Computer Science and Application 计算机科学与应用, 2020, 10(12), 2419-2424
Published Online December 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.1012256

Hans汉斯

基于动态网络的蛋白质复合物识别研究综述

于　杨，孔德宙

沈阳师范大学软件学院，辽宁　沈阳
Email: yuyangsd1204@126.com

收稿日期：2020年11月29日；录用日期：2020年12月23日；发布日期：2020年12月30日

摘　要

真实细胞系统中蛋白质及其相互作用是随时变化的，具有一定的动态性。构建蛋白质–蛋白质相互作用(PPI)动态网络并识别蛋白质复合物是揭示细胞功能关键问题。本文介绍了近年来基于动态网路的复合物识别的研究方法，分析了此类复合物预测方法所面临的挑战。

关键词

动态网络，复合物识别，基因表达，聚类

Review of Computational Method for Protein Complex Identification Based on Dynamic PPI Networks

Yang Yu, Dezhou Kong

Software College, Shenyang Normal University, Shenyang Liaoning
Email: yuyangsd1204@126.com

Received: Nov. 29th, 2020; accepted: Dec. 23rd, 2020; published: Dec. 30th, 2020

Abstract

In real cell system, proteins and their interactions are always changing and dynamic. It is one of key issues to constructing (PPI) dynamic network and to identify protein complexes for understanding cell function. We summarize the research methods of identification based on dynamic network in recent years, and analyze the challenges in this field.

文章引用: 于杨, 孔德宙. 基于动态网络的蛋白质复合物识别研究综述[J]. 计算机科学与应用, 2020, 10(12): 2419-2424.
DOI: 10.12677/csa.2020.1012256

Keywords

Dynamic Network, Protein Complex Identification, Gene Expression, Clustering

1. 引言

蛋白质复合物是蛋白质相互作用(PPI)网络中刻画细胞内许多生物活性的基本模块，是由一组具有相同功能的活动蛋白质组成，具有动态性[1]。它能实现各类细胞过程，进而完成各种各样的生物功能。因此，识别蛋白质复合物有助理解分子功能和致病基因的研究。从 PPI 中识别蛋白质复合物是现代生物信息学研究的热点问题之一。较早的研究主要是基于生物实验的方法识别复合物，但是此类方法代价较高。随着生物实验技术、文本挖掘技术和其它预测技术的发展，大量的(PPI)数据不断涌现。这为研究人员从 PPI 网络中检测蛋白质复合物的方法提供了方便的数据支持。

PPI 网络随着环境、时间和细胞周期的不同阶段不断变化，早期的大量蛋白质复合物挖掘算法主要基于蛋白质相互作用静态网络的，忽略了蛋白质之间发生相互作用时具有时空动态特性，因而导致不能有效反映细胞内蛋白质的动态变化，影响蛋白质复合物识别算法的性能[2]。因此，研究从静态 PPI 网络向动态 PPI 网络转变对于蛋白质复合物的准确识别至关重要。本文将从近 10 年基于动态网络的蛋白质复合物预测的方法进行综述，并对蛋白质复合物预测所面临的挑战与未来的研究方向进行探讨。试图为基于 PPI 动态网络的复合物识别勾画出一个清晰的概况，希望对此方面的研究提供有益的参考。

2. 基于动态网络的复合物预测方法

动态性是蛋白质的固有属性，它们之间的彼此作用都会受到外界刺激或条件改变而变化，因而首先要对 PPI 动态网络进行建模。在大部分的此类研究中，动态蛋白质网络通常用一个无向图表示，即 $DG=\{G_1, G_n, \cdots, G_n\}$ 表示，其中 $G_i=(V_i, E_i)$ 表示第 i 时刻或者条件下的子网，V_i 表示图中第 i 时刻或者条件下的节点的集合，E_i 表示相应的边的集合。基于动态网络的蛋白质复合物主要分成四个部分(见图 1)，PPI 和其它生物信息的融合、基于各种策略的 PPI 动态网络构建、有效地聚类算法设计和聚类结果的优化处理。本部分将以时间顺序介绍近 10 年复合物识别的研究现状。

