面向 PCP-MS 数据的 PPI 网络推断算法

陈 征 田 博 何增有
(大连理工大学软件学院 辽宁 大连 116000)

摘 要 随着蛋白质组学的发展,研究者们开始聚焦于人类的全部蛋白质相互作用(Protein-Protein Interaction,PPI)网络的建立,质谱分析技术已成为预测蛋白质相互作用的代表方法。质谱技术是构建蛋白质相互作用网络的主要实验手段之一,基于质谱技术产生了大量的蛋白质纯化数据,如 AP-MS 数据和 PCP-MS 数据等。这些数据为 PPI 网络的构建提供了重要的数据支持,但是通过人工的手段来构建 PPI 网络不仅低效,而且很不现实。因此,面向 PCP-MS 数据的网络推断算法是生物信息学研究的一个热点问题。文中针对一类主流的质谱(PCP-MS)数据的 PPI 网络构建算法问题开展研究,从解决目前存在的瓶颈问题出发,达到构建高质量 PPI 网络的目的。现有的面向 PCP-MS 数据的 PPI 网络推断算法的研究还处于初级阶段,相关方法较少。同时,算法结果的质量还存在着一些问题:1)很多错误的相互作用被包含在不同的推断算法结果中,同时一些正确的相互作用在结果中被遗漏;2)不同的推断算法在同一数据集上的表现差异较大;3)对于不同的数据集,同一算法表现性能的波动方差较大。因此,为了从 PCP-MS 数据中推断出结构可靠、质量较高的 PPI 网络,文中提出一种基于相关性分析与排序整合的 PPI 评分方法。该方法基于无监督学习,包括以下两个步骤:1)计算蛋白质之间的相关系数,得到多组相关性结果;2)采用排序整合的方法对多组结果进行整合,得到整合后的 PPI 分数。实验结果表明,所提方法在不使用参考标准的情况下,可以达到与有监督学习方法接近的结果。

关键词 MS 数据,PPI 网络,蛋白质直接相互作用,相关性分析,排序整合
中图法分类号 TP391.41 **文献标识码** A **DOI** 10.11896/jsjkx.181102215

PPI Network Inference Algorithm for PCP-MS Data

CHEN Zheng TIAN Bo HE Zeng-you
(School of Software Technology, Dalian University of Technology, Dalian, Liaoning 116000, China)

Abstract With the development of proteomics, scholars begin to pay more attention to the construction of Protein-Protein Interaction (PPI) network. Mass spectrometry(MS) has become a representative method for protein-protein interaction (PPI) inference, and it is one of the main experiment method to construct PPI network. Based on the technology of mass spectrometry, a large amount of experimental protein MS data is generated, such as affinity purification-mass spectrometry (AP-MS) data and protein correlation profiling-mass spectrometry (PCP-MS) data, which provide important data support for the construction of PPI networks, but constructing PPI networks by hand is impracticable and time consuming. Thus, PPI network inference algorithm for PCP-MS data has begun to become the research hotspot in bioinformatics. This thesis focused on the problem of PPI network inference for two main types of mass spectral data (AP-MS data and PCP-MS data), and designed effective methods respectively to solve the issue of current bottlenecks, achieving the construction of high-quality PPI network. The existing algorithms for PPI network interface from PCP-MS data are still in infancy, and there is a few of related algorithms. The existing method have several problem. Specifically: 1)many error interaction is contained in the results produced by the different algorithms, and the correct interaction is omitted in the results. 2)Different algorithms may produce very different results when they face the same data set. 3)For different data sets, the performance variance of the same algorithm is larger. For the problem of PPI network inference for PCP-MS data, this paper proposed a PPI scoring method based on correlation analysis and rank aggregation. The method is based on unsupervised learning and includes two steps. Firstly, correlation coefficient between protein pairs is computed, and multiple results of PPI scores can be obtained. Secondly, multiple results for each pair of proteins

到稿日期:2018-11-29 返修日期:2019-01-14 本文受国家自然科学基金项目(61572094)资助。
陈 征(1994-),男,硕士生,主要研究方向为机器学习、数据挖掘、生物信息,E-mail:chenzheng_dut@163.com;**田 博**(1994-),女,硕士生,主要研究方向为机器学习、数据挖掘、生物信息;**何增有**(1976-) 男,博士,教授,博士师导师,CCF 高级会员,主要研究方向为生物信息、机器学习、数据挖掘,E-mail:zyhe@dlut.edu.cn(通信作者)。

are combininect via rank aggregation to a single PPI score. The experimental results show that this method is comparable with those supervised learning methods using standard reference set.

Keywords MS data, PPI network, Protein direct interaction, Correlation analysis, Rank aggregation

1 引言

1.1 PPI 网络构建问题

蛋白质是细胞活动及功能的最终执行者。每个蛋白质并非独立完成其功能，通常与其他蛋白质相互作用形成临时或稳定的复合体以执行特定的功能。因此，针对大规模、高通量的蛋白质相互作用的研究应运而生，其目的是在细胞特定的生理条件下描绘出整个蛋白质组中蛋白质间相互作用的网络图[1]。基于这些作用关系，可以真正意义上阐明一个蛋白质的功能，并进一步研究细胞内各种生理反应的发生及调节机制，最终揭示生命的本质。

随着 2014 年人类蛋白质组草图的初步完成[2-3]，未来的研究重点必将转移到人类的全部 PPI 网络的建立上[4]。质谱(Mass Spectrometry，MS)分析技术是用于构建 PPI 网络最重要的生化实验之一。基于质谱技术的 PPI 网络构建的具有代表性的方法有(见图 1)："亲和-纯化"质谱(Affinity Purification-Mass Spectrometry，AP-MS)分析、蛋白质相关性概要质谱(Protein Correlation Profiling Mass Spectrometry，PCP-MS)分析以及交联质谱(Cross-linking Mass Spectrometry，XL-MS)[5-6]分析等。其中，AP-MS 技术通过不断地选取"诱饵"蛋白质(Bait)，捕获与该蛋白质具有相互作用关系的多个"猎物"蛋白质(Prey)，再通过质谱技术识别这些蛋白质，从而获取蛋白质之间的相互作用关系，最终建立 PPI 网络(图 1(a))。PCP-MS 技术通过层析(Chromatography)或者密度梯度离心(Density Gradient Centrifugation)等方式对蛋白质复合物进行分离，然后通过质谱对共洗脱的蛋白质进行识别，获取蛋白质之间的关系，实现 PPI 网络的构建(图 1(b))。不同于 AP-MS 技术，PCP-MS 技术无需亲和纯化过程，取而代之的是通过不同的方式对蛋白质复合物进行分离，通常采用高效液相色谱法(High Performance Liquid Chromatography，HPLC)的分离方式，如蔗糖密度梯度离心法(Source Density Gradient Centrifugation，SGF)、空间排阻色谱法(Size-Exclusion Chromatograph，SEC)、离子排斥色谱法(Ion Exchange Chromatography，IEX)和等电位聚焦法(Isoelectric Focusing，IEF)等[6]。XL-MS 技术使用交联剂对蛋白质复合物进行处理，将空间距离足够近且与交联剂反应的氨基酸并通过共价键连接。该方法使用交联剂，通过共价键捕获具有相互作用关系的蛋白质，然后应用质谱分析技术对交联蛋白质进行识别[7](图 1(c))。近年来，基于 MS 的方法被广泛用于构建不同类型的 PPI 网络，并取得了良好的效果[8-13]。由于生化实验产生了大量的蛋白质 MS 数据，因此通过人工方式构建 PPI 网络是不现实的，面向质谱数据的 PPI 网络推断算法也就成为生物信息学与蛋白质组学领域中研究的重要问题之一[14]。只有彻底解决了 PPI 网络推断问题，才能真正地解析蛋白质之间的关系及其功能。同时，构建高质量的 PPI 网络也是很多相关问题的研究基础：蛋白质复合体/复合物(Protein Complex)与功能模块识别[15-16]、蛋白质功能预测[17]、疾病基因与药物靶点预测[18]、疾病标记物检测[19-21]等。这些问题都是生物信息学课题中的重点和热点。虽然基于质谱技术的方法在构建 PPI 网络方面取得了良好的效果，但现有的面向 MS 数据的 PPI 网络推断算法的求解质量还存在很大问题[14,22-23]：1)不同推断算法的结果中包含很多错误的相互作用，遗失很多正确的相互作用；2)对于同一数据集，不同推断算法的结果差别较大；3)针对不同的数据集，同一推断算法的性能波动很大。因此，为了从 MS 数据中推断出高质量的 PPI 网络，必须研制更有效的算法，这不仅能决定人类全蛋白质组 PPI 网络的准确刻画，还能推动相关领域的很多核心技术的发展。

