大规模蛋白质相互作用数据的分析与应用

孙景春[①] 徐晋麟[①*] 李亦学[②] 石铁流[②*]

(① 上海交通大学生命科学技术学院, 上海 200240; ② 中国科学院上海生命科学院生命信息中心生物信息中心, 上海 200031.
* 联系人, E-mail: tlshi@sibs.ac.cn; xujinlin2001@hotmail.com)

摘要 蛋白质相互作用在生命活动中起着重要的作用. 目前已开发出几种实验和计算方法能够得到大规模蛋白质相互作用数据. 但是, 与传统的实验结果相比, 蛋白质相互作用大规模数据中存在着比例较高的假阳性. 为了能够充分利用这些数据, 需要建立生物信息学方法对这些数据进行系统的评价, 进而提高数据的可信度, 并从中挖掘出有价值的生物信息. 本文对目前蛋白质相互作用大规模数据的计算分析和应用进行了总结, 包括蛋白质相互作用数据评估方法、与蛋白质其他信息的关系以及在生物学研究中的应用, 并提出了开发分析和挖掘蛋白质相互作用数据工具的主要方向, 以期有助于这些数据的研究和应用.

关键词 蛋白质相互作用 生物信息学 蛋白质组 蛋白质功能预测 生物途径

蛋白质相互作用在生命活动中起核心作用[1,2], 不仅是正常生理过程如DNA复制、转录、翻译、物质代谢、信号传导以及细胞周期控制的基础[3,4], 也在病理过程中起着重要的作用[5~7]. 蛋白质相互作用不仅为注释未知蛋白的生物学功能提供了线索, 也为了解生命活动的的机制, 提供了必要的信息.

越来越多的基因组大规模测序为在基因组范围内研究蛋白质相互作用提供了基础, 因为基因组全序列提供了基因组中所有蛋白的信息. 这些信息使得同时研究一个生物体内所有蛋白质相互作用的特征成为可能. 一个生物体内所有蛋白质相互作用被称为是蛋白质相互作用网络(protein-protein interaction network)或相互作用组(interactome)[8]. 目前, 已经建立了几个大规模技术来研究蛋白质相互作用. 因为同时研究大量的蛋白质相互作用, 所以其结果必然存在比例较高的假阳性. 尽管如此, 研究蛋白质相互作用网络仍为基因功能的研究提供了一条新的途径, 尤其是对那些应用序列相似性不能注释的基因. 此外, 在后基因组时代, 另一个挑战就是理解蛋白质之间是如何通过相互作用来完成生命活动的[9]. 因此, 研究蛋白质组内蛋白质相互作用对于研究蛋白质功能、整体上理解细胞机制是非常重要的.

尽管由实验和计算产生的蛋白质相互作用数据在不断增加, 然而这并不意味着生物学知识也会大幅度增加. 从这些技术得到的原始数据中提取生物学知识成为应用这些数据的一个瓶颈. 为了有效地应用这些数据, 需要计算和统计模型来处理数据、进行质量控制, 如可靠性的评估和证实. 因此, 本文对目前蛋白质相互作用大规模数据的计算分析进行了总结, 其中包括评估方法、与蛋白质其他信息的关系以及蛋白质相互作用数据在生物学研究中的应用, 并提出了开发和挖掘蛋白质相互作用数据工具的主要方向, 以期有助于蛋白质相互作用大规模数据的研究和应用.

1 蛋白质相互作用的评估

尽管目前已经有多种实验方法或计算方法能够对蛋白质相互作用进行研究和预测, 并且每一种方法在应用时都尽量避免假阳性. 但是, 目前所有的研究方法都具有技术上的偏向性或缺陷, 从而导致结果中错误数据比例较高. 但是在这些蛋白质相互作用数据中必然潜藏着具有生物学意义的信息. 因此为了能够从这些数据中有效地挖掘出具有生物学意义的信息, 有必要对数据的质量进行评价.

1.1 假阳性和假阴性

假阳性是指能够被实验技术检测到的、但在细胞中并不存在的蛋白质相互作用. 假阴性是指不能被实验技术检测到的、但在细胞中确实存在的蛋白质相互作用. 每一种实验技术或计算方法都存在一定程度的假阴性和假阳性. 因此, 在实验或预测结果后, 任何方法得到的结果都不可避免地要进行假阴性和假阳性的评估. 目前对蛋白质相互作用数据中假阴性和假阳性的评估主要是根据已有的相关蛋白质的功能、亚细胞定位、代谢途径、功能注释以及蛋白质

复合物相关信息来进行评估[10,11]. 尽管这些数据并不全面, 但是也能够在一定程度上反映预测结果的质量.