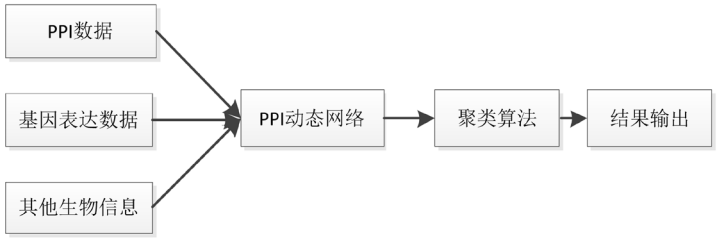

Figure 1. The process of complex identification based on dynamic network
图 1. 基于动态网络的复合物识别基本流程

2011 年，汤等人[3]利用时间序列基因表达数据和 PPI 网络构造了基于时序过程的 PPI 网络(TC-PINs)。该方法首先利用 PPI 网络中存在直接相互作用的两个蛋白质节点，若它们在某一时间点的基因表达值都

超过某一固定阈值，则认为它们在此时间点具有共表达的特性，基于此构建动态 PPI 网络，并应用于复合物的识别。

2012 年文献[4]基于静态 PPI 网络是由不同时间不同空间的叠加这种不足，提出了利用 3sigma 原则将获得的不同时刻蛋白质的活跃点集英社到静态 PPI 网络中，进而构建动态 PPI 网络进行复合物识别。

如果一个蛋白质处于活跃的时间点对应于基因的表达水平处于高峰点。基于这一假设，2013 年 Wang 等人[5]基于 3-sigma 准则根据基因表达曲线的自身特性为每个基因单独设置阈值，将基因表达信息加入到静态网络对动态相互作用网络进行建模，并从中预测复合物。实验证明蛋白质复合物的预测的结果比证明比在 TC-PINs 和原始的静态网络都更准确有效。

2014 年，李敏等人结合用基因表达谱数据和 PPI 数据构建一种高质量的动态网络，将设计的基于核－附属复合物识别算法应用于此网络[6]。Ou-Yang 等通过整合蛋白质相互作用数据和基因表达数据，通过检测稳定相互作用和瞬时相互作用来构建一系列动态 PPI 网络。在不同的时间点上，稳定的相互作用被保留起来作为蛋白质相互作用网络的主干，而在某个时间点上是否存在短暂的相互作用取决于两个相关蛋白所需的特定活性和功能，并将提出了一种新的时间平滑重叠复合物检测模型(TS-OCD)应用于复合物识别[7]。

2015 年胡等人在研究已有的动态网络构建基础之上，为基因表达周期的 12 个时刻构建相互作用网络，每三个周期的基因表达的平均值分别赋给 12 个时刻的基因表达值，改进了 36 个时刻的动态网络的构建方式；然后结合蛋白质的结构阈信息，补充缺失的相互作用信息对动态网络建模，并将提出复合物预测算法应用其中，提升了复合物的识别的综合性能[8]。

2016 年张等人首先利用基因表达数据计算每个蛋白质和 PPI 的活性时间点和活性概率，然后将动态活动信息集成到高通量的 PPI 数据中构建动态 PPI 网络。其次，提出了一种基于核心－附属结构特征的动态蛋白质复合物预测方法应用于动态网络模型[9]。

2017 年赵等人首先根据基于 3-sigma 准则，分别结合基因表达数据，PPI 数据和 GOTerms 数据中 PCC、ECC 和 GSM 三个技术指标的构建动态蛋白质相互作用网络。其中基因表达数据中有三个连续的代谢周期，每个周期有 12 个时间点，因此构建 12 个动态子网络。最后利用改进的布谷鸟搜索算法识别复合物[10]。沈等人[11]结合基因表达和静态网络构建动态网络，并将复合物内部结构应用到动态网络中识别核复合物。梁等人结合基因表达数据和静态 PPI 数据构建 12 个动态子网络，提出基于蚁群优化算法进行基于核-附属的复合物识别方法[12]。

