人类蛋白质组表达谱蛋白质鉴定的分步搜索策略

吴松锋,朱云平,贺福初

(军事医学科学院放射医学研究所,北京 100850)

摘 要:大规模蛋白质组表达谱研究的蛋白质鉴定一般采取基于数据库搜索的策略,因此数据库的选择及搜索策略在蛋白质鉴定中非常重要。现有的人类蛋白质数据库远不够完善,而从其他物种的蛋白质数据库中所能得到的补充非常有限,但人类基因组数据库中却可能存在很大的补充空间。在对国际人类蛋白质数据库充分调研、比较的基础上,提出了一种分步搜索的策略。这种策略首先利用一个质量较高、覆盖率相对较大的非冗余数据库进行基本鉴定,随后利用其他蛋白和核酸数据库进行补充鉴定和新蛋白挖掘。该策略能有效地鉴定尽可能多的高可靠蛋白,并能进一步充分利用质谱数据进行补充鉴定和新蛋白挖掘,对大规模蛋白质组表达谱研究具有重要的意义。

关键词:蛋白质组;蛋白质鉴定;蛋白质数据库;质谱

中图分类号:Q39 **文献标识码**:A **文章编号**:0253－9772(2005)05－0687－07

Strategy for the Protein Identification of Human Proteome Expression Profile: Selection of Searching Database

WU Song-Feng, ZHU Yun-Ping, HE Fu-Chu

(*Beijing Institute of Radiation Medicine*, 27 *Taiping Road*, *Beijing* 100850, *China*)

Abstract: Widely used method of protein identification for high-throughout proteome expression profile studies was database-dependent, so the selection of databases for the protein identification was very important. Despite the deficiency of available human protein databases, the complementarity of human proteins could be got mainly from human genome but not from the protein databases of other organisms. According to the comparison of the current protein databases from different aspects, IPI was recommended for the basic identification for the studies of human proteome expression profile, and other human protein or nucleic acid databases were needed for the complementary identification and novel protein mining.

Key words: proteome; protein identification; protein database; mass spectrum

收稿日期:2004－09－10;**修回日期**:2005－03－17

基金项目:国家 863 计划(编号:2002BA711A11,2004BA711A21);国家 973 计划(编号:2001CB510209);北京市科技计划项目(编号:H030230280590);国家自然科学基金资助(创新研究群体科学基金)(编号:321003)[The Chinese National Key Program of Basic Research (No. 2001CB510209), the National High Technology Research and Development Program of China (No. 2002BA711A11, 2004BA711A21), Beijing Municipal Science and Technology Project (No. H030230280590), and National Natural Science Foundation of China (No. 321003).]

作者简介:吴松锋(1977—),福建人,在读博士生,研究方向:基因组学与蛋白质组学

通讯作者:贺福初,Tel:010-66931246;E-mail: hefc@nic.bmi.ac.cn

朱云平,Tel:010-66932248;E-mail:zhuyp@hupo.org.cn

致 谢:感谢本实验室蛋白质组研究人员应万涛、姜颖、郭立海和王晶兰博士等对本工作提供的各种帮助、支持以及批评;感谢生物信息学组荔建琦博士对这些分析所提出的宝贵的批评意见;此外,特别感谢 EBI 的 Paul Kersey 博士对去除 IPI 数据库构建方法及其相关问题所提供的耐心、快速的解答

蛋白质组表达谱的蛋白质鉴定方法主要有两种:数据库搜索和从头(*De novo*)测序。从头测序相当于在所有可能的序列空间中寻找最可能的肽段或蛋白,这种方法费时费力,难以实现高通量,而且对质谱结果要求很高。然而自然界中实际存在的蛋白质氨基酸序列的组合方式相对少多了,如果只在少量的蛋白质序列中选择,即使图谱质量不高,也可以很明确地鉴定某个蛋白质。因此在大规模的蛋白质组表达谱研究中一般采用数据库搜索的策略。质谱数据的数据库搜索结果强烈依赖于数据库中所收录的蛋白质序列,即如果数据库中不存在相应的序列,即使图谱再好,也无法鉴定出相应的蛋白质;而如果数据库过于庞大,则蛋白鉴定的概率会下降,可能会增加假阳性鉴定结果。因此数据库的选择尤为重要。现有主要的质谱数据解读软件有 Mascot、Sequest 等,本文主要基于用 Mascot 软件搜索 Q-TOF 数据的结果进行讨论。