假阳性和假阴性的存在主要由以下结果之一引起的[12]: (ⅰ) 蛋白质相互作用的动力学本质. 蛋白表达和相互作用模式在不同生物学条件下是不同的, 而目前所有的实验方法或计算方法都不能做到动态检测或预测, 因此只能对真实存在的蛋白质相互作用, 得到一个粗略的描述. (ⅱ) 实验方法或计算方法的局限性. 每一种实验或计算方法所依据的生物学原理不同, 因此每一种方法预测的结果也只能部分描述真实的相互作用. (ⅲ) 在实验或计算过程中产生的错误. 这 3 个因素使得应用不同方法得到的蛋白质相互作用网络不同, 或者不同的实验室应用相同的方法也不能得到相同的蛋白质相互作用网络.

1.2 蛋白质相互作用数据的重叠和相互补充

到目前为止, 酵母作为模式生物, 其蛋白质相互作用得到了广泛的研究, 并已得到了大量的蛋白质相互作用数据, 这些数据为发展生物信息学方法来分析蛋白质相互作用提供了基础. von Mering等人[13]比较分析了不同来源的蛋白质相互作用大规模数据, 如酵母双杂交分析、蛋白质复合物的质谱分析、遗传相互作用、相关的mRNA表达以及通过基因组分析进行计算预测. 结果表明, 在近 80000 对蛋白质相互作用中, 只有约 2400 对是能够被 2 种或 2 种以上方法检测到, 也就是各种方法得到的结果之间的重叠很小. 这种现象可能是由于方法的不成熟和各自的偏向性而造成的. 同时结果还表明, 由不同方法产生的数据, 对应的相互作用蛋白质的功能分类具有不同的分布, 说明这些方法都具有各自的优点和缺点.

除了上面分析表明不同技术得到的蛋白质相互作用之间的重复性很低以外, 由不同实验室应用相同的技术得到的数据差别也很大. 以酵母双杂交实验为例, 在Uetz等人[14]和Ito等人[15]的结果中只有 141 对相互作用是相同的. 这 2 种方法鉴定的结果不到前人已发表的相互作用的15%[16]. 在最近发表的有关果蝇蛋白质相互作用的研究结果表明, 分别来自可靠性高的 2 个数据之间的重叠只有 26 对相互作用, 甚至比随机网络之间的重叠还要小[17,18]. 计算方法同实验方法一样, Strong等人[19]在应用系统发生谱[20]、基因邻近[21]、操纵子以及基因融合方法[22,23]对结核分枝杆菌(*Mycobacterium tuberculosis*)进行蛋白质功能关联预测时, 也出现了同样的结果, 即 4 种方法之间的重叠很小.

这些结果表明, 不仅不同技术之间的覆盖率不同, 而且每一个技术产生的蛋白质相互作用都具有独特的特征. 同时这种数据之间低覆盖率的现象证明目前大规模检测蛋白质相互作用技术远远没有成熟, 也表明大规模技术之间是相互补充的. 因此, 把不同来源的蛋白质相互作用数据进行整合会大大地扩充蛋白质相互作用网络的信息. 同时也说明为了尽可能多地鉴定真正存在的蛋白质相互作用, 需要多种不同的方法相互补充. 早在 1999 年Marcotte等人[24]就意识到这个问题, 并把多种计算方法整合起来预测蛋白质相互作用. Snel等人[25]在构建String数据库也是将几种计算方法进行整合来预测蛋白质相互作用. 但是, 在方法的整合上还需要进行深入的研究.

1.3 可信度的评估和提高

尽管目前还没有建立能系统评价蛋白质相互作用的统计方法, 然而已经有几个启发式方法被用来评价蛋白质相互作用, 这些方法能够在一定程度上对蛋白质相互作用进行评估, 并且为进一步提高这些数据的可信度提供了一定的线索.