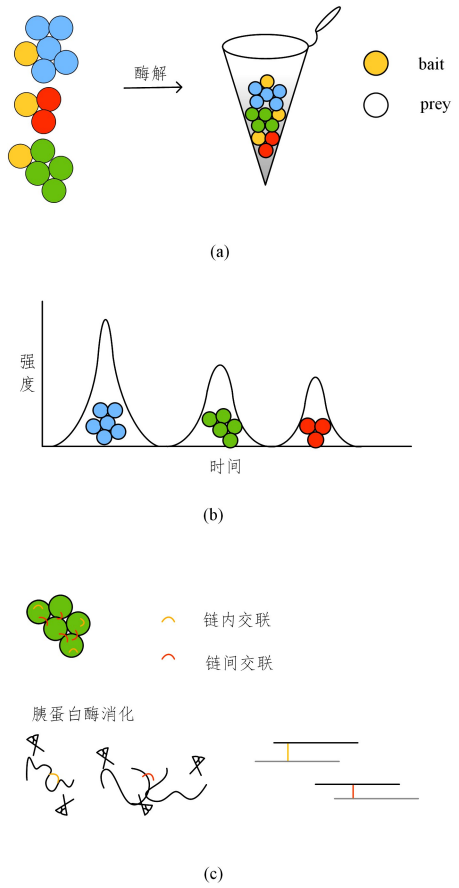


图 1 PPI 推断的 MS 实验方法

Fig. 1 MS experimental method of PPI inference

近年来，PCP-MS 已被成功应用于目前规模最大的 PPI 网络的构建中[9,24]，因此，文中是基于 PCP-MS 产生的质谱数据来研究 PPI 网络推断问题。

1.2 相关工作

PCP-MS 技术是通过蔗糖密度梯度离心法(SGF)、等电聚焦法(IEF)以及离子排斥色谱法(IEX)等方式，分别依据密度、等电点以及疏水性等特性对蛋白质复合物进行分离，从而获取蛋白质之间的关系。近年来，研究者开始趋向于针对

PCP-MS 数据的 PPI 网络构建研究。

为了更好地说明研究现状和存在的问题,本节对 PCP-MS 实验数据的产生流程和关键步骤进行说明。图 2 给出 PCP-MS 实验的流程。首先,细胞裂解得到蛋白质复合物后,通过 HPLC 对蛋白质复合物进行分离,通常使用 SGF,IEX 和 IEF 等技术。首先,基于密度、空间等不同生化特性分离蛋白质复合物后,收集其片段;然后,通过 LC-MS/MS 技术对蛋白质复合物片段进行质谱分析,再根据质谱数据识别蛋白质(在同一个片段中被识别的蛋白质可能来源于同一个蛋白质复合物,也就是说在同一个片段中的蛋白质之间可能存在相互作用)。最后,将实验得到的"蛋白质-片段"(Protein-Fraction)数据作为 PPI 网络构建算法的输入数据,从而 PPI 网络。

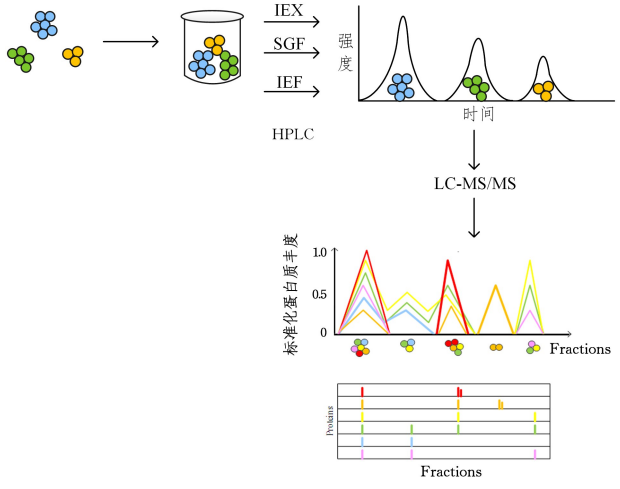


图 2　PCP-MS 实验流程图法

Fig. 2　PCP-MS experiment flowchar

近年来,针对 PCP-MS 数据的 PPI 网络推断的研究刚刚起步,成果尚少。2012 年,Havugimana 等[25]实现了人类蛋白质的 PPI 网络的构建,显著扩大了人类蛋白质组学蛋白质相互作用关系的认知范围,并在预测 PPI 网络基础上进一步实现了对蛋白质复合物的推断,为蛋白质组学数据库提供了高质量的 PPI 数据及蛋白质复合物数据。该方法进行 PPI 网络推断的流程如下:首先,通过生化实验(IEX-HPLC,IEF-HPLC 和 SGF-HPLC 技术)得到"蛋白质-片段"数据;然后,计算两两蛋白质的相互作用分数,包括皮尔森(Pearson)相关系数、基于噪音模型下的皮尔森 Noised-Pearson 相关系数,以及加权互相关系数(Weighted cross correlation,Wcc)等;接着,以每对蛋白质的相关系数作为特征,同时利用 CORUM 数据库中的部分蛋白质复合物信息构建正例集和反例集,并使用随机森林的分类方法训练模型;最后,应用得到的模型对数据集中蛋白质的相关性进行预测,从而过滤数据集中假的相互作用关系,实现高质量 PPI 网络的构建。

