Dynamic protein-protein interaction networks and the detection of protein complexes: an overview

Eileen Marie Hanna and Nazar Zaki
College of Information Technology, United Arab Emirates University, Al Ain, Abu Dhabi, UAE

Abstract - *Developing computational approaches for the detection of protein complexes in protein-protein interaction networks continues to be an evolving area of research. These approaches seek to complement the experimental methods which are usually expensive in terms of time and cost. A protein-protein interaction dataset is typically modeled as a static network whose vertices and edges respectively represent all the proteins and their interconnections. Despite the agreeable accuracies attained by various computational methods when applied on such networks, their additional improvements seem to face some limitations. It is believed that the more enrichment with biological information is added to the interaction networks and complex-detection algorithms, the better will be the overall quality of the results. In this paper, we stress on the importance of reflecting the dynamic nature of protein interaction networks as a primary enhancement phase and we highlight possible aspects by which it could be acquired.*

Keywords: protein-protein interactions, protein complex, dynamic protein-protein interaction network.

1 Introduction

From metabolism to signal transduction, transport, cellular organization and ultimately all biological processes, proteins are the key players. Their interconnections shape interaction networks which define highly-organized cellular systems [1]. Biological functions are often acquired through collaborations of interacting protein groups referred to as protein complexes [2]. The progress in identifying protein complexes, involved in normal molecular events as well as phenotypes associated with diseases, allows the progressive development of effective cures. Accordingly, various experimental methods were designed to identify complexes given protein-protein interaction (PPI) data. However, in addition to their high computational cost, they are also susceptible to high error rates [3]. Therefore, several computational approaches came into the picture to complement the experimental activities. For instance, protein complexes detected by computational algorithms with suitable accuracy and quality could guide the experimental examinations and expectantly reduce the necessary biological explorations.

In a computational setting, a PPI dataset is usually modeled as a graph whose vertices and edges represent all the proteins and their interactions respectively. In this context, the majority of the computational approaches are based on the concept by which protein complexes correspond to dense subgraphs. These methods include, but are not limited to, Markov Clustering (MCL) [4] which uses random walks in protein interaction networks; the molecular complex detection (MCODE) algorithm [5] which identifies complexes as dense regions grown from highly-weighted vertices; the clustering based on maximal cliques (CMC) method [6]; the Affinity Propagation (AP) algorithm [7]; ClusterONE [8] which identifies protein complexes through clustering with overlapping neighborhood expansion; the restricted neighborhood search (RNSC) algorithm [9,10]; the RRW algorithm which generates complexes by using repeated random walks [11]; and CFinder [12] which is based on the clique percolation method. Other approaches which are not based on the density notion include ProRank [13,14] which mainly uses a protein ranking algorithm to identify essential proteins in a PPI network; ProRank+ [15] which is an improved version of ProRank, it reflects the fact that proteins can be multifunctional and thus could belong to multiple complexes and it applies a merging procedure to improve the detected complexes; and finally PEWCC [16,17] which assesses the reliability of PPI data based on the weighted clustering coefficient notion prior to detecting protein complexes. When evaluated based on reference sets of biologically-identified protein complexes, these algorithms were on the right track. Nevertheless, their improvements towards reducing false positive and false negative outcomes seem to be bounded by the way in which PPI data is originally utilized and by the false positive and false negative interactions as well. The traditional experimental approaches used to study PPIs, such as yeast two-hybrid (Y2H) [18] and TAP-MS [19], do not provide temporal, spatial or contextual information

across which a PPI occurs. In contrast, recent methodological advances, such as ChIP-chip [20] and ChIP-seq [21] can make such informative data available. Consequently, advances in the computational approaches developed to analyze PPI networks, including those designed to detect protein complexes, ought to relate to such diversity of information that is currently presented. PPI networks are dynamic in nature [22]. Accordingly, modeling the dynamicity of PPI networks is a necessary shift in the way such networks are viewed and studied [23]. It is actually essential and allows us to expand our knowledge about how cellular processes occur. In this paper, we highlight the advantages, potential approaches and possible bottlenecks of this emerging construal of PPI networks.

2 The advantages of shifting to dynamic PPI networks

2.1 Enhancing the replication of real biological events

The shift to dynamic PPI networks in computational approaches of systems biology comes as a natural response to advances in experimental methods by which novel types and increased amounts of biological data are generated. As an interdisciplinary area of research, the more representative are its building models and methods, the better is its aptitude. Moreover, when cellular interactions are reproduced in a more realistic manner, the accountability and accuracy of the results produced by computational methods will certainly augment.