最简单的方法就是, 应用不同方法对同一物种的蛋白质相互作用进行研究得到大量不同的数据, 然后对这些数据进行整合就可以显著提高蛋白质相互作用的可信度. 例如, 如果同一对蛋白质相互作用能够被 2 个不同的实验检测到, 这种联合观测就提高了这个特定相互作用的可信度. 在标准条件下应用多个独立的检测方法, 大范围双杂交检测能够鉴定系统性的假阳性. Ito等人[15]对酵母蛋白质相互作用数据进行分析, 观察到 4 次蛋白质相互作用数据的EPR指数(反应大规模数据的整体质量)为 0.60, 3 次的只有 0.55, 2 次只有 0.4, 1 次只有 0.2. 这表明随着观察次数的增多, 数据的可信度增强.

文献也是用来评估由大规模技术产生的蛋白质相互作用的一个有价值资料, 其基本思想就是, 如果 2 个蛋白质的名字出现在同一篇文章中, 这 2 个蛋白质就有可能具有相互作用. 尽管这种有关蛋白质相互作用的信息不可靠, 但是能够证实蛋白质相互作用或者至少为判断相互作用提供一定的线索. 基于文献挖掘的方法, Jenssen等人[26]最近通过分析 1000

万MEDLINE记录生成人类 13712 个基因-基因共引用的网络.

计算方法也被用来评价大规模蛋白质相互作用的观测结果. 为了验证蛋白质相互作用数据, Deane 等人[27]开发了 2 种方法, 也就是表达谱可靠性指数(expression profile reliability index)和平行进化同源确认方法(paralogous verification method, PVM). 通过比较相互作用的 2 个蛋白质的基因表达谱, 用表达谱可靠性指数来估计这种相互作用具有生物学意义的可能性. 这种方法的思想就是具有高度相关表达模式的蛋白质更可能具有相互作用. 平行进化同源确认方法是基于下面的原理, 如果 2 个蛋白质是同源的, 他们相互作用的蛋白质也趋向于是同源的. PVM 评价了 8000 对酵母中蛋白质相互作用, 其中 3003 对得到可信度鉴定. 其他计算方法主要是应用其他生物学信息来评价蛋白质相互作用的可信度水平, 包括应用蛋白质相互作用数据和其他类型的数据, 以及预测方法也有助于证实蛋白质相互作用.

2 蛋白质相互作用数据和其他生物学数据之间的关系

由于蛋白质功能、亚细胞定位、蛋白质结构、基因表达数据和蛋白质相互作用数据不断增加, 随之而来的就是各种数据之间的内在关系. 目前一些研究人员不仅研究了这种关系, 而且对这些数据之间进行了交叉验证并分析这些关系得到单一数据所不能得到的信息. 例如, 蛋白质功能和亚细胞定位信息可以用来评价蛋白质相互作用数据[10]. 考虑到通过计算得到的相互关系可靠性相对来说比较低, 我们应用蛋白质相互作用数据和其他生物学数据之间的关系来提高蛋白质相互作用的可信度. 因此, 对不同来源的生物学数据进行整合分析能够为研究细胞生命机制提供更深入、广泛的信息. 在这部分, 我们将讨论蛋白质相互作用数据和其他类型数据的相互关系, 包括亚细胞定位、功能分类、基因表达谱.

2.1 亚细胞定位

蛋白质的亚细胞分布对于从整体上理解细胞的分子机制是很有用的. 定位数据不仅可以用于评价由其他资源推导得到的信息, 而且一个蛋白质的亚细胞定位通常与其功能有关. 在蛋白质相互作用中, 涉及的 2 个蛋白质通常位于同一个亚细胞区间. 如果鉴定的相互作用是发生在已知具有相同亚细胞定位的 2 个蛋白质之间, 该相互作用可信度水平就会提高. 因此研究蛋白质相互作用和蛋白对的亚细胞定位之间的关系能够为来自大规模方法得到的蛋白质相互作用数据确证提供一个评价方法. 另一方面, 一个蛋白质可能具有多个亚细胞位置. 例如, NK-κB转位能够把该蛋白质从细胞质中转移到细胞核中[10]. 在这种情况下, 相互作用的 2 个蛋白质具有不同的亚细胞定位. Schwikowski等人[10]对酵母 2709 对相互作用数据中具有已知亚细胞定位的 1203 个蛋白质进行分析, 结果表明, 有约 78%的相互作用对具有相同的亚细胞分布. Chen等人[12]整理了存在于MIPS中的 2301 个已注释的二元蛋白质相互作用并提取了相关蛋白质亚细胞定位信息, 发现其中 2124 个相互作用(92%)所涉及到的蛋白质都具有相同的亚细胞定位. 这些结果表明具有相互作用蛋白质偏向于属于同一个特定的亚细胞位置, 即蛋白质相互作用和蛋白质亚细胞分布具有正相关性, 能够相互提供验证的数据.