为了尽可能得到较好的鉴定结果,Mascot 软件的数据库搜索策略采取了概率打分和容错性搜索的方法[1]。

Mascot 用概率打分的方法搜索匹配质谱数据的数据库记录。一个理想的 MS/MS 谱图包含一个或多个完全的碎片离子系列:没有噪音峰,质量数准确。如果有这样完美的数据,是不需要基于概率打分算法的。但是,实际的质谱数据不是理想化的,也不可能得到非常完美的匹配,因此 Mascot 软件采用一种概率打分的方法判断匹配的结果。

同时,Mascot 采取容错性搜索的策略。在进行质谱数据的数据库搜索时,经常会有大量的谱峰找不到匹配。假定一个给定的质谱结果中含有足够的信息(比如,有足够的信噪比很好的碎片离子峰),这种数据找不到匹配肽段的可能原因有:低估质量测量误差;电荷估计错误;酶的非特异性切割;未预料的化学或翻译后修饰以及数据库中不存在对应的肽序列等。前两个问题可以由其他方法解决,后 3 个问题则可以通过 Mascot 的容错性搜索解决。Mascot 容错性搜索是通过查询一个化学和翻译后修饰的列表以及一个碱基替换矩阵来解决在搜索一个序列数据库时碰到的修饰、变异等问题[1]。

虽然 Mascot 软件利用概率打分和容错性搜索等策略可以辅助质谱数据的蛋白鉴定,但对于蛋白鉴定而言,数据库的选择仍然是一个无法绕过的重要问题,数据库的数据质量和蛋白覆盖率直接决定了蛋白质鉴定的效果。

本文在对国际蛋白质数据库调研的基础上,结合各数据库的数据来源及构建方法,比较了多种数据库的冗余度、覆盖率以及实际搜索效果,提出了一种分步搜索的策略,并应用于本实验室的人胎肝蛋白质组研究中。

1 材料和方法

NCBI 的 Entrez 人类蛋白质数据库从 NCBI(http://www.ncbi.nlm.nih.gov)的 TaxBrowser 中下载,这些蛋白质基本上包括了 nr 数据库中的所有人类蛋白质,因此本文提到的 NCBI nr 指从 Entrez 提取的人类蛋白质数据。文中所用的其他数据库下载地址详见表 1。此外,人和鼠直系同源(ortholog)蛋白比较所用的全基因组预测蛋白质数据从 Ensembl 下载。

表 1 蛋白质数据库的下载地址

Table 1 Website of protein databases

Databases	URL
NCBI's Entrez proteins	http://www.ncbi.nlm.nih.gov
Swissprot	ftp://ftp.ebi.ac.uk/pub/databases/uniprot/
RefSeq	http://www.ncbi.nlm.nih.gov/RefSeq/
Trembl	ftp://ftp.ebi.ac.uk/pub/databases/uniprot/
GenPept	ftp://ftp.ncifcrf.gov/pub/genpept/
Ensembl	http://www.ensembl.org/
Hinv-DB ORF	http://www.jbirc.aist.go.jp/hinv/index.jsp
UniRef100	ftp://ftp.ebi.ac.uk/pub/databases/uniprot/
IPI(International protein index)	ftp://ftp.ebi.ac.uk/pub/databases/IPI/current/

