PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2016, 43(7): 661~672
www.pibb.ac.cn

目标-诱饵库搜索策略在蛋白质组质谱鉴定和质控中的应用及研究进展 *

冯晓东[1,2] 马 洁[2] 常 乘[2] 白明泽[1] 朱云平[2)**] 舒坤贤[1)**]

([1]) 重庆邮电大学生物信息学研究所，重庆 400065；
[2)] 军事医学科学院放射与辐射医学研究所，蛋白质药物国家工程研究中心，北京蛋白质组研究中心，蛋白质组学国家重点实验室，国家蛋白质科学中心(北京)，北京 102206)

摘要 基于串联质谱的蛋白质组研究会产生海量的质谱数据，这些数据通常使用数据库搜索引擎进行鉴定分析，并根据肽段谱图匹配(PSM)反推真实的样品蛋白质．对于高通量蛋白质组数据的处理，其鉴定结果的可信是后续分析应用的前提，因此对鉴定结果的质量控制尤为重要，而基于目标 - 诱饵库(target-decoy)搜索策略的质量控制是目前应用最为广泛的方法．本文首先介绍了基于目标 - 诱饵库搜索策略搜库和质量控制的实施流程，然后综述了基于目标 - 诱饵库搜索策略的质量控制工具，并提出了目标 - 诱饵库搜索策略的不足及改善方法，最后对目标 - 诱饵库搜索策略进行了总结与展望．

关键词 串联质谱，蛋白质鉴定，目标 - 诱饵，质量控制
学科分类号 Q51，Q811.4
DOI: 10.16476/j.pibb.2016.0067

近年来，基于质谱平台进行蛋白质组测序已经成为主流，随着质谱仪器的进步和发展，实验产出质谱数据的精度和通量也日益提高．海量的质谱数据通常会被送入数据库搜索引擎进行谱图鉴定，之后，通常选取打分值最高的肽段谱图匹配(peptide spectrum matches，PSM)作为最佳匹配，而有研究指出肽段谱图匹配中仅有 20%～40%为可信匹配[1]．因此从所有 PSM 中分离出真正正确的鉴定结果显得尤为重要．为解决这个问题，Benjamini 等[2]提出了假阳性率的概念．假阳性率表征了所有通过打分阈值的匹配中假阳性匹配所占的比例，它是针对整体数据置信度水平的一个估计值．基于目标 - 诱饵库(target-decoy)搜索策略的质量控制[3]方法是目前应用最广泛的估计假阳性率的方法，其主旨是构建实际并不存在的诱饵库(decoy)并与目标库(target)一起送入搜库引擎进行搜索，通过在序列库中加标签的方式来识别 PSM 究竟是来自诱饵库还是目标库，并应用于后续质控方法模型的构建中．目标 - 诱饵库搜索策略的一个基本假设是目标库中的错误匹配数目与诱饵库中的错误匹配数目相等[3]，该策略不仅可以估计数据集的假阳性率，也可以指导研究人员设定合适的过滤候选肽段的打分阈值．通过目标 - 诱饵库策略发展的质控方法由于可移植性强，适应于多种实验平台，灵敏度高而广泛应用于蛋白质组高通量质谱研究中[4-5]．本文首先介绍了基于目标 - 诱饵库搜索策略搜库和质量控制的实施流程，然后介绍了基于目标 - 诱饵库搜索策略的质量控制工具，其次提出了目标 - 诱饵库搜索策略的不足及改善方法，最后对目标 - 诱饵库搜索策略进行了总结与展望．

1 基于目标-诱饵库策略搜库和质量控制的实施流程

基于目标 - 诱饵库策略搜库和质量控制的实施流程如图 1 所示．首先根据质谱数据下载相应的目

* 国际合作计划(2014DFB30010)、国家重点基础研究发展计划(973)(2013CB910800), 国家自然科学基金(21475150, 21105121, 61501071)资助项目和重庆市研究生科研创新项目(CYS14154)资助.
** 通讯联系人.
朱云平. Tel: 010-61777058, E-mail: zhuyunping@gmail.com
舒坤贤. Tel: 023-62468319, E-mail: shukx@cqupt.edu.cn
收稿日期：2016-03-01，接受日期：2016-05-06

标蛋白序列数据库，并以此构建诱饵蛋白序列数据库，构库方法大致分为反转、随机和混编三种．之后利用搜库引擎进行鉴定，质控工具根据搜库之后的结果计算假阳性率并以此进行肽段层面和蛋白质层面的质量控制，得到最终的结果．

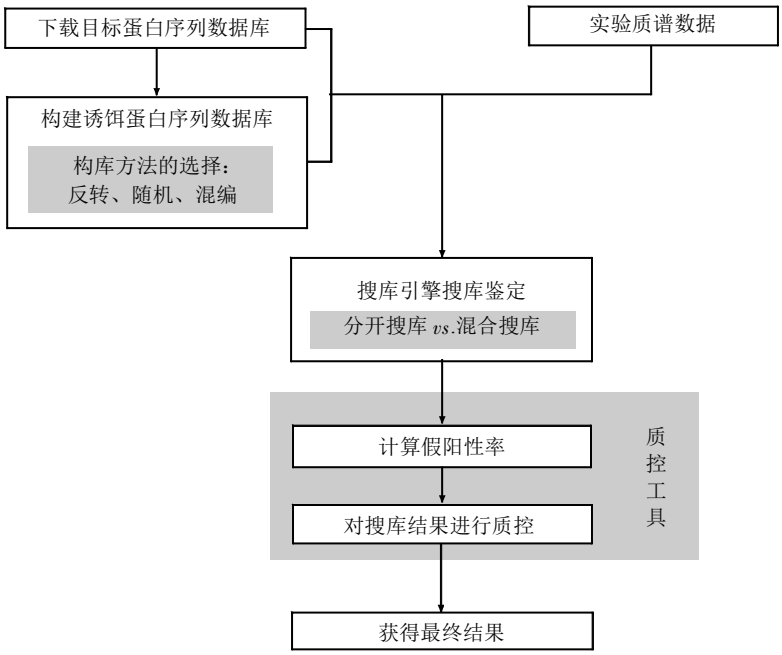

Fig. 1 Flow chart of target-decoy searching strategy

图 1 目标-诱饵库策略的实施流程图

将构建好的蛋白质序列数据库与质谱数据一起送入搜库引擎进行搜库鉴定，然后利用质控工具对搜库结果进行质量控制，获得最终结果.