2018 年 Janani 等人提出一种从基因表达矩阵中取出动态 PPI 子网络的新技术，使用一种称为混合蛙跳算法的优化方法从输入基因表达矩阵中识别双聚类，该算法采用一种新的适应度函数——离散公共评分(DCS)来评价双聚类。其次，对于每一个双聚类，将由双聚类中的基因集合组成的子图作为一个动态的 PPI 子网络，采用 PCD-DPPI 方法对蛋白质复合物识别。刘等人[13]首先将不同时间点的基因表达数据与传统的静态 PPI 网络相结合构建不同的动态子网络。其次，将基于基因本体的语义相似度作为网络权重与主成分分析法结合滤除数据噪声，并进行权重计算，并将提出得基于核－附属的结构特征的蛋白质复合物预测算法应用于动态网络。

2019 年张等人首先利用基本基因表达水平以上的波动度代替基因表达水平来确定蛋白质的活性时间点，然后结合 PPI 和基因表达信息构建的时序 PPI 网络(TPNs)，最后提出了一种融合多源生物数据的复合物识别方法[14]。Rani 等人首先通过整合基因表达数据将静态 PPI 数据转换为动态 PPI 数据，然后在每个子网络应用 MCL 和象群优化算法进行网络聚类识别蛋白质复合物[15]。

2020 年 SabziNezhad 等人利用 TAP 和 GO 数据来生成一个加权 PPI 网络进而降低 PPI 中的噪声数据，同时通过基因表达数据将 PPI 网络生成的动态子网络，然后采用模因算法对基因表达数据进行双聚类，并为每个双聚类集合创建一个动态子网络，最后识别蛋白复合物[16]。Chellal 等人[17]利用基因表达方差

信息和共表达相关信息计算动态 PPI 网络中每个 PPI 的活动时间点和活动概率，构建动态 PPI 网络，算法采用逐层检测核心簇，结合哈里斯鹰算法蛋白质复合物。

3. 常用数据库

复合物识别的常用数据集，如下表 1 所示。

Table 1. Related databases
表 1. 常用数据库

分类	数据库名字	URL
蛋白质相互作用数据库	BioGRID [18]	https://thebiogrid.org
	IntAct [19]	http://www.ebi.ac.uk/intact
	DIP [20]	http://dip.doe-mbi.ucla.edu
	MINT [21]	https://mint.bio.uniroma2.it
	HRPD [22]	http://www.hprd.org/
	MIPS [23]	http://mips.helmholtz-muenchen.de/proj/ppi
	STRING [24]	http://string－db.org/
蛋白质复物数据库	CYC2008 [25]	http://wodaklab.org/cyc2008/
	CORUM [26]	https://mips.helmholtz-muenchen.de/corum
基因表达数据	GSE3431 [27]	http://www.ncbi.nlm.nih.gov/geo/
	GSE4987 [27]	

4. 问题与研究展望

近年来虽然基于动态网络的蛋白质复合物识别的研究算法得到了一定的改进，但由于相关生物概念的了解不够深入，PPI 数据的不完整性，基因表达数据的噪声，因此在以下几个方面仍有待提高，需要深入的研究。

1) 构建真实可靠的动态网络模型

一方面在动态网络中刻画低表达的蛋白质，弥补基于阈值方法的不足；另一方面有效控制基因表达的噪声，补充的基于原则的方法的不足。

2) 融合多源生物信息的动态网络构建

虽然基因表达数据的时序性对于构建动态网络模型具有一定的帮助，但是实际蛋白质复合物的形成过程还要受多个因素的影响，例如亚细胞定位信息、空间信息、结构信息和其他的时序数据等。因此今后的研究可以从多数据源分析，整合相关信息，构建多层次多维度的动态网络模型。

3) 有效聚类算法的设计

一方面针对多层次多维度的 PPI 动态网络模型，设计快速有效的复合物聚类算法；另一方面挖掘更多的生物数据的内在的特征，结合复合物现有的多种拓扑形态和不同生物特征，设计基于节点多维向量化的聚类算法尤为重要。

