文章编号：0258－5898(2009)04－0465－05

·综 述·

蛋白质相互作用及互作网络的生物信息学分析

谢 超，郜 尽，袁运生 综述 俞 雁 审校
（上海交通大学 农业与生物学院 上海兽医生物技术重点实验室，上海 200240）

摘 要：后基因组时代的最终目标是认识真正执行生命活动的全部蛋白质的表达规律和生物学功能，即蛋白质组学研究。关键的挑战之一就是对蛋白质相互作用网络的研究，将有助于从系统角度加深对细胞结构和功能的认识，并且为新药靶位点的发现和药物设计提供理论基础。目前，用生物信息学的手段来研究蛋白质相互作用网络显示出巨大的优势，主要包括蛋白质相互作用网络的绘制与显示、网络的拓扑结构分析、网络结构模块的研究及网络的比对等。文章试图从生物信息学的角度认识蛋白质相互作用，并总结蛋白质相互作用网络的生物信息学分析方法。

关键词：蛋白质相互作用；蛋白质相互作用网络；生物信息学
中图分类号：R318.5 文献标志码：A

Protein-protein interactions and their network analysis in bioinformatics

XIE Chao, GAO Jin, YUAN Yun-sheng reviewer YU Yan reviser
(Shanghai Key Laboratory of Veterinary Biotechnology, School of Agriculture and Biology, Shanghai Jiaotong University, Shanghai 200240, China)

Abstract: The ultimate goal of post-genome research is to understand a complete set of proteins in a living organism for their expression pattern and biological function, which is called proteomics. One of the major challenges in proteome research is to study the protein-protein interactions, and the emerging bioinformatics approaches present us tremendous advantages when dealing with protein interaction networking and data analysis. Useful bioinformatics tools include protein-protein interaction network mapping, topology of the network, structure of the module and comparison of the network. The technology advancement in this field brings further understandings to the structure and function of cells at the proteome level, which may eventually lead to the discovery of new drug targets and design methods. This paper attempts to review the current researches on protein-protein interaction with an emphasis on bioinformatics intervention, and also summarizes some widely used methods for network analysis.

Key words: protein-protein interactions; protein-protein interactions networks; bioinformatics

当人类基因组计划完成时，人类对生命世界的理性认识达到了前所未有的深度与广度，然而人们也清楚地认识到一项更为艰巨的任务即蛋白质组学的研究已经摆到面前。蛋白质组学是研究细胞内所有蛋白质及其动态变化规律的科学，其目标是为了阐明生物体全部蛋白质表达和功能模式。蛋白质组学研究热点之一是蛋白质相互作用，主要包括揭示蛋白质之间相互作用，建立蛋白质相互作用关系的网络图并且对网络进行生物信息学分析。在某种程度上可以说，细胞进行生命活动是蛋白质在一定条件下相互作用的结果，若蛋白质相互作用网络被破坏或稳定性丢失，会引起细胞的功能性障碍。

随着蛋白质相互作用实验技术平台的发展，积累了大量蛋白质相互作用的实验数据，如何分析和处理这些数据成为关键。我们已经获得若干模式生物的蛋白质相互作用网络，而生物信息学的发展又极大促进了以蛋白质组学支撑的技术平台的开发，便于更深层次地挖掘蛋白质相互作用网络中的特性。研究蛋白质相互作用的最终目标就是建立模式生物中全部蛋白质相互作用的网络，阐明蛋白质相互作用的完整网络结构。在某种程度上可以说，蛋白质相互作用网络是功能基因组学、蛋白质组学和

基金项目："十五"国家科技攻关项目(2004BM711A19)("Tenth Five-Year" National Scientific and Technological Project, 2004BM711A19)。
作者简介：谢 超(1982－)，男，博士生；电子信箱：xiechao@sjtu.edu.cn。
通讯作者：俞 雁，电子信箱：yanyu@sjtu.edu.cn

系统生物学的桥梁。目前,已有蛋白质相互作用应用于实践的报道,如 Archakov等[1]已经将蛋白质相互作用应用于药物开发。

1 蛋白质相互作用

1.1 蛋白质相互作用的研究进展 蛋白质相互作用是指蛋白质在其生命周期中相互接触(或短暂接触后分离)后形成特异的复合体共同起作用,也就是说相互作用的蛋白质参与同一个代谢途径或生物学过程。蛋白质相互作用可以分为物理相互作用和功能意义相互作用。

目前,若干模式生物如酿酒酵母(*Saccharomyces cerevisiae*)、果蝇(*Drosophila melanogaster*)和线虫(*Caenorhabditis elegans*)的大规模蛋白质相互作用的网络相继构建成功,并且有人类蛋白质相互作用网络的报道。