2015 年,Wan 等[24]研究了除人类外其他 9 种生物的 PPI 网络构建及蛋白质复合物推断。在 PPI 网络的构建过程中,首先计算 PPI 分数,然后结合 CORUM 数据集中的复合物信息,应用机器学习中的 SVM 分类器来训练模型,通过该模型得到的 PPI 网络将作为下一步蛋白质复合物推断的基础。上述两种方法具有一定的共同点:1)基于 PCP-MS 实验数据计算蛋白质之间的相关系数,并以该系数作为 PPI 的特征;2)引入参考标准 CORUM 数据库中的 PPI 信息构建训练集;3)通过有监督学习方法对 PPI 存在与否进行分类,最终得到蛋白质相互作用关系。

目前,上述方法虽然已被用于不同物种和规模的 PPI 网络构建,但是还存在以下问题:1)面向 PCP-MS 数据进行 PPI 网络推断的研究较少,除上述 Havugimana 等[25]和 Wan 等[24]的方法以外,尚未见到其他相关研究;(2)目前的方法引入了标准数据库的信息来预测 PPI,而在评估最终结果时也采用该标准数据库作为参考集,这种方式的合理性值得商榷。因此,基于 PCP-MS 的 PPI 网络推断算法的研究尚处于初级阶段,还有很大的研究空间。

1.3　研究动机与本文主要贡献

近年来,面向 MS 数据的 PPI 网络推断算法已成为 PPI 网络推断的主流方法,面向 PCP-MS 实验数据的 PPI 网络推断研究才刚刚起步,现有算法的求解质量还远远不够。已有的针对 PCP-MS 数据的网络推断方法通常将蛋白质相互作用关系推断转化为分类问题,即利用分类器来推断蛋白质之间是否存在相互作用关系。在分类器模型的训练过程中,已有方法借助部分参考集数据训练模型,可能存在过拟合以及性能评估不合理等问题。因此,本文旨在解决针对 PCP-MS 数据的 PPI 网络构建问题。

本文提出了一种面向 PCP-MS 数据的 PPI 网络构建算法。该方法以初始 PCP-MS 实验数据作为输入,包括两个步骤:首先,计算蛋白质之间的相互作用分数,衡量蛋白质的相关性(由于实验过程通常会采用不同的分离技术来分离蛋白质复合物,因此会有多组实验结果);然后,通过排序整合的方式将多组实验结果进行整合,得到更为综合、全面的蛋白质相互作用关系,并根据这个结果构建 PPI 网络。

综上所述,本文主要的创新点有:针对 PCP-MS 数据的 PPI 网络构建问题,提出了一种基于相关性分析和排序整合的 PPI 网络构建方法,与已有的有监督学习方法相比,所提方法没有引入参考集的信息来训练模型,因此得到的结果更有说服力;此外,该方法本质上为基于无监督学习的 PPI 评分策略,避免了参数选取调优和过拟合等问题。通过对多组结果的整合获取更全面、可靠的 PPI,以构建高质量的 PPI 网络。

2　面向 PCP-MS 数据的 PPI 网络构建算法

2.1　方法描述

目前,针对 PCP-MS 数据的 PPI 网络构建采用有监督学习的方法,其在推断 PPI 过程中借助部分参考集的信息来训练模型。这种方式的可解释性存在一定的问题,且可能导致过拟合。本文提出了一种基于相关性分析与排序整合的 PPI 网络构建方法,该方法的流程如图 3 所示。

(1)相关性分析。PCP-MS 数据包含蛋白质分离实验过程中每个蛋白质在不同片段中的数量信息。如果两个蛋白质之间具有相互作用关系,则它们很有可能同时出现在一个片段中。因此,首先计算蛋白质之间的相关系数,以衡量两两蛋白质的相关程度。本文选取 Pearson 相关系数、Noised-Pearson 相关系数以及 Wcc 这 3 种相关系数。对于每组初始

PCP-MS 数据，通过计算蛋白质之间的相关系数，可以得到多组结果，每个实验组对应一个 PPI 排序列表，系数越大表明蛋白质之间越可能存在相互作用关系。

(2)排序整合。对于上一步得到的多组 PPI 排序列表，采用排序整合的方式将其整合为一个综合的 PPI 分数，以全面、综合地表示蛋白质相互作用关系。本文使用 Reinforce 方法对多个排序列表进行整合。

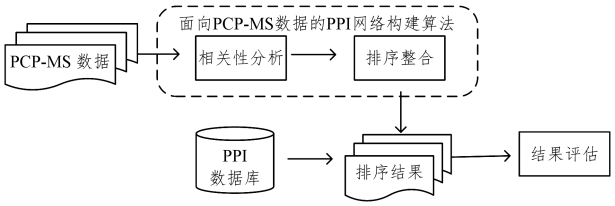


图 3 面向 PCP 数据的 PPI 网络构建算流程图

Fig. 3 Flowchart of construction algorithm of PPI network based on PCP data1

本文提出的基于相关性分析与排序整合的 PPI 评分方法是一种无监督学习方法，该算法通过上述两个步骤，能够有效地预测蛋白质相互作用关系。与有监督学习的方法相比，在不使用“标准答案”的情况下，所提方法同样可以构建出稳定且可靠的 PPI 网络。

2.2 相关性分析

在 PCP-MS 实验中，利用生化分馏技术使得同一个复合物的蛋白质被共洗脱，通过质谱分析，得到“蛋白质-片段”数据。为了更好地说明数据的处理及计算过程，给出相关性分析过程图，如图 4 所示。

采用分馏技术对蛋白质进行分离，得到 PCP-MS 实验数据，如图 4 所示。该实验数据共有 M 个片段(Fraction)，每个片段记录了 N 种蛋白质在其中出现的数量信息。具体而言，每个蛋白质与 Fraction 构成一个“Protein-Fraction”向量的形式。由于在同一个 Fraction 被识别的蛋白质可能来源于同一个蛋白质复合物，即蛋白质之间可能存在相互作用。因此，首先通过相关系数的计算来度量蛋白质之间的相关性。虽然存在很多不同的相关系数计算方法，但为了便于比较和说明，本文采用了文献[25]中使用的 3 种相关系数。

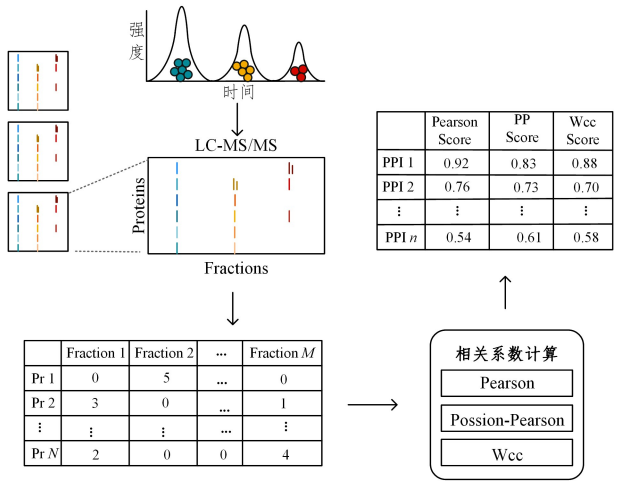


图 4 相关性分析过程图

Fig. 4 Flowchart of correlation analysis

(1)Pearson 相关系数

给定一个包含 N 种蛋白质和 M 个 Fraction 的 PCP-MS 数据，可以用一个 $N\times M$ 的矩阵 $\boldsymbol{A}$ 表示，矩阵 中的元素表示实验中蛋白质 i 在 Fraction k 中的数量。

