 31000591, 31000587, and 31171266).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bbrc.2013.09.093.

References

[1] H.W. Mewes, K. Heumann, A. Kaps, K. Mayer, F. Pfeiffer, S. Stocker, D. Frishman, MIPS: a database for genomes and protein sequences, Nucleic Acids Res. 30 (2002) 31–34.

[2] U. Peter, G. Loic, C. Gerard, A.M. Traci, S.J. Richard, R.K. James, L. Daniel, N. Vaibhav, S. Maithreyan, P. Pascale, Q. Alia, L. Ying, G. Brian, C. Diana, K. Theodore, V. Govindan, Y. Meijia, J. Mark, F. Stanley, M.R. Jonathan, A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae, Nature 403 (2000) 623–627.

[3] 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 (2001) 4569.

[4] S. Li, C.M. Armstrong, N. Bertin, A map of the interactome network of the metazoan C. elegans, Science 303 (2003) 540–543.

[5] L. Giot, J.S. Bader, C. Brouwer, A. Chaudhuri, A protein interaction map of Drosophila melanogaster, Science 302 (2003) 1727–1736.

[6] S. Peri, J.D. Navarro, R. Amanchy, Development of human protein reference database as an initial platform for approaching systems biology in humans, Genome Res. 13 (2003) 2363–2371.

[7] K. Sachs, O. Perez, D. Pe'er, D.A. Lauffenburger, G.P. Nolan, Causal protein-signaling networks derived from multiparameter single-cell data, Science 308 (2005) 523–529.

[8] D. Silverbush, M. Elberfeld, R. Sharan, Optimally orienting physical networks, J. Comput. Biol. 18 (2011) 1437–1448.

[9] M. Steffen, A. Petti, J. Aach, P. Dhaeseleer, G. Church, Automated modeling of signal transduction networks, BMC Bioinf. 3 (2002) 34.

[10] D. Silverbush, M. Elberfeld, R. Sharan, Optimally orienting physical networks, J. Comput. Biol. 18 (2011) 1437–1448.

[11] A. Gitter, J. Klein-Seetharaman, Z. Bar-Joseph, Discovering pathways by orienting edges in protein interaction networks, Nucleic Acids Res. 39 (2010) e22.

[12] T. Shlomi, D. Segal, E. Ruppin, R. Sharan, QPath: a method for querying pathways in a protein–protein interaction network, BMC Bioinf. 7 (2006) 199.

[13] W. Liu, D. Li, J. Wang, H. Xie, Y. Zhu, F. He, Proteome-wide prediction of signal flow direction in protein interaction networks based on interacting domains, Mol. Cell. Proteomics 8 (2009) 2063–2070.

[14] W. Liu, D. Li, Y. Zhu, H. Xie, F. He, Reconstruction of signaling network from protein interactions based on function annotations, IEEE/ACM Trans. Comput. Biol. Bioinform. 10 (2013) 514–521.

[15] M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (2000) 27–30.

[16] P. Baldi, S. Brunak, Y. Chauvin, C.A. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (2000) 412–424.

[17] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1982) 29–36.

[18] K. Xia, D. Dong, J.D. Han, IntNetDB v1.0: an integrated protein–protein interaction network database generated by a probablistic model, BMC Bioinf. 7 (2006) 508.

[19] U. Stelzl, U. Worm, E.E. Wanker, A human protein–protein interaction network: a resource for annotating the proteome, Cell 122 (2005) 957–968.

[20] J.F. Rual, K. Venkatesan, M. Vidal, Towards a proteome-scale map of the human protein–protein interaction network, Nature 437 (2005) 1173–1178.

[21] W.C. Hahn, R.A. Weinberg, Modelling the molecular circuitry of cancer, Nat. Rev. Cancer 2 (2002) 331–341.