Uetz等[2]用阵列法研究了大约 6 000个酵母转化株的开放阅读框(open reading frame, ORF),由于存在许多自激活现象,最终得到了 87个蛋白质之间的 281对蛋白质相互作用。同时还将酵母的 5 341个 ORF作为诱饵和 6 000个 ORF作为猎物,最终得到 817个蛋白质之间的 691对相互作用,其中 88对相互作用是已经报道过的。Ito等[3]构建了含酵母约 92% ORF的诱饵载体和猎物载体,分别形成诱饵载体亚库和猎物载体亚库(各 60个亚库),初次筛选进行 430次亚库交配后发现 175对蛋白质相互作用,其中有 163对是新发现的蛋白质相互作用;随后将所构建的全部亚库检测相互作用,共检测到 3 278个蛋白质的 4 549对相互作用。但是实验结果与 Uetz等[2]获得实验数据重叠的很少。Ho等[4]用高通量蛋白复合物质谱鉴定技术(high-throughput mass spectrometric protein complex identification, HMS-PCI)获得 3 617对蛋白质相互作用,并且该方法的假阳性率低于通过酵母双杂交(yeast two hybrid, Y2H)技术所获得的数据。Giot等[5]通过 Y2H法对果蝇 cDNA文库进行筛选,获得 7 048个蛋白质的 20 405对相互作用,并且通过严格的计算方法限制后获得一个置信度较高的蛋白质相互作用网络,其中包括 4 679个蛋白质和 4 789对相互作用。Li等[6]采用 Y2H技术对线虫的蛋白质相互作用进行研究,得到 4 027对蛋白质相互作用。随后 Li等[6]选取 Y2H方法筛选所得到的部分相互作用进行亲和纯化法验证,并且进行更深入的分析。

在模式生物研究基础上发现,几乎所有的蛋白质都会与其他蛋白质结合,这种现象出现在高等哺乳动物中的频率又高于低等动物。Stelzl等[7]首次采用 Y2H法大规模研究人类蛋白质相互作用,研究用 4 456个诱饵载体和 5 632个猎物载体进行矩阵交配试验,鉴定阳性克隆后从 1 705种蛋白质中发现 3 186对蛋白质相互作用,第一次构建了较为全面的人类蛋白质相互作用网络;并且采用了免疫共沉淀(co-immunoprecipitation, Co-IP)法验证了 Y2H的可靠性。Rual等[8]也使用酵母双杂交法对人的 8 100个 ORF进行研究,将诱饵载体编号分配形成 45个亚库,每个诱饵亚库进行交配,发现存在 2 800对相互作用;随后也采用了亲和纯化实验验证部分酵母双杂交实验结果,进一步评估所得相互作用真实性。Parrish等[9]用 Y2H技术发现并且构建了一个由 770个蛋白质组成的相互作用网络,这个网络涉及蛋白质均与运动性共济失调相关。这说明 Y2H技术已经不仅仅是研究蛋白质相互作用实验方法之一,而且已经在阐述特定疾病相关蛋白或某种信号通路的确定过程中发挥特定作用。

1.2 蛋白质相互作用数据的收集和整理

1.2.1 实验获得: 目前,广泛应用于鉴定蛋白相互作用的技术主要可以分成以下四类。①基于文库的方法:主要优点是高度并行的实验格式、候选相互作用的蛋白质及其 cDNA之间的直接关联。缺点是存在较高的假阳性和假阴性率,所以需要进一步实验证据验证蛋白质相互作用。方法有在转录水平上进行的 Y2H及其相关技术、标准的文库表达和噬菌体展示文库(phage display library),但是这些方法都是在体外进行实验,蛋白质折叠和识别情况并不能反映细胞内情况。②亲和性方法:检测蛋白质相互作用的物理方法通常依赖一个蛋白质对另一个蛋白质的亲和,这些方法可以识别直接或间接的蛋白质相互作用。主要包括在体外验证蛋白质相互作用的谷胱甘肽 S转移酶融合蛋白的沉降技术、免疫共沉淀和蛋白质微阵列法等。③分子和原子的方法:X射线晶体学和核磁共振技术有助于在原子水平识别蛋白质相互作用;蛋白质相互作用的分子方法包括:荧光共振能量转移、表面基元共振谱、蛋白质复合体的质谱技术、原子力显微镜技术、BIAcore表面等离子体共振分析技术以及石英体微平衡生物传感器技术。