2.2 Potentially uncovering previously unknown biological facts

A PPI dataset is conventionally represented as a comprehensive graph which includes the proteins along with all their interactions. However, not all the interconnections happen at the same time. In fact, the occurrence of a PPI is subject to various temporal, spatial and contextual conditions. Obviously, encompassing such conditionality parameters elucidates the dynamics of PPIs. In view of that, by combining biological information, we would reach a computational visualization level of protein interaction events that could verify or even contradict biological concepts. Furthermore, previously unknown facts may be learned, such as the characterization of hub proteins in [24] as “party hubs” which interact with their partners at the same time or “date hubs” which connect to their partners at different times and locations.

2.3 Possibly overcoming data limitations

The biological methods used to identify protein interactions are very sensitive to experimental settings. Therefore, the PPI datasets that they generate are always liable to high error rates. Many algorithms were developed to filter protein interactions according to their reliability levels. For example, some of these methods use weighting schemes based on the number of common neighbors of interacting proteins such as CDdistance [25], FSWeight [26] and AdjustCD [27]. Similarly, the PE-measure introduced in [16,17] reduces the level of noise in protein interaction networks by looking for subgraphs that are closest to maximal cliques based on the weighted clustering coefficient measures. In addition, possible enrichment data that can be used to model the dynamicity of PPI networks, such as gene expression profiles [28] and gene ontology [29], suffer from low gene coverage in contrast with most PPI datasets, in which the number of interacting proteins is typically very high [30]. The recurrence of information and/or inferences that are drawn from different types of biological data can be seen as a confidence indicator. In view of that, combining various datasets, although not fully-credible, in the direction of modeling PPI dynamics could potentially reduce data limitations such as the effect of false positives and false negative rates, as well as low coverage issues.

2.4 Increasing the ability to categorize the information deduced from PPI networks

Dynamic PPI networks, once modeled, can provide a closer view of their corresponding cellular events. Accordingly, in contrast with static PPI networks, the information revealed by dynamic networks is at a higher level of details. For instance, in the problem of identifying protein complexes in protein interaction networks, most of the presented algorithms do not differentiate between functional modules and protein complexes. That is mainly due to the absence of embedded information in the networks that could guide the search. In fact, complexes are formed by proteins which interconnect at the same time and place, whereas the members of functional modules may interact at different times and places [31]. Accordingly, when PPIs are bounded by spatiotemporal conditions inferred by gene expression and gene ontology datasets for example, the detected components could more likely be categorized as protein complexes or functional modules. Likewise, dynamic PPI modeling may highly contribute to the detection of protein subcomplexes in PPI networks. Various approaches were developed to solve

this important research problem, but all based on static networks [32]. As dynamic modeling could reveal the mechanisms of protein-complex formation and could yield better complex-detection approaches, it could also provide the same for the detection of subcomplexes.