蛋白质的相似性比较用本地 BLAST 软件(从 NCBI 下载)。为了获得直系同源基因,人和鼠数据库的蛋白质间相互最佳匹配的确定直接用相互间相似性搜索(BLAST)结果的第一项,如果相互最佳匹配均为对方,则为一个相互最佳匹配对。数据库内蛋白质的去冗余方法是直接将完全一致的序列去除,只保留其中一个。Mascot 搜索参数的设置根据仪器和实验结果确定,而后取分值高于 Mascot 提供的阈值的结果作为鉴定结果。

2 结 果

2.1 数据库的物种选择

数据库物种选择争论根源在于人类蛋白质数据

库尚不完整,而且进化关系近的生物其对应的蛋白质序列一般较相似,因此,在鉴定人类蛋白质时可能通过搜索其他生物的蛋白质数据库实现对人类蛋白质数据库的补充。本节就此问题作简要的论证。

2.1.1　跨物种蛋白质鉴定的可行性

对于进化关系上较近的物种,其对应的蛋白质有些非常相似,甚至完全一致。对人和鼠的基因组编码蛋白质(数据来自于 Ensembl 网站,见材料与方法)的分析,其相互最佳匹配(可以粗略认为这些相互最佳匹配的蛋白质对是直系同源蛋白质)的相似性分布如图 1。

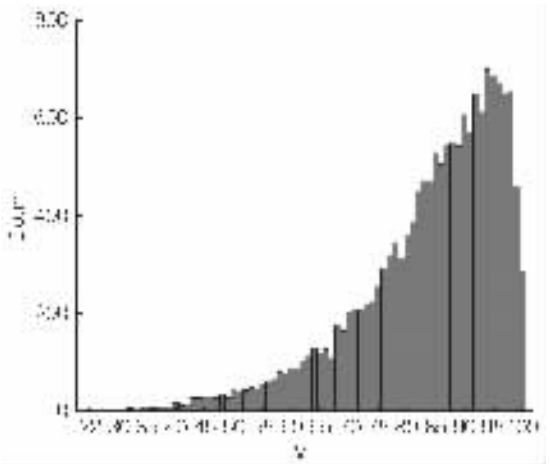

图 1　人-鼠基因组编码的蛋白质相互最佳匹配的相似性分布图

注:横坐标为序列间 Identities(BLAST 结果),纵坐标为蛋白质数目。

Fig.1　The distribution of protein similarity of reciprocity match between the coding proteins of human and mouse

Note: The x-axis indicates the identities (result of BLAST), and the y-axis indicates the number of proteins.

分析表明,人和鼠预测蛋白质序列中共有 16 112个蛋白质是互相最佳匹配的。其中相似性大于等于 95% 的有 3 410 个,占相互最佳匹配对的 21.2%。而鼠的蛋白质中相似性大于等于 95% 的有 5 308 个,占鼠总蛋白质(32 281)的 16.4%。因此,某些鼠的蛋白质可以作为人类蛋白质组研究中搜索其对应蛋白质的数据库记录,即可以用于人类蛋白质的鉴定。而 Mascot 软件也带有容错性(Error Tolerant Search)搜索的功能,可以允许相似物种的蛋白质,以及同一个物种由于测序错误、多态性、突变导致的差异。因而,根据蛋白质数据库的本质和 Mascot 软件设计原理,用其他生物(亲缘关系较近)的蛋白质数据库作为鉴定数据库在某种程度上是完全可行的。

此外,本实验室的实践也证明了能用人类样品的质谱结果搜索非人类的相应蛋白质数据;而已有的报道也证明,用模式生物的数据库鉴定非模式生物的蛋白质在某种程度上是完全可行的[2]。

2.1.2　人类蛋白质数据库的构建方法和不完整性

人类基因组测序完成后,对人类基因组的基因注释用了 3 种基本的方法[3]:(a) 有 EST 或 mRNA 支持的转录物。(b) 依赖于已鉴定的基因或蛋白质序列的相似性搜索。(c) 用隐马尔可夫模型(HMMs)进行的外显子从头预测(如 Genscan、Genie 和 FGENES 等)。