1.1 实验数据的准备

1.1.1 质谱数据

质谱数据既可以通过质谱仪器产出，也可以从质谱公共数据库中下载．目前常用的质谱仪包括 ThermoFisher 公司生产的 LTQ 系列，AppliedBiosystems 公司生产的 QSTAR 系列以及 Bruker Daltonics 公司生产的 FLEX 系列等．常用的质谱公共数据库包括 PRIDE Archive[6]、PeptideAtlas[7]、Open Proteomics Database[8]和 iProX[9]等．

1.1.2 目标蛋白序列数据库

质谱数据需要在搜库引擎中与蛋白质序列数据库进行比对，在选取蛋白质序列数据库时需要与质谱数据相吻合．蛋白质序列库可以从多个公共数据库下载，其中最常用的有 UniProt/Swissprot[10]和 NCBI 的 RefSeq[11]．

1.2 构建目标-诱饵序列库

创建诱饵序列的方法有很多[3, 12-13]，大致可分为反转、混编和随机三种类型．理想的诱饵序列应具备与目标序列相似的氨基酸分布、蛋白质长度分布，并且诱饵序列与目标序列之间没有重叠[14]．Elias 等[3]比较了反转法和随机法两种构建诱饵库的方法，发现这两种方法构建的诱饵库能够产生相近的真阳性匹配，但由于反转法保留了序列的氨基酸频率、蛋白质和肽段长度分布，因此推荐采用反转法来构建诱饵库．其他研究也指出，用反转法构建的诱饵库与目标库中的氨基酸组成和共享肽段分布一致，而且反转法容易实现、更具确定性[14-15]．另有研究指出，当采用单一过滤指标(如只考虑 SEQUEST 搜库引擎的 $Xcorr$ 特征参数)时采用随机化的方式构建诱饵库会高估假阳性率和假阴性率，但采用多过滤指标时，使用反转法或者随机法构建诱饵库的区别不大[16]．

1.3 使用搜库引擎对质谱数据进行分析

1.3.1 搜库引擎

质谱数据通常会被送入搜库引擎进行分析，常用的搜库引擎有 SEQUEST[17]、Mascot[18]、

X!Tandem[19]、Omssa[20]、InsPect[21]和 MS-GF+[22]等. 目标 - 诱饵策略的一个优势在于它几乎适用于所有的搜库引擎.

1.3.2　两种数据库搜索策略

根据是将诱饵序列和目标序列整合还是分开进行搜库，目标 - 诱饵策略分为混合搜库和分开搜库两大类.

a．混合搜库．混合搜库策略将诱饵序列和目标序列整合为一个序列库送入搜库引擎进行搜库．对于每一个谱图，搜库引擎必须在目标库和诱饵库中选择最佳匹配．正确的匹配必定来源于目标库，而错误的匹配既可能来源于目标库也可能来源于诱饵库.

b．分开搜库．分开搜库策略使用搜库引擎对目标序列库和诱饵序列库分别进行搜索．所有高于打分阈值的来自目标序列库和诱饵序列库的 PSM 都被用来计算假阳性率(*FDR*).

1.3.3　两种数据库搜索策略的比较

混合搜库允许目标序列库的 PSM 和诱饵序列库的 PSM 相互竞争，从而淘汰掉部分诱饵序列库的 PSM，因此有研究人员推荐使用混合搜库策略[15]; 但也有研究人员指出，这种竞争机制会导致部分高打分值的目标序列库的 PSM 被更高打分值的诱饵序列库的 PSM 竞争掉，因而采用混合搜库的方法相较于分开搜库的方法会产生更高的假阳性率，因此推荐采用分开搜库策略[23]．分开搜库或混合搜库孰优孰劣并没有定论，从用户的角度来讲，可以按照搜库引擎的推荐选择适宜的搜库方式．比如 Mascot 搜库引擎可以根据目标序列自动生成诱饵序列并对目标序列和诱饵序列进行分开搜库，省去了用户手动构建目标 - 诱饵混合库的麻烦，而且其内嵌的质控程序 Mascot Percolator 也是针对分开搜库的结果进行质控的.

1.4　根据肽段谱图匹配数目计算具体参数

1.4.1　目标 - 诱饵策略中的 4 类 PSM 分布及由此产生的 4 种参数

如图 2 所示的圆形包含目标 - 诱饵策略中搜库之后的 4 类 PSM 分布．右下角的扇形代表真阳性匹配(true positive，TP)，即过质控标准的，实际上也是正确的匹配．左下角的扇形代表假阳性匹配(false positive，FP)，即过质控标准的，但实际上是错误的匹配．右上角的扇形代表假阴性匹配(false negative，FN)，即未过质控标准的，但实际上是正确的匹配．左上角的扇形代表真阴性匹配(true negative，TN)，即未过质控标准的，实际上也是错误的匹配．实线矩形和虚线矩形分别圈出了正确匹配总数(FN+TP)和错误匹配总数(TN+FP)．质量控制就是在保证假阳性鉴定数(FP)低于一定水平的基础上(比如卡 FDR 低于 1%)，尽可能降低假阴性鉴定数(FN)，提高真阳性鉴定数(TP).

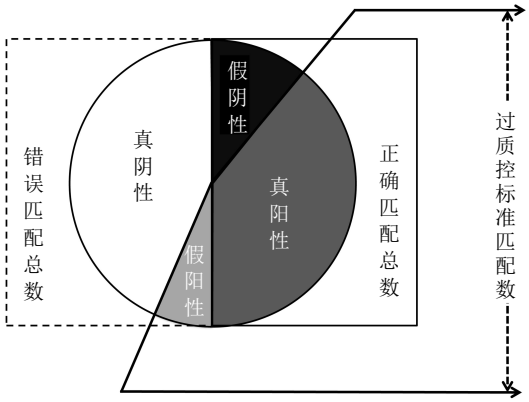

Fig. 2　PSM distribution in target-decoy database searching results

图 2　基于目标-诱饵策略搜库之后的 PSM 分布

基于目标 - 诱饵策略搜库之后的 PSM 分为真阴性 PSM(TN)、假阳性 PSM(FP)、真阳性 PSM(TP)和假阴性 PSM(FN)．目标 - 诱饵策略中涉及的参数依据这 4 类 PSM 计算而来.

目标 - 诱饵策略中常用的 4 种参数根据图 2 所示的 4 类 PSM 计算而来，其参数的定义及计算公式如表 1 所示．其中精确度表征了所有通过质控标准的匹配中真阳性匹配所占的比例，假阳性率表征了所有通过质控标准的匹配中假阳性匹配所占的比

Table 1　Definition and calculation formula of relative parameter in target-decoy strategy

表 1　目标-诱饵策略中相关参数的定义及计算公式

参数	计算公式	定义
精确度(Precision)	TP/(TP+FP)	真阳性匹配数除以真阳性匹配数与假阳性匹配数之和
假阳性率(FDR)	FP/(TP+FP)或者 1-Precision	假阳性匹配数除以真阳性匹配数与假阳性匹配数之和
敏感性(Sensitivity)	TP/(TP+FN)	真阳性匹配数除以真阳性匹配数与假阴性匹配数之和
准确性(Accuracy)	(TP+TN)/(TP+FP+TN+FP)	真阳性与真阴性匹配数之和除以所有匹配数之和

例，敏感性表征了真阳性匹配在所有正确匹配中所占的比例，准确性表征了所有匹配中被正确鉴定的匹配所占的比例.