首先，计算每个蛋白质在片段中出现的频率[25]，计算公式如下：

$$B(i,k)=\frac{A(i,k)}{\sum_{i=1}^{N}A(i,k)} \tag{1}$$

其中，$\boldsymbol{B}$ 为矩阵 $\boldsymbol{A}$ 的归一化矩阵，矩阵中的每一行为每个蛋白质的归一化向量。

然后，计算两个蛋白质之间的 Pearson 相关系数，计算公式如下：

$$PScore(i,j)=\frac{\sum_{k=1}^{M}B(i,k)*B(j,k)-\frac{(\sum_{k=1}^{M}B(i,k))*(\sum_{k=1}^{M}B(j,k))}{M}}{\sqrt{((\sum_{k=1}^{M}B(i,k)^2)-\frac{(\sum_{k=1}^{M}B(i,k))^2}{M})((\sum_{k=1}^{M}B(j,k)^2)-\frac{(\sum_{k=1}^{M}B(j,k))^2}{M})}} \tag{2}$$

其中，$PScore(i,j)$ 表示蛋白质 i 和蛋白质 j 的 Pearson 相关系数，$B(i,j)$ 和 $B(j,k)$ 分别为蛋白质 i 和蛋白质 j 在 Fraction k 中的频率信息，M 为 Fraction 总数。

(2)Noised-Pearson 相关系数

如果两个蛋白质多次同时出现在同一个 Fraction 中，并且二者同时出现时数量均较大，那么 Pearson 相关系数可以较为有效地度量它们之间的相互作用关系。然而，如果两个蛋白质同时出现时数量偏少，Pearson 相关系数的度量效果则不太理想。为了解决上述问题，可以在原始 MS 数据中引入噪声来计算 Pearson 相关系数[25]，具体过程如下。

步骤 1 对初始 MS 数据加噪声。假设每个 Fraction 观测到的蛋白质数量建模由一个泊松(Possion)过程决定，其中设置参数 λ 为对应初始纯化数据中蛋白质 i 在 Fraction k 中的数量，即 $A(i,k)$ 的值。在矩阵 $\boldsymbol{A}$ 中加入噪声项，如式(3)所示：

$$\boldsymbol{C}=\boldsymbol{A}+1/M \tag{3}$$

选择将常量 $1/M$ 加入到模型中的原因如下：1)保证矩阵中每个蛋白质在 Fraction 上都有相应的数量信息；2)背景概率在 M 个片段上是均匀分布的。

步骤 2 噪声 MS 数据的泊松过程。模拟生化实验中的数据生成过程，生成以 $C(i,j)$ 为参数的泊松分布随机数 $D(i,j)$，得到矩阵 $\boldsymbol{D}$。

$$D(i,j)=\text{Possion}(C(i,j)) \tag{4}$$

步骤 3 相似性度量。此过程与上述 Pearson 相关系数的计算方法相同，对 $\boldsymbol{D}$ 进行归一化处理后，计算蛋白质之间的 Pearson 相关系数，计算公式如下：

$$E(i,k)=\frac{D(i,k)}{\sum_{i=1}^{N}D(i,k)} \tag{5}$$

$$PPScore(i,j)=\frac{\sum_{k=1}^{M}E(i,k)*E(j,k)-\frac{(\sum_{k=1}^{M}E(i,k))*(\sum_{k=1}^{M}E(j,k))}{M}}{\sqrt{((\sum_{k=1}^{M}E(i,k)^2)-\frac{(\sum_{k=1}^{M}E(i,k))^2}{M})((\sum_{k=1}^{M}E(j,k)^2)-\frac{(\sum_{k=1}^{M}E(j,k))^2}{M})}} \tag{6}$$

其中,$PPScore(i,j)$表示基于噪音模型下蛋白质 i 和蛋白质 j 的 Pearson 相关系数。上述步骤 2 和步骤 3 可以重复多次(比如,文献[25]中设置为 1 000 次,本文亦重复 1 000 次),并取其均值作为最终的 PPI 分数。

(3)Wcc 系数

除上述两种计算相关系数的方法外,WCC 方法[26]也是生物信息学领域中一种衡量蛋白质之间相关性的方法。给定 $N\times M$ 的 PCP-MS 数据矩阵 $\boldsymbol{A}$,$A(i,k)$为蛋白质 i 在 Fraction k 中的数量。

对于一对蛋白质 i 和 j,其 Wcc 相关系数为 i 和 j 的互相关系数除以 i 的自相关系数与 j 的自相关系数之乘积。Wcc 相关系数的计算公式如式(7)所示。

$$WccScore(i,j)=\frac{ccoss[i,j]}{acors[i]*acors[j]} \tag{7}$$

$$ccors[i,j]=\sum_{k=1}^{M}A(i,k)*A(j,k)+\sum_{\Delta=1}^{tr-1}\sum_{k=1}^{M-1}(A(i,k)*A(j,k+\Delta)*wg_{\Delta}$$

$$acors[i]=2*\sqrt{(\sum_{k=1}^{M}A(i,k)^2)+\sum_{\Delta=1}^{tr-1}\sum_{k=1}^{M-1}A(i,k)*A(i,k+\Delta)*wg_{\Delta}}$$

$$acors[j]=2*\sqrt{(\sum_{k=1}^{M}A(j,k)^2)+\sum_{\Delta=1}^{tr-1}\sum_{k=1}^{M-1}A(j,k)*A(j,k+\Delta)*wg_{\Delta}}$$

其中,$WccScore(i,j)$表示蛋白质 i 和 j 的 Wcc 的系数;$ccors[i,j]$为其互相关系数;$acors[i]$和 $acors[j]$ 分别为蛋白质 i 和 j 的自相关系数;参数 wg 表示权重向量,与三角宽度 tr 有关,可以根据 tr 计算得到。

本文应用 WCC 方法计算得到蛋白质之间的 Wcc 系数。与 Pearson 相关系数相比,该方法考虑到了两个蛋白质质谱的相对偏移,即可以比较一个蛋白质在某点/片段处的谱图与另一个蛋白质相应点/片段附近的谱图。实验中,Wcc 系数可由 R 语言 wccsom 包中的 Wcc 函数[1)]计算得到。关于参数的选取,tr 取值为 1,其余均为默认参数,且 Wcc 系数值在 0～1 之间。

由于在计算相关系数的过程中会产生 $N(N-1)/2$ 对 PPIs,其中蛋白质的数量 N 一般为几千,因此所有可能的候选 PPI 的数量是非常庞大的;同时,这些候选 PPI 中包含大量假阳性的结果,因此对 3 种相关系数的计算结果,只保留其系数大于 0.5 的 PPI 做进一步的分析。