2.2 功能分类

蛋白质相互作用通常跟特定的生物学途径有联系. 因此, 看到一对相互作用的蛋白质具有相同的细胞功能并不奇怪. Schwikowski等人[10]在研究酵母的对蛋白质相互作用数据和蛋白质功能分类关系时发现, 有约 63%的相互作用对具有相同的功能分类. 为了进一步分析具有相互作用的蛋白质之间细胞功能的关系, Chen等人[12]把具有细胞功能的 3936 个酵母ORFs根据Mering等人[13]提出的分类方法归到 11 个大的功能类中. 结果发现, 在 2301 个MIPS注释的蛋白质相互作用, 所涉及到的蛋白质都属于同一个功能类. 该结果表明, 蛋白质相互作用和蛋白质功能类别具有一定的正相关. 但是如此高的比例(100%)也可能是由于任何涉及到属于不同功能类的蛋白质相互作用对已经从这个数据中剔除出去了以及功能分类比较宽泛. 尽管如此, 这些研究结果在一定程度上说明了蛋白质相互作用和蛋白质功能类别具有一定的正相关. Marcotte等人[28]和Strong等人[19]应用蛋白质的功能注释信息来评价蛋白质相互作用的可靠性.

2.3 基因表达谱数据

目前基因表达谱数据分析是在基因组研究中最热门的研究领域中之一, 为研究生物学分子机制提供了一个新的途径. 例如, Eisen等人[29]基于基因表达

聚类分析研究的提出基于假定共表达的一组基因可能调节相关功能假说. 目前已有研究对蛋白质相互作用数据和基因表达数据之间的整体关系进行了系统研究.

Grigoriev[30]基于T7 噬菌体和酵母的分析表明, 共表达基因编码的蛋白质之间存在相互作用的概率比随机产生的蛋白质之间相互作用的概率要高得多. Ge等人[31]给出了一个更为全面的证据, 他们应用转录组-相互作用相关作图来比较相同表达类的基因所编码的蛋白质之间的相互作用和不同表达类的基因所编码的蛋白质之间的相互作用, 结果表明具有相同表达谱的基因更可能编码相互作用的蛋白质.

由此看来, 相互作用蛋白质在细胞中应该同时表达. 然而蛋白质相互作用和基因表达之间的关系是很复杂的, 这是因为基因表达水平未必能够真正代表蛋白质丰度. 而且, 由高通量方法得到的大规模蛋白质相互作用数据存在较大的假阳性. 尽管存在这些问题, 基因表达和蛋白质相互作用之间的相互关系的研究仍能够揭示出这些数据内部之间关系的总体趋势. 因此, 通过整合基因表达和蛋白质相互作用数据对提出更多有意义的假设是很重要的.

3 蛋白质相互作用数据的应用

蛋白质相互作用网络包含了蛋白质的相关信息, 如与之相互作用的蛋白质、蛋白质功能分类、相互作用的复合物组成以及所参与生物途径的有关信息. 因此, 挖掘蛋白质相互作用数据能够预测未知蛋白质可能的生物功能、发现潜在的蛋白质复合物亚基、构建蛋白质参与的生物途径. 然而, 考虑到蛋白质相互作用网络的复杂性和其所包含的假阳性, 从蛋白质相互作用网络得到的生物学推测必定存在着错误, 这也是不能忽略的. 尽管如此, 分析蛋白质相互作用数据进行生物学分子机制的研究仍是很重要的.

3.1 蛋白质功能预测

蛋白质通常通过与细胞内其他具有相同功能的蛋白质相互作用来完成其功能, 这种情况在 2 个相互作用蛋白质通常具有相同功能分类的统计学研究中有所体现[32]. 在蛋白质组学的研究中, 一个重要的挑战就是对那些不能应用同源预测方法进行功能注释的蛋白质进行注释, 而应用蛋白质相互作用信息对未知蛋白质预测其可能的功能, 是一种非常重要的非同源方法[33]. 因此, 针对这一目的, 已经开发出几个基于蛋白质相互作用数据进行蛋白质功能预测的方法.