④遗传方法: 对于细菌和酵母等可以控制处理遗传的物种, 遗传方法也可以分析蛋白质相互作用。主要有抑制子突变、合成致死性筛选和显性负突变等。

1.2.2 预测获得: 蛋白质相互作用数据除了通过实验获得外, 还可以应用生物信息学手段对蛋白质相互作用进行预测, 通过模拟和计算方法获得的蛋白质相互作用数据比通过实验获得数据快很多。①基于基因组上下文关系: 主要方法包括系统发生谱法、基因邻居法、基因融合和转录谱。②基于蛋白质序列系统发生过程: 主要方法包括镜像树、关联突变、同源模建、多体串线、支持向量机和贝叶斯网络。③基于蛋白质结构的方法: 综合多种方法进行预测。

1.3 蛋白质相互作用数据评估及数据库建立 在同一物种中, 不同的高通量实验方法得到的蛋白质相互作用数据之间很难彼此覆盖。von Mering等[10]也指出, 高通量蛋白质相互作用数据比小规模的数据有着更高的假阳性率, 因此必须提高对假阳性的辨识能力, 从噪音数据中区分出真实的蛋白质相互作用。另一方面, 必须提高检验和分析方法的灵敏度, 以避免丢失实际存在的真实相互作用, 特别是在信号转导和代谢调控等过程中出现的瞬态相互作用[11]。

对蛋白质相互作用进行评价的标准包括以下几方面。Mrowka等[12]根据蛋白质在mRNA的表达水平相关性来评价两个蛋白质相互作用的可靠性。Deane等[13]则用旁系证实方法(paralogous verification method PVM)来估计蛋白质相互作用的可靠性, 但是这种方法敏感性不高。此外, 还可以应用蛋白质相互作用网络的拓扑性、mRNA水平的微阵列实验所提供的共表达信息及共有的细胞通路或定位信息等。在以上评价的基础上, 可进行蛋白质相互作用的筛选, 得到高置信度的蛋白质相互作用网络。

大多数蛋白质相互作用具有更广泛的网状结构, 随着网络结构不断扩展, 需要更高级的数据库和数据采集工具来提取有用的信息。生物信息学快速发展的一个优势就是可以提供蛋白质相互作用的数据库, 其中可以存储、查找蛋白质相互作用数据, 评估相互作用数据可靠性等。综合大规模和小规模的蛋白质相互作用数据, 蛋白质相互作用的数据库应运而生。BIND(http://bind.ca)在一个面对对象的数据库里收录了各种生物分子之间的相互作用, 数据库中的内容来自高通量的实验数据提交以及手动从科技文献中整理出的蛋白质相互作用数据; BIND将生物分子(包括蛋白质、核酸和小配体)之间相互作用分为相互作用、复合体和通路三类。DIP(http://dip.doe-mbi.ucla.edu)数据库的蛋白质相互作用是经过人工审核和计算方法验证后加入数据库的; DIP是采用计算方法分析蛋白质相互作用的优秀数据库; 同时, DIP发展了全基因组范围的数据质量监测工具, 保证了数据的可信性。MIPS(http://mips.gsf.de)是一个包括酵母和哺乳动物的蛋白质相互作用的数据库, 可靠性和准确性很高, 其中包括了子数据库MPPI(MIPS mammalian protein-protein interaction database)。BioGRID(http://www.thebiogrid.org)是一个由酵母蛋白质相互作用、线虫蛋白质相互作用、果蝇蛋白质相互作用以及人类蛋白质相互作用共4个子库的数据库。因此, 蛋白质互作数据库是蛋白质相互作用研究水平的标志[14]。

计算方法对高通量蛋白质相互作用结果进行验证, 为蛋白质相互作用数据的可靠性提供保障。各种预测蛋白质相互作用方法的计算方法有着不同的适用范围, 而相应的评价标准还未建立。因此, 在实际应用中应当考虑组合现有的不同方法进行预测, 并且将这些方法与实验方法结合, 挖掘蛋白质相互作用网络中更多的相互作用节点, 这样就可以更完整地描述蛋白质相互作用的生物学过程[15]。

2 蛋白质相互作用网络的生物学分析

2.1 蛋白质相互作用的可视化 在蛋白质相互作用的技术领域中, 生物信息学最终的挑战在于对蛋白质相互作用的重新构建。在一个界面友好的清晰格式中表达蛋白质相互作用是十分困难的, 所以将蛋白质相互作用可视化, 便于人类理解大量复杂数据背后所隐藏的信息, 因为网络图要比数据更容易发现其中所蕴含的信息。