2.3　排序整合

每对 PPI 对应多个相关系数,如何将其整合为一个分数作为其排名的可靠依据是问题的关键。该分数整合问题的本质是一个排序整合(Rank Aggregation)问题,即给定一组排序列表,通过某种方式将其整合为一个"更好"的列表。

排序整合方法大多基于以下策略:是否依赖排名信息和是否依赖分数信息。对应的方法称为基于排名方法和基于分数方法。在基于排名的方法中,元素是按照其排名或者等级排列的。而在基于分数的方法中,列表中的元素是按照分数大小或者对分数进行转换后的分值进行排序的。显然,每个 PPI 的多个相关系数相当于多个分数,因此本文采用基于分数的整合方法。

目前已存在多种不同的分数排序整合方法,从简单的加权平均到各类复杂的统计方法等。针对质谱数据存在较多噪声的特点,提出了基于假设检验和错误率控制的 Reinforce 方法,该方法不仅可以用于 PCP-MS 数据的 PPI 网络推断,也可以用于解决 AP-MS 数据的 PPI 网络构建问题[27]。

Reinforce 方法包含以下几个步骤。

(1)数据预处理。假设有 F 组 PPI 相关系数列表,将每组列表按相关系数降序排列后,以排名作为每对 PPI 的新分数。然后对每个列表进行归一化处理,即除以每个列表的 PPI 总数,使得分数值在[0,1]之间。对于可能存在缺失值的情况,将缺失值设置为 1。将上述多个预处理后的列表合并,可以得到一个 $n\times F$ 的矩阵 $\boldsymbol{S}$(见图 5(b)),其中 n 是在所有 F 个列表出现过的蛋白质的总数。每个 PPI 对应一个归一化排序向量 $\boldsymbol{s}=(s_1,s_2,\cdots,s_F)$。

(2)排序整合。此步骤采用 RAA[28]的改进方法 adjustedRRA来对多组排序列表进行整合,如图 5(c)所示。首先对 $\boldsymbol{s}=(s_1,s_2,\cdots,s_F)$中的元素按照从小到大进行排序,得到向量 $\boldsymbol{s}^*=(s_1^*,s_2^*,\cdots,s_F^*)$,其中 $s_i^*<s_j^*$。adjustedRRA 方法假设每个列表中的排序是随机的,即每个 PPI 归一化排序向量的元素服从[0,1]上的均匀分布。基于此假设,计算归一化排序向量 $\boldsymbol{s}=(s_1,s_2,\cdots,s_F)$的显著程度,即 $p_i=P(s_i'<s_i^*)(i=1,\cdots,F)$,$s'$为空假设前提下能生成随机排序向量 $\boldsymbol{s}'=(s_1',\cdots,s_F')$,每个元素 s_i'服从[0,1]均匀分布。按照下列公式依次计算每个 p_i 的值:

$$p_i=\sum_{f=i}^{F}C_F^f(s_i^*)^f(1-s_i^*)^{F-f} \tag{8}$$

以此得到向量 $\boldsymbol{p}=(p_1,p_2,\cdots,p_F)$。然后采用置换检验的方法计算最终的 P-value,具体步骤如下:

1)根据式(8)计算得到向量 $\boldsymbol{p}=(p_1,p_2,\cdots,p_F)$;

2)计算统计值 $T=-2\sum\log p_i$;

3)置换检验,随机生成 m 个服从[0,1]均匀分布的排序向量 $\boldsymbol{r}''$,经步骤 1)和 2)计算得到统计值 T''。假设统计 m 个 T''值中满足"$T''>$"的数量为 c,则最终的 P-value$=c/m$。

(3) 错误发现率控制。假设有 n 对 PPI,排序整合后得到了每对 PPI 的 P-value,即为 $\boldsymbol{Q}=(Q_1,Q_2,\cdots,Q_n)$。

1) https://cran.r-project.org/web/packages/wccsom/index.html

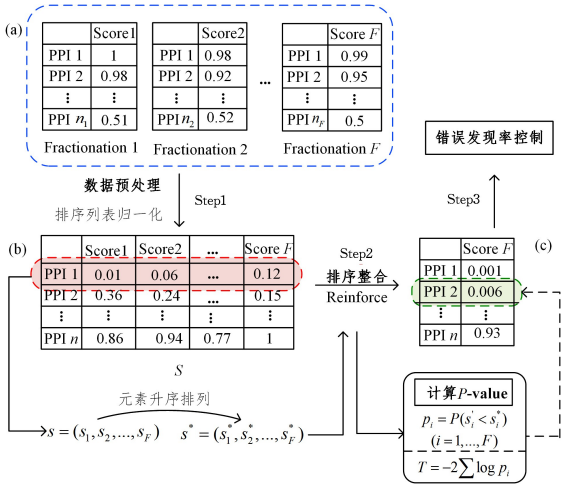


图 5 Reinforce 方法过程图

Fig. 5 Procedure of Reinforce method

然后采用 Storey[29] 等提出的方法估计错误发现率，步骤如下：

1)对向量 $\boldsymbol{Q}=(Q_1,Q_2,\cdots,Q_n)$ 中的每个元素按升序排列，得到一个向量 $\boldsymbol{Q}^*=(Q_1^*,Q_2^*,\cdots,Q_n^*)$；

2)令 $\bar{q}(Q_n^*)=FDR(Q_n^*)$，其中 $FDR(\alpha)$ 为阈值 α 的错误发现率估计；

3)基于 $\bar{q}(Q_i^*)=\min\{FDR(Q_i^*),\bar{q}(Q_{i+1}^*)\}$ 得出计算结果，其中 $i=n-1,n-2,\cdots,1$，以此作为最终的 FDR 估计值。

Reinforce 方法通过以上 3 个步骤对多个 PPI 排序列表进行整合，每个 PPI 得到单一 PPI 分数，该分数能够综合、全面地表示 PPI 的关联程度。

3 实验结果

3.1 数据集和参考集

本文的实验数据来源于人类可溶性蛋白质复合物数据库[1)]。选取初始 LC-MS/MS 实验数据作为数据集，其中包括 13 组分馏实验中 5 584 种蛋白质在 1 163 个不同片段的 MS 数量信息。实验中 13 组原始 PCP-MS 数据集的相关信息如表 1 所列。

表 1 13 组 PCP-MS 数据集的相关信息

Table 1 Information about 13 groups PCP-MS data set

	细胞	分馏方式	MS 技术	片段数
1	HEK293	HCW	LTQ	120
2	HeLaS3	SAF	LTQ	48
3	HeLaS3	HCW	LTQ	120
4	HeLaS3	HCW	LTQ-Orbitrap	120
5	HeLaS3	TCS	LTQ	375
6	HeLaS3	SGF	LTQ-Orbitrap	14
7	HeLaS3	IEF(PH 5-8)	LTQ-Orbitrap	10
8	HeLaS3	IEF(PH 3-10)	LTQ-Orbitrap	10
9	HELaC3	IEF(PH 5-8)	LTQ-Orbitrap	10
10	HELaC3	IEF(PH 3-10)	LTQ-Orbitrap	10
11	HELaC3	SGF	LTQ-Orbitrap	14
12	HELaC3	WAX	LTQ	43
13	HELaC3	TCS	LTQ	269