应用最早的也是最简单的方法是主要连接数方法(majority rule' assignment)[10], 也被称为guilt by association[34], 其原理很简单. 例如, 如果蛋白质X(功能未知)被发现与 5 个蛋白质具有相互作用, 如果这 5 个蛋白质都具有相同的功能, 则蛋白质X也具有此功能; 如果其中 3 个具有相同的功能, 则蛋白质X则被认为可能具有该功能. 基于这种方法, 大规模蛋白质相互作用数据能够为应用序列比对不能进行功能注释的许多新的蛋白质提供一定的功能信息. Schwikowski等人[10]收集了已发表的 2709 个酵母蛋白质相互作用, 基于在酵母蛋白质组数据库(YPD)有关细胞功能和亚细胞定位注释的信息对相关蛋白质进行功能分类. 他们列出了 370 个功能未知的蛋白质至少具有一个功能已知的作用对. 其中, 有 29 个蛋白质具有 2 或 2 个以上的功能相同的蛋白质对. 但是该方法的缺点是不能预测相互作用的 2 个未知蛋白质的功能, 而且对于如果一个蛋白质相连的蛋白质分属与不同功能的个数相同的情况, 也很难对该蛋白质进行功能预测. 为此Hishigaki等人[35]提出了一种对邻近的几个蛋白进行统计的方法而不只用与之有直接相关的蛋白质, 但是这个距离的度很难把握. 随后Zhou等人[36]提出一种方法, 即应用表达谱之间的相关性得到的网络图中的最短距离来预测蛋白质功能.

上面的方法主要是根据蛋白质相互作用网络的局部进行的功能预测, Deng等人[37]、Vazquez等人[38]以及Karaoz等人[39]开发了几种从全局的角度进行功能预测的方法. Deng等人[37]基于马尔可夫随机模型开发了一个数学模型, 该模型不再是对相互作用蛋白质功能之间简单的查询, 而是应用贝叶斯方法对未知蛋白质赋予一个功能给一个概率, 也就是相当于给了一个可信度. Vazquez等人[38]应用模拟退火算法, 对整个网络中的未知蛋白质赋予功能, 然后找到能量最低的组合, 这样对每一个未知蛋白质都赋予功能, 并给出一个概率. Karaoz等人[39]在整合了基因表达数据的基础上应用Hopfield 网络技术的局部阈值原则(local-threshold rule)使得整个网络达到一个稳态, 这样对整个网络中的未知蛋白都进行了功能预测.

前面提到的简单的方法, 虽然有缺陷但是应用

起来比较简单，而后面几个从全局出发的方法，虽然能够对所有未知蛋白质进行功能预测，但是实现起来比较麻烦，而且这几种方法对整个网络中所有的未知蛋白质都进行了功能预测，因为是从总体上进行平衡，这样必然以降低应用简单方法得到高可信度的蛋白质功能预测来增加应用简单方法不能预测蛋白质的个数，也就是提高了覆盖率，降低了准确率．因此，如果把这两种思路结合起来，可能效果会更好．

3.2 生物途径的构建

生物途径的研究是一个具有挑战性的研究课题．生物途径网络是一个复杂协同的系统，包括代谢途径、信号传导途径以及遗传调控途径，而且途径之间还具有相互交叉．在所有这些途径中，都会不同程度地涉及到蛋白质相互作用．因此，研究蛋白质相互作用网络能够为生物途径提供一个框架．目前应用蛋白质相互作用数据进行生物途径的构建的研究还不是很多，主要是因为蛋白质相互作用数据的不成熟和途径的复杂性．这也为开发基于蛋白质相互作用数据构建生物途径提出了一个严重的挑战．目前应用蛋白质相互作用构建生物途径主要是整合了已有的基因表达数据[40,41]以及途径相关数据库KEGG[12]．

Chen等人[12]应用KEGG 细丝MAPK信号传导途径作为一个参照，依据蛋白质相互作用图构建细丝MAPK信号传导途径．其中首要的一步就是整合了不同来源的蛋白质相互作用数据来提高蛋白质相互作用数据的可靠性．Liu等人[41]通过整合基因表达和蛋白质相互作用数据，进而对生物途径成分的次序进一步验证和补充. Ideke等人[42]应用DNA芯片、质谱分析以及蛋白质-DNA相互作用来分析酵母半乳糖代谢途径．