表 1 所列 4 种参数都可以从整体水平上评估 PSM 的质量. 而其中应用最广泛的是假阳性率(FDR). 下一节将详细介绍 5 种假阳性率的计算公式.

1.4.2 假阳性率的 5 种计算公式及适用范围

根据搜库策略的不同，假阳性率的计算公式有所差异. 其中公式 1 是混合搜库常用的计算公式，公式 2 是分开搜库常用的计算公式，公式 3、公式 4 是在公式 2 的基础上进行改进而来，公式 5 是在公式 1 的基础上进行改进而来.

公式 1(concatenated *FDR*，FDR_c)是搜索混合的目标 - 诱饵数据库时常用的假阳性率(*FDR*)计算公式[14]. 这种方法认为，通过某一打分阈值的来自诱饵库的 PSM 数目(*D*)与来自目标库的 PSM 数目(*T*)中的错误匹配是相等的. 因此将来自诱饵库的 PSM 数目(*D*)与来自目标库的 PSM 数目(*T*)相加，其中所包含的错误匹配数应该是诱饵库 PSM 数目的 2 倍(2**D*).

$$FDR_c = \frac{2\times D}{T+D} \tag{1}$$

公式 2(separate *FDR*，FDR_s)是搜索分开的目标 - 诱饵数据库时常用的假阳性率(*FDR*)计算公式[13]. 这种方法同样认为通过某一打分阈值的来自诱饵库的 PSM 数目(*D*)与来自目标库的 PSM 数目(*T*)中的错误匹配是相等的. 于是认为 *T* 中的错误匹配数为 *D*.

$$FDR_s = \frac{D}{T} \tag{2}$$

公式 3(*FDR* with *PIT* correction，FDR_{pit})是在公式 2(FDR_s)的基础之上做了修正，因为并非所有来自目标库的 PSM 都是正确的，所以在估计假阳性率时应该乘以一个修正系数 *PIT*(the percentage of incorrect targets)，*PIT* 代表目标序列库 PSM 中不正确 PSM 所占的比例[13].

$$FDR_{pit} = PIT\times\frac{D}{T} \tag{3}$$

公式 4(refined seperate *FDR*，FDR_{rs})也是在公式 2(FDR_s)的基础上做了修正. Navarro 等[24]指出，如果一个谱图对应的诱饵库 PSM 打分值并没有超过对应的目标库 PSM，那么即使它的打分值高于打分阈值，也不应该被考虑，否则会高估假阳性率. 其中 *DO*(decoy only)代表只存在于诱饵库中且打分值高于打分阈值的 PSM 个数，*TO*(target only)代表只存在于目标库中且打分值高于打分阈值的 PSM 个数，*TB*(target better)代表同时存在于目标库和诱饵库中但目标库中分值更高的 PSM 个数，*DB*(decoy better)代表同时存在于目标库和诱饵库中但诱饵库中分值更高的 PSM 个数.

$$FDR_{rs} = \frac{2\times DB+DO}{TB+TO+DB} \tag{4}$$

公式 5(refined concatenated *FDR*，FDR_{rc})是在公式 1(FDR_c)的基础之上做了修正. Cerqueira[25]指出，既然诱饵库 PSM 必然是错误的，那么它们在 FDR 估计过程中就不该被考虑.

$$FDR_{rc} = \frac{D}{T-D} \tag{5}$$

1.4.3 其他常用参数

不同的质控工具中还经常使用下面几种参数表征结果的置信度水平，以弥补仅使用 *FDR* 进行质控的不足.

P 值是常用的表征某一 PSM 是由随机匹配产生的概率大小的参数. *P* 值越小，说明该 PSM 为随机匹配的概率越小. 例如：采用 SEQUEST 搜库引擎进行搜库，打分阈值卡 *XCorr* 为 3.0 时，匹配诱饵库的所有 34 492 个 PSM 中只有 219 个 PSM 的 *XCorr* 打分值高于 3.0，那么 *XCorr* 为 3.0 的 PSM 对应的 *P* 值为 219/34 492=0.0063.

通常情况下，随着搜库引擎主打分的下降，鉴定结果 *FDR* 的计算值逐渐增大. 但由于 *FDR* 不是关于肽段图谱匹配分值的单调函数，不同的打分值可能对应相同的 *FDR*，为了解决这个问题，引入了 *q*-value 来考察单个 PSM 的置信度[13]. *q*-value 定义为某一个 PSM 能够被接收的最小 *FDR*.

PEP(posterior error probability，后验错误概率)[26]表示通过打分阈值的某一 PSM 错误的可能性大小. 举例来说，如果一个 PSM 的 *PEP* 为 1%，那么它有 99%的可能性是正确的匹配. *PEP* 代表了单个 PSM 的错误概率，因此也被称作 Local*FDR*.

1.5 基于目标-诱饵策略对搜库结果进行质量控制

目标 - 诱饵策略出现前，研究人员常基于经验规则确定打分阈值来判定某一 PSM 的正误，比如采用搜库引擎 SEQUEST 搜库，打分阈值卡特征参数 *XCorr* 为 3.0，这个 3.0 即是根据经验确定的. 而基于目标 - 诱饵策略进行质控的思路是将 PSM 按照某一特征参数打分值进行排序. 对于每一个打分值，统计通过该打分值的目标 PSM 和诱饵 PSM 数目并根据 **1.4.2** 节中 5 种公式的其中一种计算其

对应 *FDR*．当某一个打分值对应 *FDR* 达到预设的 *FDR*(比如 1%)时，确定该打分值为打分阈值，并将通过打分阈值的 PSM 认定为正确 PSM，未通过打分阈值的 PSM 认定为错误 PSM．由于手工计算量过大，所以通常借助质控工具实现．目前有多种质控工具被开发出来以实现质量控制功能，算法各不相同，详见下文介绍．

2　基于目标-诱饵策略的质控工具

使用搜库引擎进行质谱数据分析的一个重大难题是如何在保证数据集置信度的基础上有效区分正确 / 错误鉴定结果．大部分搜库引擎仅按照其模型输出一个或多个打分参数作为衡量鉴定结果标准的参考，并不能保证匹配的正确性，往往需要利用第三方质控工具对直接给出的鉴定结果进行进一步过滤．采用目标 - 诱饵策略评估不同 PSM 质量的优劣，是目前常用的质控方法，根据其质控对象的不同，可以分为肽段层面的质控工具和蛋白质层面的质控工具，目前已经发展的主要质控工具参见表 2．