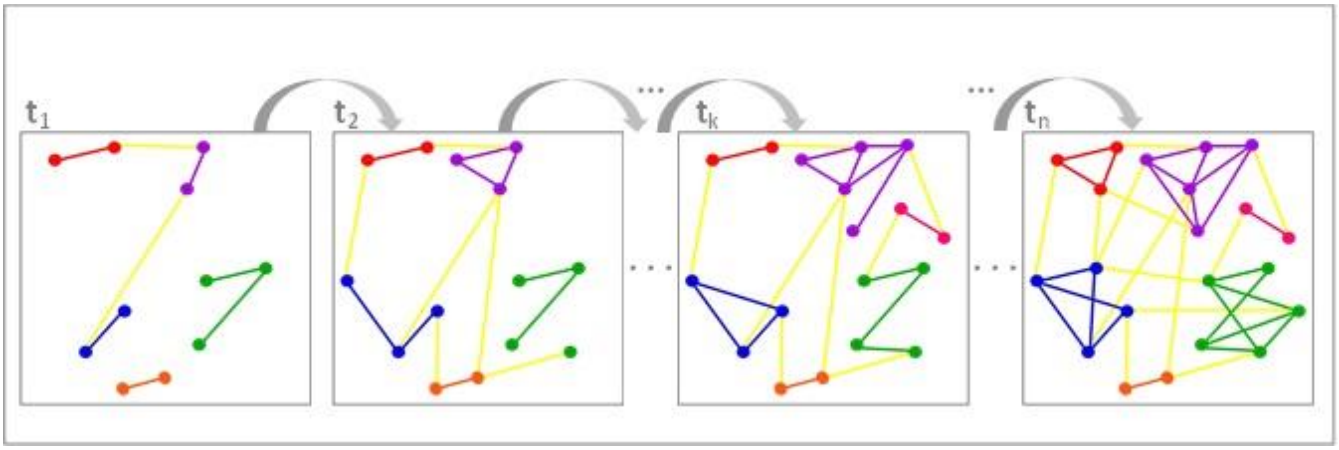

Fig. 1 Snapshots of a hypothetical PPI network, capturing its dynamics at different time points/stages. Each schema includes the available proteins at a certain stage, along with their interconnections. Nodes and edges of similar colors correspond to the same protein complexes, whereas the rest of the edges are represented in yellow.

2.5 Increasing the accountability and the accuracy of the results produced by computational methods

Undeniably, dynamic PPI networks describe cellular interactions in a more realistic manner. Therefore, the computational methods, customized to suit such networks, would certainly produce analytical results with higher accuracy and accountability. Here, we namely consider the algorithms designed to detect protein complexes in protein interaction networks. The integration of temporal, spatial or contextual biological information with PPI data as a means to show the PPI dynamics, can be viewed as a kind of clustering based on temporal, spatial and/or contextual attributes. Hence, the proteins and their interconnections can be grouped based on the integrated conditions and a protein complex-

detection method shall be applied accordingly and with a generalization capability indeed. Once this is achieved, the rates of false positives and false negatives will certainly decrease at the level of the detected complexes and at the level of their protein members as well. Consequently, the overall accuracy of the results will be higher than those scored by methods applied on static networks. The former potentially applies to other exploratory approaches of PPI networks.

3 Modeling Dynamic PPI Networks

A single scheme is usually used to represent a static PPI network with all its components. In contrast, a dynamic PPI network can be visualized by a series of schemes representing snapshots of the network state corresponding to different stages and/or locations of molecular activities, as shown in Fig. 1. The interpretation of a dynamic interaction network and its state transitions depends on the types of data which are used to biologically-condition PPI events. We will hereafter highlight some of the concepts and the approaches to model the dynamicity of protein interaction networks and we will particularly relate them to the problem of detecting protein complexes in PPI networks.

The advancements in experimental techniques are gradually allowing in-depth explorations of biological systems. The resultant progresses can broaden our understanding of biology through the integration of various types of generated information and by consistently developing computational tools to expand our knowledge.

Gene expression datasets are subsequent products which consist of quantitative measurements of RNA species in cellular compartments across different conditions [33]. Genome-wide expression levels can now be studied [34]. Time-series gene expression data

report quantities of RNA across different time points in cellular processes. It is believed that genes with correlated expressions across subsets of conditions most likely interact. When combined with PPI data to model the interaction dynamics, it can potentially reveal the processes which underline the formation of protein complexes. For instance, that was done in [35] where it was shown that a just-in-time mechanism elapsing through continuous time points delineates the formation of most complexes. The statistical 3-sigma principle was then used by the works presented in [35] and [36] to define the active time points of proteins based on their gene expression levels and consequently, introduce approaches to detect and refine protein complexes. The core-attachment view of complexes was recently considered in [37]; based on gene expression data, the identification of a protein complex was split into two main parts: a static core consisting of proteins expressed throughout the whole cell cycle and a short-lived dynamic attachment. The results of these approaches were better than the ones tested on static networks. Kim et al. [38] highlighted some of the computational methods used to infer dynamic networks from expression data based on statistical dependence to classify nodes and/or edges as active or inactive. These methods include: Bayesian networks [39], relevance networks [40], Markov Random Fields [41], ordinary differential equations [42] and logic-based models [43].

As they are conditioned by time, PPIs are space-dependent as well. In other words, the occurrence of a protein interaction is also subject to the co-localization of its interacting partners in cellular components [44]. Actually, a failed interaction caused by inappropriate protein localizations could be pathological. Consequently, subcellular localization annotations [45] can be used to model dynamic PPI networks based on spatial constraints. Indeed, the formation of protein complexes is also influenced by the localization settings of proteins. According to that, it is certainly beneficial to incorporate the spatial dynamics towards improving complex-detection approaches. Various methods aim at studying and collecting spatial movements about proteins [46]. However, in addition to mathematical modeling techniques, further approaches to appropriately integrate spatial protein dynamics in PPI networks are still required.