4) 人类疾病研究的应用

现有研究表明疾病的发生与一组异常的基因有关，这些基因构成基因网络。因而复合物预测的方法可以用在人类 PPI 网络和致病基因上，进而分析动态的基因模块，识别关键基因与人类疾病的关系，为疾病的研究提供技术支持。

5) 大数据应用

随着高通量技术的发展，产生大量的相互作用数据以及多源的生物其他数据，这使源数据和数据库存在异质性，导致数据缺失、数据矛盾等问题，因而需要依靠大数据思维和数据分析技术对蛋白质相互作用多源数据进行深入挖掘和融合，进而有效预测蛋白质复合物。

基金项目

辽宁省自然科学基金(20180550918)。

参考文献

[1] Banzhaf, M. and Typas, A. (2014) Dynamic Protein Complexes for Cell Growth. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 4355-4356. https://doi.org/10.1073/pnas.1402016111

[2] Srihari, S. and Leong, H.W. (2012) Temporal Dynamics of Protein Complexes in PPI Networks: A Case Study Using Yeast Cell Cycle Dynamics. *BMC Bioinformatics*, **13**, Article No. S16. https://doi.org/10.1186/1471-2105-13-S17-S16

[3] Tang, X., Wang, J., Liu, B., Li, M., Chen, G. and Pan, Y. (2011) A Comparison of the Functional Modules Identified from Time Course and Static PPI Network Data. *BMC Bioinformatics*, **12**, Article No. 339. https://doi.org/10.1186/1471-2105-12-339

[4] 彭小清. 动态蛋白质网络的构建和蛋白质复合物识别研究[D]: [硕士学位论文]. 长沙: 中南大学, 2012.

[5] Wang, J., Peng, X., Li, M. and Pan, Y. (2013) Construction and Application of Dynamic Protein Interaction Network Based on Time Course Gene Expression Data. *Proteomics*, **13**, 301-312. https://doi.org/10.1002/pmic.201200277

[6] Li, M., Chen, W., Wu, J., Wu, F. and Pan, Y. (2014) Identifying Dynamic Protein Complexes Based on Gene Expression Profiles and PPI Networks. *Biomed Research International*, **2014**, Article ID: 375262. https://doi.org/10.1155/2014/375262

[7] Ou-Yang, L., Dai, D.Q., Li, X.L., Wu, M., Zhang, X.F. and Yang, P. (2014) Detecting Temporal Protein Complexes from Dynamic Protein-Protein Interaction Networks. *BMC Bioinformatics*, **15**, Article No. 335. https://doi.org/10.1186/1471-2105-15-335

[8] 胡赛熊, 赵碧海, 李学勇, 王晶. 动态加权蛋白质相互作用网络构建及其应用研究[J]. 自动化学报, 2015, 41(11): 1893-1900.

[9] Zhang, Y., Lin, H., Yang, Z., Wang, J., Liu, Y. and Sang, S. (2016) A Method for Predicting Protein Complex in Dynamic PPI Networks. *BMC Bioinformatics*, **17**, Article No. 229. https://doi.org/10.1186/s12859-016-1101-y

[10] Zhao, J., Lai, X.J. and Wu, F.X. (2017) Predicting Protein Complexes in Weighted Dynamic PPI Networks Based on ICSC. *Complexity*, **2017**, Article ID: 4120506. https://doi.org/10.1155/2017/4120506

[11] Shen, X.J., Yi, L., Jiang, X.P., He, T.T., Yang, J.C., Xie, W., Hu, P. and Hu, X.H. (2017) Identifying Protein Complex by Integrating Characteristic of Core-Attachment into Dynamic PPI Network. *Plos One*, **12**, e0186134. https://doi.org/10.1371/journal.pone.0186134

[12] Liang, J., Lei, X., Guo, L. and Tan, Y. (2018) ACO Based Core-Attachment Method to Detect Protein Complexes in Dynamic PPI Networks. In: Tan, Y., Shi, Y. and Tang, Q., Eds., *Advances in Swarm Intelligence. ICSI* 2018. *Lecture Notes in Computer Science*, Vol 10941, Springer, Cham, 101-112. https://doi.org/10.1007/978-3-319-93815-8_11