Table 2　List of quality control tools based on the target-decoy strategy

表 2　基于目标-诱饵策略发展的质控工具列表

类别	工具	网址	参考文献
肽段层面单搜库引擎质控	PeptideProphet	https://www.systemsbiology.org/resources/software-downloads/	[27−29]
	FlexiFDR	http://sourceforge.net/projects/mssuite/files/	[30]
	APE	http://omics.pnl.gov/software/APE.php	[31]
	Percolator	http://noble.gs.washington.edu/proj/percolator/	[32]
	Qranker	http://cruxtoolkit.sourceforge.net/	[33]
	MascotPercolator	http://www.sanger.ac.uk/resources/software/mascotpercolator/	[34−35]
	PepDistiller	http://www.bprc.ac.cn/pepdistiller/	[36]
	CRanker	无下载网址但可联系 liangxijunsd@163.com 索取	[37]
	X!Tandem Percolator	–	[38]
	MS-GF+Percolator	http://proteomics.ucsd.edu/software-tools/	[39]
	OmssaPercolator	http://sourceforge.net/projects/omssapercolator/	[40]
	De-Noise	–	[41]
	FC-Ranker	–	[42]
	Mfs	–	[43]
	RockerBox	https://trac.nbic.nl/rockerbox/	[44]
肽段层面多搜库引擎质控	FDRAnalysis	http://www.ispider.manchester.ac.uk/FDRAnalysis/FDR_analysis_home.html	[45]
	Oscore	–	[46]
	iProphet	http://www.proteomecenter.org/software.php	[47]
蛋白质层面质控	Proteinprophet	https://www.systemsbiology.org/resources/software-downloads/	[48]
	Mayu	–	[49]
	Raid	ftp://ftp.ncbi.nlm.nih.gov/pub/qmbp/qmbp ms/RAId	[50]
	Provalt	https://kiwi.rcr.uga.edu/tcprot	[51]
	Fido	http://noble.gs.washington.edu/proj/fido/	[52]
	Barista	http://cruxtoolkit.sourceforge.net/	[53]
	IDpicker	http://fenchurch.mc.vanderbilt.edu/	[54]
	BuildSummary	https://github.com/shengqh/RCPA.Tools/releases	[55]

2.1　肽段层面的质控工具

肽段层面基于目标 - 诱饵策略的质控工具可以分为单搜库引擎质控工具和整合多搜库引擎质控工具．而单搜库引擎质控工具按照算法又可以细分为简单参数模型法、概率模型法和机器学习法三类．

2.1.1　肽段层面单搜库引擎质控工具

a．基于简单参数模型法的质控工具

早期的基于目标 - 诱饵策略的质控工具大多为简单参数模型法，通过随机肽段匹配数估计正常数据库鉴定肽段中随机匹配的分布，以此确定搜索引

擎主打分在不同假阳性率条件下的阈值，并将其作为数据集的过滤标准[12, 56]. 2010 年发表的 HLPP 成人肝表达谱数据集就采用了如此的肽段过滤标准：要求所有谱图 - 肽段匹配满足 $\Delta C_n \geqslant 0.08$，基于诱饵数据库的匹配结果，改变 Xcorr 值来确定不同电荷条件下肽段的卡值[57].

Rudnick 等[58]将 Mascot 分值与根据随机匹配肽段数计算的假阳性率进行直线拟合，可外推得到任何假阳性率要求下的过滤标准. 张纪阳等[59]2007 年构建了线性判别模型，可以根据 SEQUEST 的 2 个打分参数 Xcorr 和 ΔC_n 从目标 PSM 中过滤掉假阳性匹配. 相较于对数据进行过滤的方法仅采用 1 个特征参数而言，基于简单参数模型的质控工具由于可以整合多个特征参数，因而其质控效果有了提升.

b. 基于概率模型法的质控工具

Keller 等[60]于 2002 年发展的 PeptideProphet 是典型的基于概率模型过滤肽段结果的方法. 该方法将 SEQUEST 的 4 个特征参数进行线性加权，得到了一个判别函数. 利用贝叶斯公式和先验分布得到图谱 - 肽段之间正确匹配和错误匹配的概率，在此基础之上获得最终鉴定的肽段列表. 为突破 PeptideProphet 在建模估计时使用经验模型的限制，Choi 等[61]提出了 2 种易适应的混合模型：成分变量混合模型和半参数混合模型，并提供了基于目标 - 诱饵数据库搜索结果使用非参建模的分析选项.

FlexiFDR[30]最初是为搜库引擎 MassWiz[62]设计的质控工具，同时适应于分开搜库和混合搜库. FlexiFDR 对不同电荷下的诱饵 PSM 构建线性回归模型，并采用动态校准的方法估计 FDR 以得到合适的打分阈值. APE[31]质控工具首先需要用户定义过滤数据的特征参数集和 *FDR*，统计在给定 *FDR* 条件下采用不同特征参数过滤数据所获得的真阳性匹配数目，选择获得真阳性匹配数目最多的特征参数作为过滤数据的指标.

概率分布模型需要预设概率分布的形式，但是实际应用中总会存在常规分布难以描述的情况. 非参概率密度函数拟合提供了一种通用的概率密度描述方式，其基本思想是利用一系列核函数(如高斯函数)的叠加来任意精度地拟合观测数据的分布. 张纪阳等[63]将目标 - 诱饵策略与非参模型相结合以进行质量控制，最初仅应用非参密度方法估计随机数据库匹配的分布函数，以获得过滤假阳性结果判别函数的候选等高线，在 2009 年的模型中又构建了贝叶斯非参模型以计算后验概率[64]，并于 2010 年将该模型用于处理 Mascot 搜库后的结果[65].

c. 基于机器学习法的质控工具

2002 年的 *Q*score[5]采用半监督式机器学习的方式针对真实的实验数据调整模型参数. 2007 年的 Percolator 进一步发展了半监督式机器学习方式[32]. 它采用支持向量机(SVM)[66]的方式将 SEQUEST 的 20 个特征参数加以综合考虑，最终对于每一个 PSM 给定一个综合评定的 q-value 值. 2009 年，*Q*ranker[33]在 Percolator 基础之上进行了改进，与 Percolator 半监督式机器学习方式不同的是，它采用了全监督式机器学习方式. RockerBox[44]针对大规模数据提供过滤以减少存储和计算量，与 Percolator 保留所有 PSM 数据不同的是，它直接将低质量的数据过滤掉.