Gene ontology annotations [47], which provide information about genes that are shared across species, can also infer the dynamics of PPI networks [48]. As an indicator of interaction probability, various weighting schemes were introduced to assign PPI weights based on the similarity degrees of gene ontology terms between interacting partners. Among these approaches are SWEMODE [49], which detects communities within PPI networks based on weighted clustering coefficient and weighted average nearest-neighbors degree measures, and OIIP [48], which is a method to detect protein complexes in PPI networks by assigning node and edge weights based on the size of gene annotations.

Gene expression, spatial annotation and gene ontology annotation data could credibly contribute to the incremental attempts to model dynamic PPI networks.

Forthcoming approaches are expected to profit from these data among other types of biological information. Specifically, the integration of biological attributes enhances the computational methods designed to detect protein complexes in protein interaction networks. It not only participates in uncovering the mechanisms of protein-complex formation but also points out useful details for the design of such methods. In addition, the former may help categorize protein complexes and could be informative regarding their building blocks as well.

4 Datasets and Evaluation Measures

The datasets which could be used to enrich PPI networks in order to model their dynamic aspects, such as gene expression and gene ontology data, typically describe the variations of protein activities and/or quantities across sets of conditions. The resulting network analysis ought to consider these conditions. For example, the detected protein complexes in a PPI network enriched by time-series gene expression data would most likely be adherent to the conditions across which the gene expression data were generated. Therefore, reference protein-complex sets which were used to evaluate previous approaches that work on static networks, such as MIPS [50] and CYC2008 [51], may not be the best choice for dynamic networks. Accordingly, reference sets tailored to match input datasets and their conditionality could be more convenient in such cases. Similarly, issues regarding the choice of evaluation measures arise when shifting to dynamic PPI networks. The formulae used to evaluate the accuracy, sensitivity and specificity in addition to other qualities of previous approaches are not the same [8, 52]. Strong evaluation scores include the number of complexes in the reference catalog that are matched with at least one of the predicted complexes with an overlap score, w, greater than a certain threshold; the

clustering-wise sensitivity (S_n);the clustering-wise positive predictive value (*PPV*); the geometric accuracy (*Acc*); and the maximum matching ratio (*MMR*) which shows how accurately the predicted complexes represent the reference complexes by dividing the total weight of the maximum matching by the number of reference complexes. Given *m* predicted complexes and *n* reference complexes, the corresponding formulae are given by the following equations, where t_{ij} represents the number of proteins that are found in both predicted complex *m* and reference complex *n*.

$$w(A,B) = \frac{|A \cap B|^2}{|A||B|} \qquad (1)$$

$$S_n = \frac{\sum_{i=1}^{n} max_{j=1}^{m} t_{ij}}{\sum_{i=1}^{n} n_i} \qquad (2)$$

$$PPV = \frac{\sum_{j=1}^{m} max_{i=1}^{n} t_{ij}}{\sum_{j=1}^{m} \sum_{i=1}^{n} t_{ij}} \qquad (3)$$

$$Acc = \sqrt{S_n \times PPV} \qquad (4)$$

5 Conclusion

The realization of dynamic protein interaction networks is a natural evolution which leverages computational methods for biology. It could typically be acquired by investing in recent biological data generated by advanced experimental techniques. These data include, but are not limited to, gene expression, subcellular localization annotation and gene ontology terms annotation datasets which provide temporal, spatial and contextual information about protein interactions throughout cellular processes. With emphasis on the algorithms for the detection of protein complexes, by modeling the dynamics of PPI networks, we could: reproduce the mechanisms of protein-complex formation more realistically; potentially uncover new biological facts about complexes; overcome data limitations existing in most experimental datasets; categorize modules deduced from PPI networks; and finally, increase the accuracy and value of the detected results. Accordingly, novel algorithms for the detection of protein complexes in dynamic protein interaction networks are expected to appear.

6 References

[1] Durbin, R.M., Abecasis, G.R., Altshuler, D.L., et al. “A map of human genome variation from population-scale sequencing”. Nature, Vol. 467, 1061–1073, Oct. 2010.