但至今,数据库收录的蛋白质仍不完全。在人类基因组测序完成后,对人类基因组的分析表明:RefSeq 数据库中被认为是"已知"的基因,只有 92% 至少部分匹配上基因组;85% 多于一半匹配;16% 匹配上多处(可能是旁系同源或假基因)[3]。因此基因组数据中存在很大的不完全性。此外,基因预测方法也是不完善的,现有的方法只能准确预测出人类 70% 的外显子,而只有 20% 的基因所有外显子都能被准确预测[3]。

2.1.3　非人类蛋白质数据对人类蛋白质数据的补充非常有限

由于基因组测序完成后对基因组编码蛋白质的注释方法其中之一是同源性比较,而能用于质谱数据跨物种鉴定的蛋白质一般需要较高的相似性。因此,可以认为,只要人类基因组相应的核酸序列已经测序完毕,而且基因组的注释能力足够强并且更新足够快,则在非人类蛋白质数据库中注释的蛋白质很快就会在人类蛋白质数据库也会有相应的注释蛋白质。所以只有非人类基因组的注释蛋白质正好对应于人类基因组的空白部分,才有可能成为人类蛋白质的补充。所以以下只考虑无法匹配上人类基因组的非人类蛋白质数据。

由上述人-鼠间序列比较可知,鼠的总蛋白质约有 16% 和人对应蛋白质相似性大于 95%,因此这些蛋白质有可能用于人类蛋白质的鉴定。另根据已有的文献报道,鼠只有少于 1% 的基因(118 个)在人类基因组中没有找到同源预测基因(但和其他物种有明显的相似性,这些基因被认为很可能在人类中缺

失了[4],但笔者认为也有可能是人类基因组测序不完全造成的)。如果这 118 个蛋白质是人类基因组中含有的,只是由于基因组测序不完全的原因未被发现,用 16%的比例计算,得到鼠中有 118×16%=19 个蛋白质(占鼠总蛋白质数的 0.06%)可能用于填补人类蛋白质数据库的空白。虽然计算过程中用 95%相似性比较苛刻,但这些数据已足以说明其他物种数据库的记录可用于补充人类蛋白质鉴定的空间极其有限。而如果大量引入非人类蛋白质数据进行数据库搜索,反而会大大增加假阳性匹配的概率。因此,在人类蛋白质组研究中,质谱数据的蛋白质鉴定用人类蛋白质的数据库已经足够了。

综上所述,虽然跨物种蛋白质鉴定被证明是可行的,而且人类蛋白质数据库也不够完整,但其他物种的蛋白质数据对人类蛋白质数据库的补充作用非常小。故而,在大规模蛋白质组研究中,为了不过多引入其他假阳性匹配结果,在人类蛋白质组研究蛋白质鉴定时,只搜索人类蛋白质数据库。

2.2 各种人类蛋白质数据库的关系和比较

由于数据来源、构建方式和目的的不同,当前国际上存在着多种人类蛋白质序列数据库。最明显的两个极端:一个是 Swiss-Prot 数据库,以极高的质量和几乎完全非冗余而闻名,但数据量较小,覆盖面也较小;另一个是 NCBI nr 数据库,该数据库将国际上最常见的数据库直接加和,一般被认为是非常全面的数据库,覆盖率较大,但总体质量不高,是高度冗余的数据库(只去除了完全一致的序列)。在准确性和覆盖率的权衡中,到底该选择什么样的数据库,曾一度在本实验室引起很大的争议。

本节对当前人类蛋白质数据库进行了充分的调研和比较,并用一批数据作了实际搜索测试,提出了分步搜索的设想。

2.2.1 现有人类蛋白质数据库的关系

由图 2 可知,现有的人类蛋白质数据主要有以下几个来源:a. 从文献或其他比较可靠的数据来源收集到的蛋白质数据,一般经过手工校对,包括 Swiss-Prot 和 RefSeq 的 NP 部分等。b. 直接由 mRNA 数据翻译得到:主要是 TrEMBL 和 GenPept,以及 H-Inv DB ORF。c. 基因组预测得到,包括 Ensembl,UCSC 提供的预测蛋白质数据。d. 其他可靠性较低的数据库,包括 TrEST 等。