Percolator 的一个缺点是不能适用于所有的搜库引擎，因为不同搜库引擎给出的数据集特征参数是不同的. 因此针对不同搜库引擎的 Percolator 相继被开发出来. Mascot Percolator 是针对 Mascot 搜库引擎的质控工具[34-35]. 而 PepDistiller[36]在 Mascot Percolator 的基础上进行了改进：引入了肽段胰酶切端个数(NTT)以提高半酶切搜库的灵敏度；与 Percolator 采用公式 3 计算 FDR 不同的是，PepDistiller 计算 FDR 时采用了公式(4). X!Tandem Percolator 是针对搜库引擎 X!Tandem 设计的质控工具[38]，MS-GF + Percolator 是针对搜库引擎 MS-GF+设计的质控工具[39]，OmssaPercolator 是针对搜库引擎 Omssa 设计的质控工具[40].

De-noise[41]是基于模糊控制的 SVM 机器学习模型，它将低打分值的目标 PSM 认定为噪音 PSM，并在迭代过程中逐渐过滤掉噪音 PSM. 但是 De-noise 需要用户输入较多的参数，FC-Ranker[42]相比 De-Noise 需要更少的来自用户的输入，它使用模糊分类模型估计每个目标 PSM 正确的可能性. CRanker[37]与 FC-Ranker 不同的是将训练误差的权重视为一个变量，同时用户可以根据 FDR 选择高打分值的 PSM. 而 Mfs[43]采用多核心机器学习模型进行质控，它提出模糊隶属关系来抑制目标 PSM 中错误匹配的影响.

2.1.2 整合多搜库引擎的肽段层面质控工具

上述 **2.1.1** 节中介绍的质控工具都是针对单搜库引擎的结果进行处理的，然而不同的搜库引擎即使是处理同一批质谱数据产生的结果也会有差异. 如果能综合考虑多种搜库引擎产生的结果会比单搜

库引擎产生的结果质量更高[67-68]．虽然整合多搜库引擎质控工具计算量大、耗费时间长，但随着现代计算机对数据的处理能力日益增强以及云计算平台的逐渐完善，上述缺陷可以被有效克服．因此，整合多搜库引擎的质控工具在未来将是一种趋势．

Oscore[46]整合了搜库引擎 Sequest 和 Amass[69]的搜库结果．通过逻辑回归模型评估每一个参数的权重，从而估计一个候选 PSM 是正确 PSM 的概率．FDRAnalysis 相较于 Oscore 支持的搜库引擎有所增加[45]，它可以将 Mascot 的 dat 文件、Omssa 的 csv 文件和 X！Tandem 的 xml 文件转化为 mazIdentML 格式或 XML 格式的文件以进行下一步的质控．iProphet 与 Peptideprophet、Proteinprophet 一起整合在 TPP 流程中，是比较成熟的整合多搜库引擎的质控工具．首先 Peptideprophet 对多种搜库引擎产出的结果分别进行肽段水平的质控，Peptideprophet 质控后的多个结果送入 iProphet 进行整合多搜库引擎的质控[47]．

2.2　蛋白质层面的质控工具

蛋白质层面质量控制存在两大挑战．挑战一是同一肽段会对应多个蛋白．以图 3 为例，肽段 2 既对应到蛋白质 *A* 中也对应到蛋白质 *B* 中．所以在反推蛋白质时需要考虑该肽段具体对应哪个蛋白质．挑战二是蛋白质层面的假阳性率并不等同于肽段层面的假阳性率．因为从谱图层面的假阳性率到肽段层面的假阳性率一直到蛋白质层面的假阳性率其数值是逐渐增大的．以图 3 为例：谱图 *1*～*6* 为正确的谱图，谱图 *7* 为错误的谱图，其谱图层面假阳性率为 14%．7 个谱图仅仅对应到 4 个肽段．其中肽段 *1*～*3* 为正确的肽段，肽段 *4* 为错误的肽段，肽段层面假阳性率为 25%．4 个肽段仅仅对应到 3 个蛋白质．其中蛋白质 *A* 和蛋白质 *B* 为正确的蛋白质，蛋白质 *C* 为错误的蛋白，蛋白质层面假阳性率为 33%．因此肽段层面的假阳性率并不适用于蛋白质层面的假阳性率．

有多种蛋白质层面的质控工具被开发出来以解决上述挑战．这些质控工具通常整合了质量控制和蛋白质装配两大功能．蛋白质装配通常采用简约原则，即在反推样品中的蛋白时，采用尽可能少的蛋白来解释该样品[70]．而在蛋白质装配之前进行质量控制是一个必须的过程，不同的质控工具其质控策略有所不同，下面将分别加以介绍．

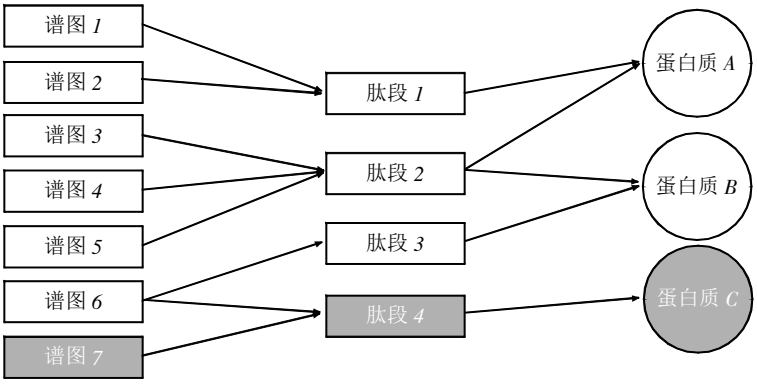

Fig. 3　FDR in the level of PSM, peptide and protein

图 3　谱图层面、肽段层面和蛋白质层面假阳性率示意图

谱图层面假阳性率是 14%，肽段层面假阳性率是 25%，蛋白质层面假阳性率是 33%．假阳性率是逐渐增大的．

Proteinprophet[48]采用概率模型将高打分值的候选肽段组装成候选蛋白并给予每一个候选蛋白一个表征其置信度的打分值．RAId 采用了同 Proteinprophet 类似的策略[50]：在候选肽段组装成蛋白质之前先计算肽段的 *P*value 值，通过计算给出精确的蛋白质层面的 *E*value 值．但是上述工具只考虑了高打分值的候选肽段而忽略了低打分值的肽段信息．如果最高打分值的 PSM 恰巧是不正确的，而第二高打分值的 PSM 才是正确的，如果只按最高打分值进行过滤，那么第二高打分值的正确 PSM 就被过滤掉了．为了解决上述问题，Provalt[51]不再采用先在肽段层面估计 *FDR* 然后再在蛋白质