[2] Gavin, A.C., Aloy, P., Grandi, P., et al. “Proteome survey reveals modularity of the yeast cell machinery”. Nature, 440, 631–636, Mar. 2006.

[3] Adelmant, G., and Marto, J.A. “Protein complexes: the forest and the trees”. Expert Rev. Proteomics, 6(1), 5–10, Feb. 2009.

[4] Dongen, S. “Graph clustering by flow simulation”. PhD Thesis. University of Utrecht, Amsterdam, 2000.

[5] Bader, G.D., and Hogue, C.W.V. “An automated method for finding molecular complexes in large protein interaction networks”. BMC Bioinformatics, 4:2, Jan. 2003.

[6] Guimei, L., Wong, L., and Chua, H.N. “Complex discovery from weighted PPI networks”. Bioinformatics, 25(15), 1891 – 1897, May 2009.

[7] Frey, B.J., and Dueck, D. “Clustering by passing messages between data points”. Science, 315(5814):972 – 976, Feb. 2007.

[8] Nepusz, T., Yu, H., and Paccanaro, A. “Detecting overlapping protein complexes in protein-protein interaction networks”. Nature Methods, 9, 471 – 472, Mar. 2012.

[9] King, A.D., Przulj, N., and Jurisica, I. “Protein complex prediction via cost-based clustering”. Bioinformatics, 20(17), 3013 – 3020, June 2004.

[10] Przulj, N., Wigle, D.A., Jurisica, I. “Functional topology in a network of protein interactions”. Bioinformatics, 20(3), 340 – 348, Feb. 2004.

[11] Macropol, K., Can, T., and Singh, A.K. “RRW: repeated random walks on genome-scale protein networks for local cluster discovery”. BMC Bioinformatics, 10:283, Sep. 2009.

[12] Adamcsek, B., Palla, G., Farkas, I.J., et al. “CFinder: locating cliques and overlapping modules in biological networks”. Bioinformatics, 22(8), 1021 – 1023, Apr. 2006.

[13] Zaki, N.M., Berengueres, J., and Efimov, D. “Detection of protein complexes using a protein ranking algorithm”. Proteins: Structure, Function, and Bioinformatics, 80(10), 2459 – 2468, Oct. 2012.

[14] Zaki, N.M., Berengueres, J., and Efimov, D. “Prorank: A method for detecting protein complexes”. In Proceedings of the 14th International Conference on Genetic and Evolutionary Computation Conference (GECCO '12), Philadelphia. Edited by Terence Soule: ACM, New York, 209 – 216, July 2012.

[15] Hanna, E.M., and Zaki, N.M. “ProRank+: A Method for Detecting Protein Complexes in Protein Interaction Networks”. In Proceedings of the 5th International Conference on Bioinformatics Models, Methods and Algorithms (BIOINFORMATICS’14), Angers, Loire Valley, France, Mar. 2014.

[16] Efimov, D., Zaki, N.M., and Berengueres, J. “Detecting protein complexes from noisy protein interaction data”. In Proceedings of the 11th International Workshop on Data Mining in Bioinformatics (BIOKDD '12), 1-7, Aug. 2012.

[17] Zaki, N.M., Dmitry, D., and Berengueres, J. “Protein Complex Detection using Interaction Reliability Assessment

and Weighted Clustering Coefficient". BMC Bioinformatics, 14:163, May 2013.

[18] Fields, S., and Song, O. "A novel genetic system to detect protein–protein interactions". Nature, 340, 245 – 246, July 1989.

[19] Collins, M.O., and Choudhary, J.S. "Mapping multiprotein complexes by affinity purification and mass spectrometry". Curr. Opin. Biotechnol., 19, 324 – 330, Aug. 2008.

[20] Kim, T.H., and Ren, B. "Genome-wide analysis of protein–DNA interactions". Annu. Rev. Genomics Hum. Genet., 7, 81 – 102, Sep. 2006.

[21] Johnson, D.S., Mortazavi, A., Myers, R.M., et al. "Genome-wide mapping of in vivo protein–DNA interactions". Science, 316, 1497 – 1502, June 2007.

[22] Levy, E.D., and Pereira-Leal, J.B. "Evolution and dynamics of protein interactions and networks". Curr. Opin. Struct. Biol., 18, 349 – 357, June 2008.