在蛋白质相互作用网络中, 节点表示蛋白质, 节点间的连线表示这两个节点所代表的蛋白质之间存在相互作用。调整点和线的位置, 把蛋白质安排在二维图或三维空间中, 以此方式构建出蛋白质相互作用网络图。常用来构建蛋白质相互作用网络的软件很多, 例如可视化软件Osprey和Pajek[16]。

目前, 蛋白质相互作用网络的构建绝大多数都是静态的和定性的, 在网络中并不能反映出相互作用强度大小。动态的蛋白质相互作用可视化很困难, 因为这个过程会涉及多个因素参与。蛋白质组

计划最终目标之一就是构建出人类蛋白质相互作用网络图的动态形式,阐述蛋白质在何时与何地发生相互作用以及它们是如何发生相互作用的。

2.2 蛋白质相互作用网络拓扑结构的分析 蛋白质相互作用网络已证实网络科学普遍存在网络拓扑性质,提出了超小世界和自由尺度标度等网络拓扑属性和动力学模型,但其隐藏的动力学机制尚不清楚。

蛋白质相互作用形成了一个超小世界,即网络中连接任意两点所需要步数远小于随机网络。这一特征有利于信号的传递和整合。相互作用蛋白质在相互作用网络中不是平均分布的,有些蛋白质与很多其他蛋白质相互作用,形成节点(hub),而另外一些蛋白质只与很少的蛋白质作用,这种性质被称为自由尺度标度,也就是说蛋白质相互作用网络中节点的度分布呈幂函数分布[16]。L等[17]在分析了酵母、线虫和果蝇3个模式生物的蛋白质相互作用网络后得出结论,蛋白质相互作用网络具有高度容忍错误的特性,并且低连接度的节点占大部分,有利于一个蛋白质相互作用网络保持稳定,这是由于过多的直接连接会使整个蛋白质相互作用对定向灭活过度敏感。这种系统的稳健性,使得整个网络在某些蛋白质出现问题时能够最大可能地保持稳定性。这也许就解释了一些生命过程中,即使一些基因缺失或突变,仍然可以保持生物性状不发生重大改变。虽然迄今的研究大部分建立在部分蛋白质相互作用网络之上,但这些局部性质依然为后续研究奠定了基础。

2.3 蛋白质相互作用功能模块的研究 蛋白质在行使某些功能时往往是多个蛋白质共同起作用,因此,蛋白质相互作用网络具有模块化性质,对网络结构模块的研究可以揭示蛋白质相互作用网络形成的内部机制。不同的网络结构模块间具有一定的独立性,但网络中具有相近的相互作用的蛋白质趋向于具有相同或相似功能,因此,可以对蛋白质相互作用网络图中未知功能的蛋白质进行功能预测。

Pereira-Lea等[18]对酵母蛋白质相互作用网络进行网络功能模块研究后认为,不同物种之间可以通过建立模型来分析蛋白质相互作用网络。这符合系统生物学(system biology)的观点,即蛋白质相互作用的研究可以用于形成系统生物学研究中的模块,进而构建更为完整的蛋白质相互作用网络图。更复杂的数学模型也被用来进行蛋白功能预测,如概率方法、马尔可夫随机场和信息传递算法等。在利用蛋白质相互作用预测蛋白质功能的方法中,蛋白质相互作用的数据决定了其预测的可靠性,数据量的增加无疑将会提高其预测的准确性。另一方面,很多蛋白质在细胞中具有多种功能,如何成功地预测蛋白质的多种功能将会是蛋白质功能预测中的下一个研究方向。

2.4 蛋白质相互作用网络的比较 目前蛋白质相互作用的研究开始关注并开展在网络水平上的比较,即将两个或多个蛋白质相互作用网络进行比较。主要类型有不同物种的蛋白质相互作用网络比较,同一物种不同生理条件下的蛋白质相互作用网络比较,同一物种不同时间点上的蛋白质相互作用网络比较。蛋白质相互作用网络比较可以揭示出不同网络之间的保守结构,揭示新的网络功能模块、新的蛋白质功能、新的相互作用、蛋白质及其相互作用在不同的物种之间的保守进化关系[19]。

蛋白质相互作用网络比较是非线性比较,这具有很大的挑战性,目前这个领域中的一些初步研究也主要集中在原核生物中。有学者[20]提出用于蛋白质相互作用网络比较的PATHBLAST家族工具,通过整合蛋白质序列和相互作用拓扑的相似性来寻找两个蛋白质网络中相似的网络结构。但是,蛋白质相互作用网络过大会导致运算量骤增,不适合大规模地运用。在PATHBLAST的基础上又提出了一种多物种网络比较的算法,但这个算法仍有局限性,实际中只能应用到4或5个网络比较的规模。