为了验证本文方法的有效性，选取 CORUM 数据集[2)] 作为参考集[30]，该数据集共包括 324 种蛋白质复合物。如果同一个复合物中的蛋白质之间存在相互作用关系，而不在同一个复合物中的蛋白质之间不具有相互作用关系，则可以按此方式构造出一个 PPI 参考集来对实验结果进行评估。

3.2 数据集和参考集

为了测试本文提出的基于相似性分析和排序整合的 PPI 评分方法的效果，分别针对无监督学习方法(本文方法)与有监督学习方法(文献[25]提出的方法)做了两组实验，并对两种方法的结果进行比较。

(1)基于有监督学习方法

首先，对于初始 MS 实验数据集，分别计算 13 个分馏实验组中蛋白质之间的相关系数，包括 Pearson 相关系数、Noise-Pearson 系数以及 Wcc 相关系数。然后，以 0.5 为阈值，保留相关系数不小于 0.5 的 PPI，得到 13 组实验和 3 种相关系数的 PPI 列表，将其组合，得到 122×10^4 对 PPI，每个 PPI 对应一个长度为 39 的相关系数向量。为了找出具有真实相互作用关系的蛋白质，应用随机森林(Random Forest，RF)分类器对 PPI 进行分类。

实验训练集的构造方式如下：首先，从 CORUM 数据集中随机选取一半的蛋白质复合物，把同在一个复合物中的每对蛋白质作为正例放入正例集，而不在同一个蛋白质复合物中的蛋白质对的集合构成反例集。这里需要说明的是，在随机选取 CORUM 中的蛋白质复合物时，由于剪接体(Spliceosome)和核糖体(Ribosome)两种复合物包含较多的蛋白质，选择它们会导致大部分训练样本来自这两个复合物，因此将其排除，不予选择，如此构造的反例集要远大于正例集。为了使得正反例集平衡，随机选取与正例集 Pos(Positive Set)等量的反例集 Neg(Negative Set)，本文设定 $|\{Pos\}|=|\{Neg\}|=2\,000$。对于 CORUM，则随机选取一半作为参考集，余下的部分将作为实验结果的“标准答案”。接下来，以上述构造的数据作为训练集，其中每对 PPI 对应一个 39 维的相关系数向量作为特征向量，应用 RF 分类器来训练模型。此处须设置 RF 分类器的参数，包括树的数目 n 和树的最大深度 $depth$。在得到分类模型之后，将其应用于 122×10^4 对 PPI 数据，每对 PPI 得到一个置信分数，据此对 PPI 数据进行排序，便得到 PPI 排序列表，分数越高说明蛋白质之间越可能存在相互作用关系。在评估阶段，分别选取排在前 1 000，2 000，…，10 000 的 PPI 与 CORUM 参考集进行对比，得到有监督学习的 PPI 预测结果。为了能够从定量角度准确、具体地对结果进行分析，计算了归一化的 AUC 值。

在实验之前，有以下两点需要特别考虑。

1)此方法采用 RF 分类器，存在参数选取的问题，参数是否会对实验结果造成重要的影响有待验证。如果存在影响，那么如何选取合适的参数以得到最优的模型。

2)模型的正反例集的构造存在随机性，这种随机性是否会对实验结果造成影响。

1) Human Soluble Protein Complex

2) http://human.med.utoronto.ca

3) http://human.med.utoronto.ca

针对上述两个问题，分别通过实验进行验证。

首先，探究参数的影响。随机生成一组正、反例集，基于此数据集，分别测试 RF 在不同参数下的结果。令 $n=100$，$depth$ 取值分别为 2，4，6，8 和 10，结果如图 6 所示。可以看出，在 RF 分类器选取不同参数的情况下，曲线的走向基本一致，且 AUC 值在 0.67 到 0.75 之间浮动，说明参数对模型有一定影响但不是最主要的因素。当树的最大深度 $depth$ 从 2 逐渐增大到 8 时，对应的 AUC 值从 0.67 逐渐增大到0.75；当 $depth$ 从 8 继续增大到 10 时，AUC 值从 0.75 逐渐减小到 0.71。因此，当 $depth$ 取值为 8 时，AUC 值最大，效果较好。

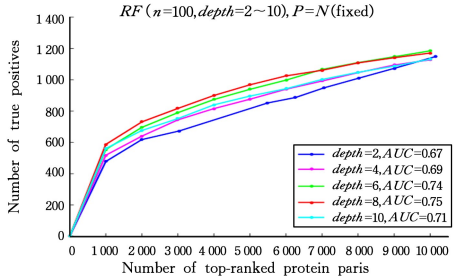
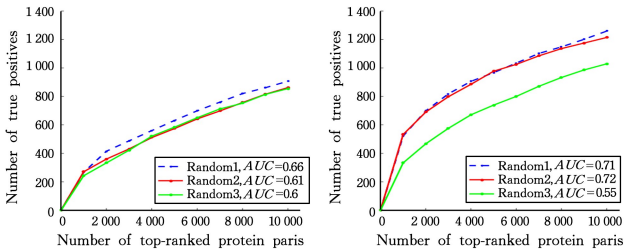


图 6　正、反例集固定时，参数 $depth$ 对结果的影响（电子版为彩色）

Fig. 6　Effect of $depth$ on the performance when positive set and negative set are fixed

接下来，探究训练集构造的随机性是否会对实验结果造成影响。固定 RF 的参数，每次随机生成正、反例集，并对结果进行比较。图 7 分别为当 $n=100$ 时，$depth=2$ 和 $depth=8$ 的实验结果。从图中可以看出，对于两组参数，当参数一定时，结果的曲线走势一致。当参数 $n=100$，$depth=2$ 时，AUC 值在 0.6 到 0.66 之间，没有较大幅度的波动；当参数 $n=100$，$depth=8$ 时，AUC 值在 0.55 到 0.72 之间，结果差异较为明显。此外，对比两组参数的结果，当随机选取正、反例集时，$depth=8$ 的结果要好于 $depth=2$ 的结果，这也与图 6 所描述的一致。



(a) RF($n=100$，$depth=2$)　　(b) RF($n=100$，$depth=8$)

图 7　正、反例集的随机选取对结果的影响

Fig. 7　Effect on performances when positive set and negative set are selected at random

综合上述实验结果，当 RF 的参数 n 为 100，$depth$ 为 8 时，实验结果最好。

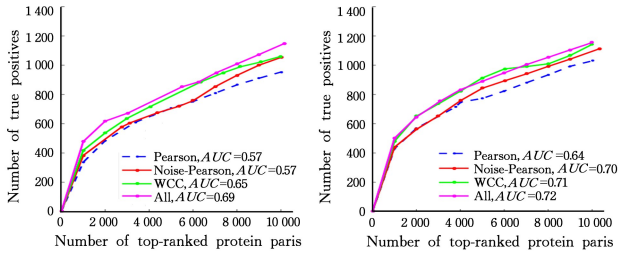
对于有监督学习方法，实验分别针对以下两种情况进行测试：

1）使用 RF 分类器依次对 3 种相关系数方法的结果进行 PPI 预测；

2）使用 RF 分类器对 3 种相关系数方法的总体结果进行 PPI 预测。

情况 1）中，每组相关系数对应 13 组结果，即每对 PPI 对应一个 13 维的 PPI 向量，对应的 PPI 对数分别为：871 273，417 997 和 1 184 620。情况 2）是 3 种相关系数的总体结果，对应 122×10^4 对 39 维 PPI 向量。