[23] Przytycka, T.M., Singh, M., Slonim, D.K. "Toward the dynamic interactome: it's about time". Briefings in Bioinformatics, 11, 15 – 29, Jan. 2010.

[24] Han, J.D., Bertin, N., Hao, T., et al. "Evidence for dynamically organized modularity in the yeast proteinprotein interactionnetwork". Nature, 430, 88 – 93, July 2004.

[25] Brun, C., Chevenet, F., Martin, D., et al. "Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network". Genome Biol., 5(1):R6, Dec. 2003.

[26] Chua, H., Ning, K., Sung, H.W., et al. "Using indirect protein-protein interactions for protein complex prediction". J. Bioinform. Comput. Biol., 6, 435 – 466, Jan. 2008.

[27] Hon, N.C., Sung, W.K., and Wong, L. "Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions". Bioinformatics, 22, 1623 – 1630, July 2006.

[28] Chen, J., and Yuan, B. "Detecting functional modules in the yeast protein-protein interaction network". Bioinformatics, 22, 2283 – 2290, July 2006.

[29] Xu, B., Lin, H., and Yang, Z. "Ontology integration to identify protein complex in protein interaction networks". Proteome Sci., 9:S7, Oct. 2011.

[30] Von Mering, C., Krause, R., Snel, B., et al. "Comparative assessment of large-scale data sets of protein-protein interactions". Nature, 417, 399 – 403, May 2002.

[31] Spirin, V., and Mirny LA. "Protein complexes and functional modules in molecular networks". PNAS, 100, 12123 – 12128, Oct. 2003.

[32] Zaki, N.M., and Mora, A. "A comparative analysis of computational approaches and algorithms for protein subcomplex identification". Scientific Reports, 4: 4262, nature group, Mar. 2014.

[33] Lovén, J., Orlando, D.A., Sigova, A.A., et al. "Revisiting global gene expression analysis". Cell, 151(3), 476 – 482, Oct. 2012.

[34] Secrier, M., and Schneider, R. "Visualizing time-related data in biology, a review". Briefings in Bioinformatics, Software Review, Apr. 2013.

[35] Wang, J., Peng, X., Xiao, Q., et al. "An effective method for refining predicted protein complexes based on protein activity and the mechanism of protein complex formation". BMC Systems Biology, 7:28, Mar. 2013.

[36] Wang, J., Peng, X., Li, M. "Active Protein Interaction Network and Its Application on Protein Complex Detection". In Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 37 – 42, Nov. 2011.

[37] Li, M., Chen, W., Wang, J., et al. "Identifying dynamic protein complexes based on gene expression profiles and PPI Networks". Biomed Research International, Mar. 2014.

[38] Kim, Y., Han, S., Choi, S., et al. "Inference of dynamic networks using time-course data". Briefings in Bioinformatics, 15(2), 212 – 228, May 2013.

[39] Friedman, N., Linial, M., Nachman, I., et al. "Using Bayesian networks to analyze expression data". J. Comp. Biol., 7, 601 – 620, July 2004.

[40] Remondini, D., O'Connell, B., Intrator, N., et al. "Targeting c-Myc-activated genes with a correlation method: detection of global changes in large gene expression network dynamics". PNAS, 102, 6902 – 6906, May 2005.

Song, L., Kolar, M., and Xing, E.P. "KELLER: estimating time-varying interactions between genes". Bioinformatics, 25, i128 – i136, June 2009.

[41] Bansal, M., Belcastro, V., Ambesi-Impiombato, A., et al. "How to infer gene networks from expression profiles". Mol. Syst. Biol, 3:122, Feb. 2007.

[42] Morris, M.K., Saez-Rodriguez, J., Sorger, P.K., et al. "Logic-based models for the analysis of cell signaling networks". Biochemistry, 49, 3216 – 3224, Mar. 2010.

[43] Park, S., Yang, J.S., Shin, Y.E., et al. "Protein localization as a principal feature of the etiology and comorbidity of genetic diseases". Mol. Syst. Biol., 7:494, May 2011.

[44] De Lichtenberg, U., Jensen, L.J., Brunak, S., et al. "Dynamic complex formation during the yeast cell cycle". Science, 307, 724 – 727, Feb. 2005.

[45] Lee, Y.H., Tan, H.T., and Chung, M.C. "Subcellular fractionation methods and strategies for proteomics". Proteomics, 10, 3935 – 3956, Nov. 2010.