[13] Liu, L., Sun, X., Song, W. and Du, C. (2018) A Method for Predicting Protein Complexes from Dynamic Weighted Protein-Protein Interaction Networks. *Journal of Computational Biology*, **25**, 586-605. https://doi.org/10.1089/cmb.2017.0114

[14] Zhang, J., Zhong, C., Lin, H.X. and Wang, M. (2019) Identifying Protein Complexes from Dynamic Temporal Interval Protein-Protein Interaction Networks. *BioMed Research International*, **2019**, Article ID: 3726721. https://doi.org/10.1155/2019/3726721

[15] Rani, R.R., Ramyachitra, D. and Brindhadevi, A. (2019) Detection of Dynamic Protein Complexes through Markov Clustering Based on Elephant Herd Optimization Approach. *Scientific Reports*, **9**, Article No. 11106. https://doi.org/10.1038/s41598-019-47468-y

[16] Sabzinezhad, A. and Jalili, S. (2020) DPCT: A Dynamic Method for Detecting Protein Complexes From TAP-Aware Weighted PPI Network. *Frontiers in Genetics*, **11**, 567.

[17] Yao, H., Guan, J. and Liu, T. (2020) Denoising Protein-Protein Interaction Network via Variational Graph Auto-Encoder for Protein Complex Detection. *Journal of Bioinformatics and Computational Biology*, **18**, 2040010. https://doi.org/10.1142/S0219720020400107

[18] Andrew, C.A., Rose, O., Lorrie, B., Jennifer, R., Christie, C., Kolas, N.K., Lara, O.D., Sara, O., Chandra, T. and Adnane, S. (2017) The BioGRID Interaction Database: 2017 Update. *Nucleic Acids Research*, **45**, D369-D379.

[19] Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C. and Del-Toro, N. (2013) The MIntAct Project—IntAct as a Common Curation Platform for 11 Molecular Interaction Databases. *Nucleic Acids Research*, **42**, D358-D363. https://doi.org/10.1093/nar/gkt1115

[20] Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K. and Bowie, J.U. (2004) The Database of Interacting Proteins: 2004 Update. *Nucleic Acids Research*, **32**, D449-D451. https://doi.org/10.1093/nar/gkh086

[21] Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A.P. and Santonico, E. (2012) MINT, the Molecular Interaction Database: 2012 Update. *Nucleic Acids Research*, **40**, D857-D861. https://doi.org/10.1093/nar/gkr930

[22] Keshava Prasad, T., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R. Shafreen, B. and Venugopal, A. (2008) Human Protein Reference Database—2009 Update. *Nucleic Acids Research*, **37**, D767-D772. https://doi.org/10.1093/nar/gkn892

[23] Mewes, H.W., Hani, J., Pfeiffer, F. and Frishman, D. (1998) MIPS: A Database for Protein Sequences and Complete Genomes. *Nucleic Acids Research*, **26**, 33-37. https://doi.org/10.1093/nar/26.1.33

[24] Szklarczyk, D., Franceschini, A., Wyder, S., *et al.* (2015) STRING v10: Protein-Protein Interaction Networks, Integrated over the Tree of Life. *Nucleic Acids Research*, **43**, D447-D452. https://doi.org/10.1093/nar/gku1003

[25] Pu, S., Wong, J., Turner, B., Cho, E. and Wodak, S.J. (2008) Up-to-Date Catalogues of Yeast Protein Complexes. *Nucleic Acids Research*, **37**, 825-831. https://doi.org/10.1093/nar/gkn1005

[26] Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C. and Mewes, H.-W. (2009) CORUM: The Comprehensive Resource of Mammalian Protein Complexes—2009. *Nucleic Acids Research*, **38**, D497-D501. https://doi.org/10.1093/nar/gkp914

[27] Tu, B.P., Kudlicki, A., Rowicka, M. and McKnight, S.L. (2005) Logic of the Yeast Metabolic Cycle: Temporal Compartmentalization of Cellular Processes. *Science*, **310**, 1152-1158. https://doi.org/10.1126/science.1120499