对于 RF 分类模型，分别取 $n=100$，$depth=2$ 和 $n=100$，$depth=8$ 两组参数，对之前 3 种方法的结果进行 PPI 预测，实验结果如图 8 所示。实验中的正、反例集是随机生成的，同时保证了正、反例集是固定的，然后用于两组参数的训练。



(a) RF($n=100$，$depth=2$)　　(b) RF($n=100$，$depth=8$)

图 8　基于有监督学习方法的实验结果

Fig. 8　Performance of supervised learning method

从图 8 中可以看出，在相同参数情况下的实验结果如下。

1）不同相关系数方法的结果有所差异。其中，Pearson 相关系数的结果最差，且与另外两种相关系数的结果差距较大。这说明此方法的准确性与相关系数的计算方法有关。这是由于用于 RF 模型分类的 PPI 数据来源于相关系数的计算，每个 PPI 的相关系数向量作为特征向量，在有监督学习方法中，结果依赖于特征向量的取值。因此，这些用于计算特征向量的相关系数方法起着关键性的作用。在这 3 种相关系数的计算方法中，Pearson 相关系数实质上为计算两个变量的协方差与标准差的商，两个蛋白质的 Fraction 向量越相近，则其对应的 Pearson 相关系数越大。由于初始 PCP-MS 数据集是大规模稀疏矩阵，许多蛋白质的 Fraction 向量是十分稀疏的，这些蛋白质之间的 Pearson 相关系数较大，从而导致大量实际上不具有相互作用关系的蛋白质对被误认为存在相互作用关系。而基于噪声的 Pearson 相关系数和 Wcc 系数对初始数据分别进行了加噪声处理和考虑了偏移的问题，因此效果好于 Pearson 相关系数。

2）与单一相关系数结果相比，总体结果更好。这是由于该方法为有监督学习方式，特征的数量对模型效果会产生一定的影响。当使用 39 维的特征时，训练得到的模型要优于使用 13 维特征的情况。

此外，两组参数结果相比，$depth=8$ 的结果更好，这也与前面的实验结果一致。

（2）基于无监督学习方法

首先，对 MS 数据进行相关性分析。分别计算 13 组实验中蛋白质之间的 Pearson 相关系数、基于噪声的 Pearson 相关系数和 Wcc 系数，保留系数不小于 0.5 的 PPI。每种相关系数方法对应 13 组 PPI 相关系数列表。然后，通过 Reinforce 排序整合方法对多组排序列表进行整合。

关于整合，针对以下两种情况进行测试：

1）使用 Reinforce 依次对 3 种相关系数方法的结果进行整合；

2）使用 Reinforce 对 3 种相关系数方法的总体结果进行整合。

基于无监督学习方法的实验结果如图 9 所示。可以看出,3 种相关系数方法及其组合数据的结果并没有表现出明显的差异,4 组曲线走势一致,且相应的 *AUC* 值也在 0.54 到 0.56 之间。实验结果表明基于无监督学习方法对相关系数的计算并不敏感,具有鲁棒性。

(3)无监督学习方法与有监督学习方法的比较

为了验证本文方法的有效性,进一步将无监督学习方法与有监督学习方法进行比较。对于用 3 种不同相关系数方法得到的结果,在 RF 方法参数 *depth* 取 2 和 8 时与 Reinforce 方法进行比较。实验结果如图 10 所示。

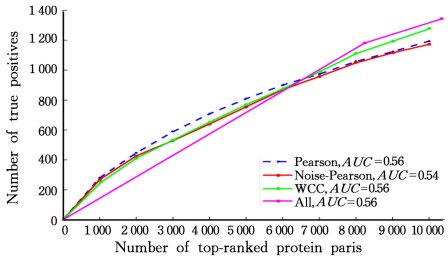


图 9 基于无监督学习方法的实验结果

Fig. 9 Performance of unsupervised learning method

根据对 *AUC* 值的比较,在分别使用 3 个相关系数(包括 Pearson,Noised-Pearson 和 Wcc)进行相关性分析时,Reinforce 均不弱于有监督学习方法在参数 *depth* 为 2 时的结果。图 10(d)选取总体数据时,由于 Reinforce 在 top-8243 时的分数相同,导致曲线从横坐标 8243 开始绘制,前面的曲线被拟合为直线。这是由于 Reinforce 方法在假设条件下统计 $T''>T$ 值时,T''为随机生成的排序向量,因此会导致产生的分数为 0。虽然此图的 *AUC* 值要低于有监督学习方法,但从图像中可以看出该方法可以预测更多的 PPI。在 PPI 网络推断研究中,高效、可靠地预测出具有真实相互作用关系的蛋白质对 PPI 网络构建起着至关重要的作用。

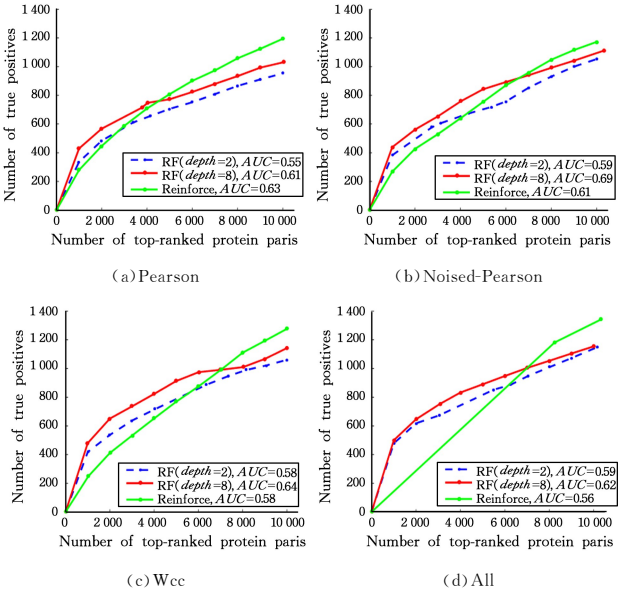


图 10 Reinforce 与 RF 方法的结果比较

Fig. 10 Effect on performances when positive set and negative set are selected at random

综合上述实验结果,本文提出的基于相关性分析与排序整合的方法在不使用参考集的情况下,可以有效地预测更多具有真实相互作用关系的 PPI。与基于有监督学习方法相比,在某些参数情况下,本文方法的结果要略胜一筹。此外,本文方法属于无监督学习方法,避免了参数选取调优和过拟合等问题。该方法具有很强的鲁棒性,能够有效地构造出可靠的 PPI 网络。