4 展望

随着生命科学研究的重点从基因组转向蛋白质组，生物学网络逐渐成为研究热点，因为生物学网络能够从整体上反应细胞内分子机制．目前研究的生物学网络主要包括调控网络、代谢网络以及蛋白质相互作用网络．因为在细胞中所发生的几乎所有的事件都涉及到蛋白质相互作用，并且是其他生物学网络的基础，所以对蛋白质相互作用的鉴定和分析会越来越重要．

尽管已经有大量的与蛋白质相互作用的相关研究，但是在蛋白质相互作用的鉴定或预测、证实和应用中还存在很多问题需要进一步研究，主要包括以下几个方面：(ⅰ) 如何综合利用已有的试验技术和预测方法以便得到高质量的蛋白质相互作用数据；(ⅱ) 如何评价已经产生的或将要生成的大规模蛋白质相互作用数据；(ⅲ) 如何充分利用蛋白质相互作用数据和其他生物学知识的关系，系统地提高蛋白质相互作用数据可信度；(ⅳ) 如何从如此大量的数据中挖掘出更多有生物学意义的信息，尤其是如何利用蛋白质相互作用数据和其他生物学数据构建生物途径．因此，开发系统评价蛋白质相互作用数据的工具以及整合相关数据的工具是应用生物信息学研究蛋白质相互作用的重要方向．

随着实验和预测技术的改进、蛋白质相互作用数据的不断完善和提高、蛋白质相互作用数据系统分析、评价和应用的生物信息学方法的建立，蛋白质相互作用的研究会变得相对容易，并在生命机制的研究中发挥更大的作用．

致谢 本工作为国家自然科学基金(批准号：90408010)，国家高技术研究发展计划(批准号：2001AA231011，2002AA231051，2003AA231011 和 2004BA711A21)和国家重点基础研究发展规划(批准号：2002CB713807，2003CB715901 和 2004CB518606)资助项目．

参 考 文 献

1 Eisenberg D, Marcotte E M, Xenarios I, et al. Protein function in the post-genomic era. Nature, 2000, 405(6788): 823~826[DOI]
2 Auerbach D, Thaminy S, Hottiger M O, et al. The post-genomic era of interactive proteomics: Facts and perspectives. Proteomics, 2002, 2(6): 611~623
3 Wang J. Protein recognition by cell surface receptors: Physiological receptors versus virus interactions. Trends Biochem Sci, 2002, 27(3): 122~126[DOI]
4 Kone B C, Kuncewicz T, Zhang W, et al. Protein interactions with nitric oxide synthases: Controlling the right time, the right place, and the right amount of nitric oxide. Am J Physiol Renal Physiol, 2003, 285(2): F178~F190
5 Cohen F E, Prusiner S B. Pathologic conformations of prion proteins. Annu Rev Biochem, 1998, 67: 793~819[DOI]
6 Loregian A, Marsden H S, Palu G. Protein-protein interactions as targets for antiviral chemotherapy. Rev Med Virol, 2002, 12(4): 239~262[DOI]
7 Selkoe D J. The cell biology of beta-amyloid precursor protein and presenilin in Alzheimer's disease. Trends Cell Biol, 1998, 8(11): 447~453[DOI]
8 Legrain P, Wojcik J, Gauthier J M. Protein--protein interaction maps: A lead towards cellular functions. Trends Genet, 2001, 17(6): 346~352[DOI]

9 Garrels J I. Yeast genomic databases and the challenge of the post-genomic era. Funct Integr Genomics, 2002, 2(4-5): 212~237[DOI]

10 Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. Nat Biotechnol, 2000, 18(12): 1257~1261[DOI]

11 Gavin A C, Bosche M, Krause R, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature, 2002, 415(6868): 141~147[DOI]

12 Chen Y, Xu D. Computational analyses of high-throughput protein-protein interaction data. Curr Protein Pept Sci, 2003, 4(3): 159~181[DOI]

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

14 Uetz P, Giot L, Cagney G, et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature, 2000, 403(6770): 623~627

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

16 Hazbun T R, Fields S. Networking proteins in yeast. Proc Natl Acad Sci USA, 2001, 98(8): 4277~4278[DOI]

17 Formstecher E, Aresta S, Collura V, et al. Protein interaction mapping: A *Drosophila* case study. Genome Res, 2005, 15(3): 376~384