层面估计 *FDR* 的做法，而是直接计算蛋白质层面的 *FDR*．Provalt 根据用户定义的蛋白质层面的 *FDR* 确定最合适用来判定 PSM 是正确还是错误的打分阈值．Mayu 也将目标 - 诱饵策略应用于蛋白质层面 *FDR* 的计算．它通过诱饵肽段推断出诱饵蛋白从而计算 *FDR*[49]，能够有效解决 FDR 数值从谱图层面到肽段层面一直到蛋白质层面逐渐增大的问题．Barista[53]采用支持向量机的方式综合考虑不同的数据特征值并给出最佳的候选蛋白．由于 Barista 在肽段鉴定环节保留了低打分值的肽段信息，因此在最终的蛋白质装配环节中可以有效地利用这些肽段信息．

大规模的数据集通常包含多个物种的同源序列．即使是采用简约原则，也会产生许多假阳性蛋白．IDpicker[54]能够有效解决这个问题，它需要用户设定仅来源于某一特定蛋白的显著性肽段匹配，从而程序不会再将这些显著性肽段匹配归类到其他蛋白质中去．该方法能大大减少同源性蛋白质的数目．Fido 质控工具采用贝叶斯模型计算错误蛋白后验概率[52]，可以有效解决同一肽段对应多个蛋白质的问题．BuildSummary[55]根据母离子电荷状态、修饰状态、漏切位点数、仪器类型等信息将肽段和蛋白质进行集群．根据用户设定的蛋白质层面 FDR 阈值从不同的分类集群中进行质控，最终给出可信蛋白质列表．它也可以整合多搜库引擎的结果进行综合质控．

3 目标-诱饵策略不足之处及改善方法

3.1 目标-诱饵策略不足之处

目前，利用目标 - 诱饵策略进行 FDR 计算时仍然没有一个公认的流程．主要分歧在于：a．构建诱饵库的方式不确定；b．采用混合搜库还是分开搜库的方式不确定；c．计算 FDR 的公式不确定．具体到蛋白质基因组学和宏蛋白质组学方面，由于其数据量大、序列相似度高，导致从匹配库中筛选掉假阳性匹配很困难．另外质量控制并非仅仅在质谱数据层面，如果产出质谱数据的仪器参数设置不同，处理时长不一样，即使是采用同一套流程进行质控，所得到的结果也有差异，因此有研究人员提出通过监控实验参数对数据进行质控的观点[71]．

3.2 改善方法

针对目标 - 诱饵策略估计 FDR 的不足，许多研究人员提出了改善建议[72-75]．两步搜库法[72]最初是针对宏蛋白质组学和蛋白质基因组学中的大规模数据(>10^6 序列)提出的改善方法．两步搜库法第一步仅使用目标序列库进行搜库，根据搜库结果产生一个较小规模的库，再用较小规模的库构建目标 - 诱饵库进行第二次搜库．另外一种两步搜库法略有不同[74]，它最初是为鉴定新蛋白质而设计的方法，由于序列数据库特别大，所以将目标库和诱饵库分开搜库．第一步搜库后将未匹配的 PSM 从序列数据库中去除掉以减少搜库空间．第二步搜库将精简后的目标库和诱饵库混合起来进行搜库．另外也有构建诱饵库思路的调整：不再人工构建诱饵库，而是根据 PSM 匹配程度将目标库中的低匹配库做诱饵库[76]．

4 总结与展望

目标 - 诱饵策略广泛应用于肽段图谱匹配的假阳性率估计中．基于目标 - 诱饵策略的质控工具也相继被开发出来以满足大批量、自动化运算的需要．当然，目标 - 诱饵策略的应用并不局限于对数据质量的评估．目前多种搜库引擎及质量控制工具如雨后春笋般被开发出来以应用于基于质谱的蛋白质组研究．而几乎每一种搜库引擎或质控工具在刚提出时都会与目前主流的搜库引擎(比如 Mascot、Sequest)或主流的质控工具(如 Peptideprophet、Percolator)进行比较，但是由于都是开发者自己进行的比较，所以其客观性有待考察．目标 - 诱饵策略可以应用于对搜库引擎或质控工具的评估[77-78]，同时目标 - 诱饵策略也可以为实验人员选取合适的搜库引擎及质控工具提供一个参考．

参 考 文 献

[1] Nesvizhskii A I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. Journal of Proteomics, 2010, **73**(11): 2092-2123

[2] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological), 1995, **57**(1): 289-300

[3] Elias J E, Gygi S P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nature Methods, 2007, **4**(3): 207-214

[4] Fenyo D, Beavis R C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. Analytical Chemistry, 2003, **75**(4): 768-774

[5] Moore R E, Young M K, Lee T D. Qscore: an algorithm for

evaluating SEQUEST database search results. Journal of the American Society for Mass Spectrometry, 2002, **13**(4): 378−386

[6] Vizcaino J A, Deutsch E W, Wang R, *et al*. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotechnol, 2014, **32**(3): 223−226

[7] Deutsch E W, Lam H, Aebersold R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. EMBO Reports, 2008, **9**(5): 429−434

[8] Prince J T, Carlson M W, Wang R, *et al*. The need for a public proteomics repository. Nature Biotechnology, 2004, **22**(4): 471−472

[9] Chen T, Zhao J, Ma J, *et al*. Web resources for mass spectrometry-based proteomics. Genomics, Proteomics & Bioinformatics, 2015, **13**(1): 36−39

[10] Uniprot C. The universal protein resource (UniProt). Nucleic Acids Research, 2008, **36**(Database issue): D190−195

[11] O'leary N A, Wright M W, Brister J R, *et al*. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research, 2016, **44**(D1): D733−745

[12] Higdon R, Hogan J M, Van Belle G, *et al*. Randomized sequence databases for tandem mass spectrometry peptide and protein identification. Omics: A Journal of Integrative Biology, 2005, **9**(4): 364−379

[13] Kall L, Storey J D, Maccoss M J, *et al*. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. Journal of Proteome Research, 2008, **7**(1): 29−34

[14] Elias J E, Gygi S P. Target-decoy search strategy for mass spectrometry-based proteomics. Methods in Molecular Biology, 2010, **604**: 55−71

[15] Jeong K, Kim S, Bandeira N. False discovery rates in spectral identification. BMC Bioinformatics, 2012, **13**(Suppl 16): S2

[16] Wang G, Wu W W, Zhang Z, *et al*. Decoy methods for assessing false positives and false discovery rates in shotgun proteomics. Analytical Chemistry, 2009, **81**(1): 146−159

[17] Eng J K, Mccormack A L, Yates J R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of the American Society for Mass Spectrometry, 1994, **5**(11): 976−989

[18] Perkins D N, Pappin D J, Creasy D M, *et al*. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 1999, **20**(18): 3551−3567