结束语 PCP-MS 技术是近年来用于 PPI 网络构建的新兴技术手段。该方法应用分馏技术对蛋白质复合物进行分离,再将共洗脱得到的蛋白质进行质谱分析,以识别这些蛋白质,获取蛋白质之间的相互作用关系,从而实现 PPI 网络的构建。目前,虽然已有一些针对 PCP- MS 数据进行 PPI 网络构建的研究工作,但是针对这一问题的研究还处于初级阶段,相关方法较少。此外,现有的基于有监督学习方式,在训练过程中引入了参考标准数据,这种方式的合理性值得商榷。

针对 PCP-MS 数据的 PPI 网络构建问题,本文提出了一种基于相关性分析和排序整合的 PPI 网络构建方法。该方法包括以下两个步骤:(1)相关性分析。以初始 PCP-MS 实验数据作为输入,计算蛋白质之间的相关系数,得到多组实验对应的 PPI 结果。(2)排序整合。采用 Reinforce 方法将多组 PPI 结果整合为一个 PPI 分数,用以表示蛋白质相互作用关系的程度。实验结果表明,该方法能够与已有的有监督学习方法相媲美。本文提出的方法为无监督学习方法,且没有借鉴标准参考数据,在方法的可解释方面比有监督学习方法更胜一筹;此外,也可以避免有监督学习方法中可能产生的过拟合等问题。在未来的工作中,如何扩展模型的应用范围是值得研究的问题。

参 考 文 献

[1] GUAN W,WANG J,HE F C. The advance in research methods for large-scale protein-protein interactions [J]. Chinese Bulletin of Life Sciences,2006,18(5):507-512. (in Chinese)
关薇,王建,贺福初. 大规模蛋白质相互作用研究方法进展[J]. 生命科学,2006,18(5):507-512.

[2] KIM M S,PINTO S M,GETNET D,et al. A draft map of the human proteome [J]. Nature,2014,509(7502):575-581.

[3] WILHELM M,SCHLEGL J,HAHNE H,et al. Mass-spectrometry-based draft of the human proteome [J]. Nature,2014,509(7502):582-587.

[4] BAKER M. Proteomics:The interaction map [J]. Nature,2012,484(7393):271-275.

[5] MIRZAEI H,CARRASCO M. Modern Proteomics-Sample Preparation,Analysis and Practical Applications[M]. Springer International Publishing,2016.

[6] MEHTA V,TRINKLE-MULCAHY L. Recent advances in large-scale protein interactome mapping[J]. F1000research,2016,5:782.

[7] FAN S B,WU Y J,YANG B,et al. A New Approach to Protein Structure and Interaction Research:Chemical Cross-linking in Combination With Mass Spectrometry [J]. Progress in Biochemistry and Biophysics,2014,41(11):1109-1125. (in Chinese)
樊盛博,吴妍洁,杨兵,等. 蛋白质结构与相互作用研究新方法——交联质谱技术[J]. 生物化学与生物物理进展,2014,41(11):1109-1125.

[8] HUTTLIN E L,TING L,BRUCKNER R J,et al. The BioPlex Network:A Systematic Exploration of the Human Interactome. [J]. Cell,2015,162(2):425-440.

[9] HUTTLIN E L,BRUCKNER R J,PAULO J A,et al. Architecture of the human interactome defines protein communities and disease networks:[J]. Nature,2017,545(7655):505-509.

[10] BEHRENDS C,SOWA M E,GYGI S P,et al. Network organization of the human autophagy system [J]. Nature, 2010, 466(7302):68-76.

[11] JÄGER S,CIMERMANCIC P,GULBAHCE N,et al. Global landscape of HIV-human protein complexes [J]. Nature,2012, 481(7381):365-370.

[12] SOWA M E,BENNETT E J,GYGI S P,et al. Defining the human deubiquitinating enzyme interaction landscape [J]. Cell, 2009,138(2):389-403.

[13] GURUHARSHA K G,RUAL J F,ZHAI B,et al. A protein complex network of Drosophila melanogaster [J]. Cell, 2011, 147(3):690-703.

[14] TENG B,ZHAO C,LIU X,et al. Network inference from AP-MS data:computational challenges and solutions [J]. Briefings in Bioinformatics,2015,16(4):658-674.

[15] CHEN B,FAN W,LIU J,et al. Identifying protein complexes and functional modules—from static PPI networks to dynamic PPI networks [J]. Briefings in Bioinformatics,2014,15(2):177-194.

[16] JI J,ZHANG A,LIU C,et al. Survey:Functional Module Detection from Protein-Protein Interaction Networks [J]. IEEE Transactions on Knowledge & Data Engineering,2014,26(2): 261-277.

[17] VARJOSALO M,SACCO R,STUKALOV A,et al. Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized AP-MS [J]. Nature Methods,2013,10(4): 307-314.

[18] SHARAN R,ULITSKY I,SHAMIR R. Network-based prediction of protein function [J]. Molecular Systems Biology,2007, 3(1):88.

[19] BARABÁSI A L,GULBAHCE N,LOSCALZO J. Network medicine:a network-based approach to human disease [J]. Nature Reviews Genetics,2011,12(1):56-68.

[20] TAYLOR I W,LINDING R,WARDE-FARLEY D,et al. Dynamic modularity in protein interaction networks predicts breast cancer outcome [J]. Nature Biotechnology,2009,27(2):199-204.

[21] HE Z,YU W. Stable feature selection for biomarker discovery [J]. Computational Biology and Chemistry,2010,34(4):215-225.

[22] NESVIZHSKII A I. Computational and informatics strategies for identification of specific protein interaction partners in affinity purification mass spectrometry experiments [J]. Proteomics, 2012,12(10):1639-1655.

[23] ARMEAN I M,LILLEY K S,TROTTER M W B. Popular computational methods to assess multiprotein complexes derived from label-free affinity purification and mass spectrometry (AP-MS) experiments [J]. Molecular & Cellular Proteomics,2013, 12(1):1-13.

[24] WAN C,BORGESON B,PHANSE S,et al. Panorama of ancient metazoan macromolecular complexes [J]. Nature, 2015, 525(7569):339-344.

[25] HAVUGIMANA P C,HART G T,NEPUSZ T,et al. A census of human soluble protein complexes [J]. Cell,2012,150(5): 1068-1081.

[26] DE GELDER R,WEHRENS R,HAGEMAN J A. A generalized expression for the similarity of spectra:application to powder diffraction pattern classification [J]. Journal of Computational Chemistry,2001,22(3):273-289

[27] TIAN B,DUAN Q,ZHAO C,et al. Reinforce:An Ensemble Approach for Inferring PPI Network from AP-MS Data [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics,2017,PP(99):1-1.

[28] KOLDE R,LAUR S,ADLER P,et al. Robust rank aggregation for gene list integration and meta-analysis [J]. Bioinformatics, 2012,28(4):573-580.

[29] STOREY J D. A direct approach to false discovery rates [J]. Journal of the Royal Statistical Society,2002,64(3):479-498.

[30] RUEPP A,WAEGELE B,LECHNER M,et al. CORUM:the comprehensive resource of mammalian protein complexes—2009 [J]. Nucleic Acids Research,2009,38(suppl_1):D497-D501.