在生物信息学高速发展的今天,又有一些新的算法引入蛋白质相互作用网络比较,如基于统计模型和贝叶斯参数推断的网络对齐算法等[21]。但是,目前比较蛋白质相互作用网络的研究仍然没有很深入,因为目前大部分网络比较方法所采用的网络相似性打分函数都没有严格的生物学基础。因此,需要以引入更合理的、以网络进化模型为基础的计算方法,使得蛋白质相互作用网络的比较可以迅速发展。这个发展必然会推动并促进整个生物学领域的快速发展。

3 小 结

目前,蛋白质相互作用的研究还远没有达到系统地理解生命现象的要求。蛋白质相互作用在深度、广度和生物学意义等方面随时在发生变化。到

目前为止,还没有建立一种生物完全的蛋白质相互作用网络,因此,蛋白质相互作用研究将是后基因组时代一个重要课题。系统地了解蛋白质相互作用将有助于人们认识蛋白质的功能和代谢通路等重要特征,对阐明疾病的发病机制以及诊断和治疗提供新思路和新途径。

参考文献:

[1] Archakov A,Ivanov YD. Analytical nanobiotechnology formedicine diagnostics[J]. Mol Biosyst 2007 3(5):336—342.

[2] Uetz P, Giot L, Cagney G, et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae[J]. Nature 2000 403(6770):623—627.

[3] Ito T, Chiba T, Ozawa R, et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome[J]. Proc Natl Acad Sci 2001 98(8):4569—4574.

[4] Ho Y, Gruhler A, Heilbut A, et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry[J]. Nature 2002 415(6868):180—183.

[5] Giot L, Bader JS, Brouwer C, et al. A protein interaction map of Drosophila melanogaster[J]. Science 2003 302(5651):1727—1736.

[6] Li S, Armstrong CM, Bertin N, et al. A map of the interactome network of the metazoan C. elegans[J]. Science 2004 303(5657):540—543.

[7] Stelzl U, Worm U, Lalowski M, et al. A human protein-protein interaction network: a resource for annotating the proteome[J]. Cell 2005 122(6):957—968.

[8] Rual JF, Venkatesan K, Hao T, et al. Towards a proteome-scale map of the human protein-protein interaction network[J]. Nature 2005 437(7062):1173—1178.

[9] Parrish JR, Gulyas KD, Finley RL Jr. Yeast two-hybrid contributions to interactome mapping[J]. Curr Opin Biotechnol 2006 17(4):387—393.

[10] von Mering C, Krause R, Snel B, et al. Comparative assessment of large-scale data sets of protein-protein interactions[J]. Nature 2002 417(6887):399—403.

[11] Stefan W, Almaas E. Peeling the yeast protein network[J]. Proteomics 2005 5(2): 444—449.

[12] Mrowka R, Patzak A, Herzel H. Is there a bias in proteome research[J]. Genome Res 2001 11(12):1971—1973.

[13] Deane CM, Salwinski L, Xenarios I, et al. Protein interactions: two methods for assessment of the reliability of high throughput observations[J]. Mol Cell Proteomics 2002 1(5):349—356.

[14] Mewes HW, Frishman D, Mayer KF, et al. MIPS: analysis and annotation of proteins from whole genomes in 2005[J]. Nucleic Acids Res 2006 34 (Database issue): D169—D172.

[15] Gray PN, Bushey K, Chappell TG, et al. A novel approach for generating full-length, high coverage allele libraries for the analysis of protein interactions[J]. Mol Cell Proteomics 2007 6(3):514—526.

[16] Maslov S, Sneppen K. Computational architecture of the yeast regulatory network[J]. Phys Biol 2005 2(4): S94—S100.

[17] Li D, Li JQ, Ouyang SG, Protein interaction networks of Saccharomyces cerevisiae, Caenorhabditis elegans and Drosophila melanogaster: Large-scale organization and robustness[J]. Proteomics 2006 6(2):456—461.

[18] Pereira-Leal JB, Audit B, Peregrin-Alvarez JM, et al. An exponential core in the heart of the yeast protein interaction network[J]. Mol Biol Evol 2005 22(3):421—425.

[19] Sharan R, Suthram S, Kelley R, et al. Conserved patterns of protein interaction in multiple species[J]. Proc Natl Acad Sci 2005 102(6): 1974—1979.

[20] Suthram S, Sittler T, Ideker T. The plasmodium protein network diverges from those of other eukaryotes[J]. Nature 2005 438(7064): 108—112.

[21] Yang Q, Sze SH. Path matching and graph matching in biological networks[J]. J Comput Biol 2007 14(1):56—67.

收稿日期:2007—11—05 本文编辑:王淑平