[19] Craig R, Beavis R C. TANDEM: matching proteins with tandem mass spectra. Bioinformatics, 2004, **20**(9): 1466−1467

[20] Geer L Y, Markey S P, Kowalak J A, *et al*. Open mass spectrometry search algorithm. Journal of Proteome Research, 2004, **3**(5): 958−964

[21] Tanner S, Shu H, Frank A, *et al*. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Analytical Chemistry, 2005, **77**(14): 4626−4639

[22] Kim S, Mischerikow N, Bandeira N, *et al*. The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. Molecular & Cellular Proteomics: MCP, 2010, **9**(12): 2840−2852

[23] Blanco L, Mead J A, Bessant C. Comparison of novel decoy database designs for optimizing protein identification searches using ABRF sPRG2006 standard MS/MS data sets. Journal of Proteome Research, 2009, **8**(4): 1782−1791

[24] Navarro P, Vazquez J. A refined method to calculate false discovery rates for peptide identification using decoy databases. Journal of Proteome Research, 2009, **8**(4): 1792−1796

[25] Cerqueira F R, Graber A, Schwikowski B, *et al*. MUDE: a new approach for optimizing sensitivity in the target-decoy search strategy for large-scale peptide/protein identification. Journal of Proteome Research, 2010, **9**(5): 2265−2277

[26] Kall L, Storey J D, Maccoss M J, *et al*. Posterior error probabilities and false discovery rates: two sides of the same coin. Journal of Proteome Research, 2008, **7**(1): 40−44

[27] Keller A, Nesvizhskii A I, Kolker E, *et al*. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Analytical Chemistry, 2002, **74**(20): 5383−5392

[28] Choi H, Nesvizhskii A I. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. Journal of proteome research, 2008, **7**(1): 254−265

[29] Ding Y, Choi H, Nesvizhskii A I. Adaptive discriminant function analysis and reranking of MS/MS database search results for improved peptide identification in shotgun proteomics. Journal of Proteome Research, 2008, **7**(11): 4878−4889

[30] Yadav A K, Kumar D, Dash D. Learning from decoys to improve the sensitivity and specificity of proteomics database search results. PloS One, 2012, **7**(11): e50651

[31] Piehowski P D, Petyuk V A, Sandoval J D, *et al*. STEPS: a grid search methodology for optimized peptide identification filtering of MS/MS database search results. Proteomics, 2013, **13**(5): 766−770

[32] Kall L, Canterbury J D, Weston J, *et al*. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nature Methods, 2007, **4**(11): 923−925

[33] Spivak M, Weston J, Bottou L, *et al*. Improvements to the percolator algorithm for Peptide identification from shotgun proteomics data sets. Journal of Proteome Research, 2009, **8**(7): 3737−3745

[34] Brosch M, Yu L, Hubbard T, *et al*. Accurate and sensitive peptide identification with Mascot Percolator. Journal of Proteome Research, 2009, **8**(6): 3176−3181

[35] Wright J C, Collins M O, Yu L, *et al*. Enhanced peptide identification by electron transfer dissociation using an improved Mascot Percolator. Molecular & Cellular Proteomics: MCP, 2012, **11**(8): 478−491

[36] Li N, Wu S, Zhang C, *et al*. PepDistiller: A quality control tool to improve the sensitivity and accuracy of peptide identifications in shotgun proteomics. Proteomics, 2012, **12**(11): 1720−1725

[37] Liang X, Xia Z, Jian L, *et al*. An adaptive classification model for peptide identification. BMC Genomics, 2015, **16**(Suppl 11): S1

[38] Xu M, Li Z, Li L. Combining percolator with X!Tandem for accurate and sensitive peptide identification. Journal of Proteome Research, 2013, **12**(6): 3026-3033

[39] Granholm V, Kim S, Navarro J C, *et al*. Fast and accurate database searches with MS-GF+Percolator. Journal of Proteome Research, 2014, **13**(2): 890-897

[40] Wen B, Li G, Wright J C, *et al*. The OMSSAPercolator: an automated tool to validate OMSSA results. Proteomics, 2014, **14**(9): 1011-1014

[41] Jian L, Niu X, Xia Z, *et al*. A novel algorithm for validating peptide identification from a shotgun proteomics search engine. Journal of Proteome Research, 2013, **12**(3): 1108-1119

[42] Liang X, Xia Z, Niu X, *et al*. Peptide identification based on fuzzy classification and clustering. Proteome Science, 2013, **11**(Suppl 1): S10

[43] Jian L, Xia Z, Niu X, *et al*. ℓ2 multiple kernel fuzzy SVM-based data fusion for improving peptide identification. IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM, 2015,

[44] Van Den Toorn H W, Munoz J, Mohammed S, *et al*. RockerBox: analysis and filtering of massive proteomics search results. Journal of Proteome Research, 2011, **10**(3): 1420-1424

[45] Wedge D C, Krishna R, Blackhurst P, *et al*. FDRAnalysis: a tool for the integrated analysis of tandem mass spectrometry identification results from multiple search engines. Journal of Proteome Research, 2011, **10**(4): 2088-2094

[46] Shao C, Sun W, Li F, *et al*. Oscore: a combined score to reduce false negative rates for peptide identification in tandem mass spectrometry analysis. Journal of Mass Spectrometry: JMS, 2009, **44**(1): 25-31

[47] Shteynberg D, Deutsch E W, Lam H, *et al*. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Molecular & Cellular Proteomics: MCP, 2011, **10**(12): M111. 007690

[48] Nesvizhskii A I, Keller A, Kolker E, *et al*. A statistical model for identifying proteins by tandem mass spectrometry. Analytical Chemistry, 2003, **75**(17): 4646-4658

[49] Reiter L, Claassen M, Schrimpf S P, *et al*. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. Molecular & Cellular Proteomics: MCP, 2009, **8**(11): 2405-2417

[50] Alves G, Yu Y K. Mass spectrometry-based protein identification with accurate statistical significance assignment. Bioinformatics, 2015, **31**(5): 699-706

[51] Weatherly D B, Atwood J A, 3rd, Minning T A, *et al*. A Heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results. Molecular & Cellular Proteomics: MCP, 2005, **4**(6): 762-772

[52] Serang O, Maccoss M J, Noble W S. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. Journal of Proteome Research, 2010, **9** (10): 5346-5357

[53] Spivak M, Weston J, Tomazela D, *et al*. Direct maximization of protein identifications from tandem mass spectra. Molecular & Cellular Proteomics : MCP, 2012, **11**(2): M111 012161

[54] Ma Z Q, Dasari S, Chambers M C, *et al*. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. Journal of Proteome Research, 2009, **8**(8): 3872-3881