18 Giot L, Bader J S, Brouwer C, et al. A protein interaction map of *Drosophila melanogaster*. Science, 2003, 302(5651): 1727~1736[DOI]

19 Strong M, Mallick P, Pellegrini M, et al. Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization: A combined computational approach. Genome Biol, 2003, 4(9): R59[DOI]

20 Pellegrini M, Marcotte E M, Thompson M J, et al. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc Natl Acad Sci USA, 1999, 96(8): 4285~4288[DOI]

21 Overbeek R, Fonstein M, D'Souza M, et al. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci USA, 1999, 96(6): 2896~2901[DOI]

22 Enright A J, Iliopoulos I, Kyrpides N C, et al. Protein interaction maps for complete genomes based on gene fusion events. Nature, 1999, 402: 86~90[DOI]

23 Marcotte E M, Pellegrini M, Ng H L, et al. Detecting protein function and protein-protein interactions from genome sequences. Science, 1999, 285(5428): 751~753[DOI]

24 Marcotte E M, Pellegrini M, Thompson M J, et al. A combined algorithm for genome-wide prediction of protein function. Nature, 1999, 402: 83~86[DOI]

25 Snel B, Lehmann G, Bork P, et al. STRING: A web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucl Acids Res, 2000, 28(18): 3442~3444[DOI]

26 Jenssen T K, Laegreid A, Komorowski J, et al. A literature network of human genes for high-throughput analysis of gene expression. Nat Genet, 2001, 28(1): 21~28[DOI]

27 Deane C M, Salwinski L, Xenarios I, et al. Protein interactions: Two methods for assessment of the reliability of high throughput observations. Mol Cell Proteomics, 2002, 1(5): 349~356[DOI]

28 Marcotte E M, Pellegrini M, Thompson M J, et al. A combined algorithm for genome-wide prediction of protein function. Nature, 1999, 402(6757): 83~86[DOI]

29 Eisen M B, Spellman P T, Brown P O, et al. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA, 1998, 95(25): 14863~14868[DOI]

30 Grigoriev A. A relationship between gene expression and protein interactions on the proteome scale: Analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. Nucl Acids Res, 2001, 29(17): 3513~3519[DOI]

31 Ge H, Liu Z, Church G M, et al. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. Nat Genet, 2001, 29(4): 482~486[DOI]

32 Huynen M A, Snel B, Mering C, et al. Function prediction and protein networks. Curr Opin Cell Biol, 2003, 15(2): 191~198[DOI]

33 Marcotte E M. Computational genetics: Finding protein function by nonhomology methods. Curr Opin Struct Biol, 2000, 10(3): 359~365[DOI]

34 Oliver S. Guilt-by-association goes global. Nature, 2000, 403(6770): 601~603[DOI]

35 Hishigaki H, Nakai K, Ono T, et al. Assessment of prediction accuracy of protein function from protein-protein interaction data. Yeast, 2001, 18(6): 523~531[DOI]

36 Zhou X, Kao M C, Wong W H. Transitive functional annotation by shortest-path analysis of gene expression data. Proc Natl Acad Sci USA, 2002, 99(20): 12783~12788[DOI]

37 Deng M, Tu Z, Chen T, et al. Mapping Gene Ontology to proteins based on protein-protein interaction data. Bioinformatics, 2004, 20(6): 895~902[DOI]

38 Vazquez A, Flammini A, Maritan A, et al. Global protein function prediction from protein-protein interaction networks. Nat Biotechnol, 2003, 21(6): 697~700[DOI]

39 Karaoz U, Murali T M, Letovsky S, et al. Whole-genome annotation by using evidence integration in functional-linkage networks. Proc Natl Acad Sci USA, 2004, 101(9): 2888~2893[DOI]

40 Steffen M, Petti A, Aach J, et al. Automated modelling of signal transduction networks. BMC Bioinformatics, 2002, 3(1): 34~44[DOI]

41 Liu Y, Zhao H. A computational approach for ordering signal transduction pathway components from genomics and proteomics Data. BMC Bioinformatics, 2004, 5(1): 158[DOI]

42 Ideker T, Thorsson V, Ranish J A, et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science, 2001, 292(5518): 929~934[DOI]

(2005-03-05 收稿, 2005-07-18 收修改稿)