[55] Sheng Q, Dai J, Wu Y, *et al*. BuildSummary: using a group-based approach to improve the sensitivity of peptide/protein identification in shotgun proteomics. Journal of Proteome Research, 2012, **11**(3): 1494-1502

[56] Huttlin E L, Hegeman A D, Harms A C, *et al*. Prediction of error associated with false-positive rate determination for peptide identification in large-scale proteomics experiments using a combined reverse and forward peptide sequence database strategy. Journal of Proteome Research, 2007, **6**(1): 392-398

[57] Chinese Human Liver Proteome Profiling C. First insight into the human liver proteome from PROTEOME(SKY)-LIVER(Hu) 1.0, a publicly available database. Journal of Proteome Research, 2010, **9**(1): 79-94

[58] Rudnick P A, Wang Y, Evans E, *et al*. Large scale analysis of MASCOT results using a Mass Accuracy-based THreshold (MATH) effectively improves data interpretation. Journal of Proteome Research, 2005, **4**(4): 1353-1360

[59] Zhang J, Li J, Xie H, *et al*. A new strategy to filter out false positive identifications of peptides in SEQUEST database search results. Proteomics, 2007, **7**(22): 4036-4044

[60] Keller A, Nesvizhskii A I, Kolker E, *et al*. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem, 2002, **74**(20): 5383-5392

[61] Choi H, Ghosh D, Nesvizhskii A I. Statistical validation of Peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. J Proteome Res, 2008, **7**(01): 286-292

[62] Yadav A K, Kumar D, Dash D. MassWiz: a novel scoring algorithm with target-decoy based analysis pipeline for tandem mass spectrometry. Journal of Proteome Research, 2011, **10** (5): 2154-2160

[63] Zhang J, Li J, Liu X, *et al*. A nonparametric model for quality control of database search results in shotgun proteomics. BMC Bioinformatics, 2008, **9**: 29

[64] Zhang J, Ma J, Dou L, *et al*. Bayesian nonparametric model for the validation of peptide identification in shotgun proteomics. Molecular & Cellular Proteomics: MCP, 2009, **8**(3): 547-557

[65] Ma J, Zhang J, Wu S, *et al*. Improving the sensitivity of MASCOT search results validation by combining new features with Bayesian nonparametric model. Proteomics, 2010, **10**(23): 4293-4300

[66] Anderson D C, Li W, Payan D G, *et al*. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. Journal of Proteome Research, 2003, **2** (2): 137-146

[67] Alves G, Wu W W, Wang G, *et al*. Enhancing peptide identification

confidence by combining search methods. Journal of Proteome Research, 2008, **7**(8): 3102−3113

[68] Searle B C, Turner M, Nesvizhskii A I. Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. Journal of Proteome Research, 2008, **7**(1): 245−253

[69] Sun W, Li F, Wang J, *et al*. AMASS: software for automatically validating the quality of MS/MS spectrum from SEQUEST results. Molecular & Cellular Proteomics: MCP, 2004, **3**(12): 1194−1199

[70] Zhang B, Chambers M C, Tabb D L. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. Journal of Proteome Research, 2007, **6**(9): 3549−3557

[71] Bittremieux W, Willems H, Kelchtermans P, *et al*. iMonDB: mass spectrometry quality control through instrument monitoring. Journal of Proteome Research, 2015, **14**(5): 2360−2366

[72] Jagtap P, Goslinga J, Kooren J A, *et al*. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. Proteomics, 2013, **13**(8): 1352−1357

[73] Helmy M, Sugiyama N, Tomita M, *et al*. Mass spectrum sequential subtraction speeds up searching large peptide MS/MS spectra datasets against large nucleotide databases for proteogenomics. Genes to Cells: Devoted to Molecular & Cellular Mechanisms, 2012, **17**(8): 633−644

[74] Bitton D A, Smith D L, Connolly Y, *et al*. An integrated mass-spectrometry pipeline identifies novel protein coding-regions in the human genome. PloS One, 2010, **5**(1): e8949

[75] Blakeley P, Overton I M, Hubbard S J. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. Journal of Proteome Research, 2012, **11**(11): 5221−5234

[76] Gonnelli G, Stock M, Verwaeren J, *et al*. A decoy-free approach to the identification of peptides. Journal of Proteome Research, 2015, **14**(4): 1792−1798.

[77] Tu C, Sheng Q, Li J, *et al*. Optimization of search engines and postprocessing approaches to maximize peptide and protein identification for high-resolution mass data. Journal of Proteome Research, 2015, **14**(11): 4662−4673

[78] Granholm V, Navarro J F, Noble W S, *et al*. Determining the calibration of confidence estimation procedures for unique peptides in shotgun proteomics. Journal of Proteomics, 2013, **80**: 123−131

The Application and Progress of Target-decoy Database Search Strategy in Identification and Quality Control of Tandem Mass Spectrometry Data in Shotgun Proteomics*

FENG Xiao-Dong[1,2], MA Jie[2], CHANG Cheng[2], BAI Ming-Ze[1], ZHU Yun-Ping[2]**, SHU Kun-Xian[1]**

([1]) *Institute of Bioinformatics, Chongqing University of Posts and Telecommunications, Chongqing* 400065, China;

[2]) *Institute of Radiation Medicine, Engineering Research Center for Protein Drugs, Beijing Proteome Research Center, State Key Laboratory of Proteomics, National Center for Protein Sciences, Beijing* 102206, China)

Abstract With the advance of mass spectrometry and experimental techniques, a huge amount of mass spectrometry (MS) data has been accumulated rapidly in the past few years. Usually, different database search engines are used to analyze the MS data for peptide and protein identification. Meanwhile, selection of all those assignments that are actually correct is one of the most daunting tasks in mass spectrometry based proteomics investigations. Target-decoy searching strategy has become one of the most popular strategies to control the fasle identification in MS/MS data analysis. In this paper, we first introduce the workflow of target-decoy database searching strategy. Then we summarize the quality control tools based on the target-decoy searching strategy. Besides, we point out the deficiency of target-decoy searching strategy and put forward mending measures. Finally, we carry on summary and outlook to target-decoy searching strategy.

Key words tandem mass spectrum, protein identification, target-decoy, quality control

DOI: 10.16476/j.pibb.2016.0067

* This work was supported by grants from Projects of International Cooperation and Exchanges(2014DFB30010), National Basic Research Program of China (2013CB910800), The National Natural Science Foundation of China (21475150, 21105121, 61501071) and Chongqing Postgraduate Scientific Research and Innovation Projects(CYS14154).

**Corresponding author.

ZHU Yun-Ping Tel: 86-10-61777058, E-mail:zhuyunping@gmail.com

SHU Kun-Xian Tel: 86-23-62468319, E-mail: shukx@cqupt.edu.cn

Received: March 1, 2016 Accepted: May 6, 2016