而在这些数据库的基础上,出现了不少复合数据库,包括 NCBI 的 nr、EBI 的 Uniprot 和 IPI 等数据库,这些数据库与前面描述的数据库间关系见图 2。

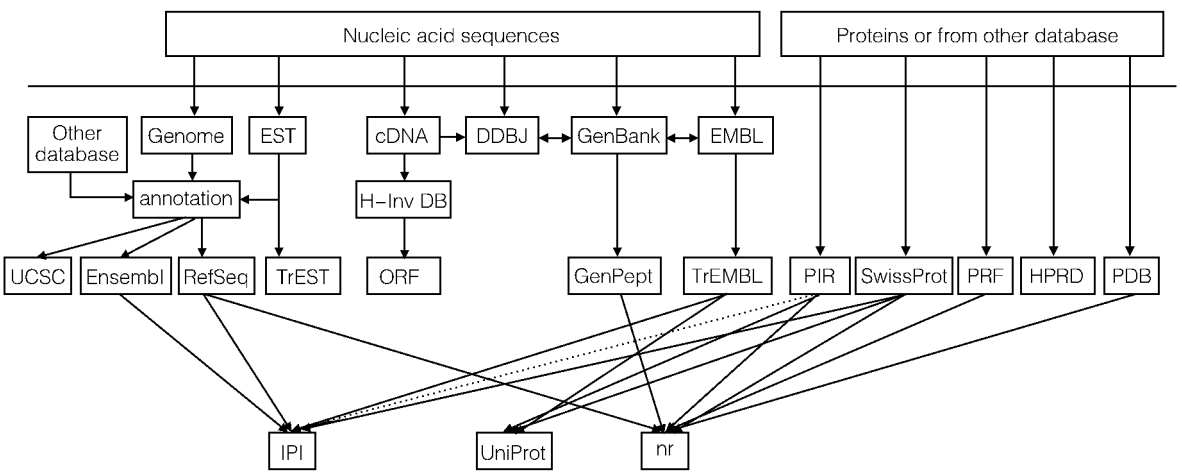

图 2 人类蛋白质数据库的组成及其关系图

Fig. 2 Composition and relationship of human protein databases

2.2.2 人类蛋白质数据库间冗余度和覆盖率比较

上述各数据库由于构建方式、目的不同,对序列的处理等也相应有所差异。因此,不同数据库的冗余度和覆盖率是有差别的,而这些差别直接影响到数据库搜索的结果,因此在数据库选择时必须考虑这些因素。

表 2 列出了不同数据库的冗余度比较。由表中结果可见,IPI、UniRef100 和 Swiss-Prot 几乎是完全非冗余的,TrEMBL 的冗余度也较小,而 GenPept 和 nr 数据库则是高度冗余的。

表 2　不同数据库冗余度比较

Table 2　Redundance comparison of different databases

Database*	Download date	Total seq	Redundance seq	Non-redundance seq
Entrez proteins	2004-04-26	214 254	85950(40.1%)	128 304
GenPept	2004-05-22	148 019	39429(26.7%)	108 590
Hinv-DB ORF	2004-04-29	39 091	3177(8.1%)	35 914
RefSeq	2004-05-03	27 724	1277(4.6%)	26 447
Ensembl	2004-02-03	29 802	1294(4.3%)	28 508
Trembl	2004-05-10	36 187	601(1.7%)	35 586
Swissprot	2004-05-10	10 880	1(0.009%)	10 879
UniRef100	2004-05-04	43 503	0	43 503
IPI	2004-05-04	41 519	0	41 519

* : All of the databases used are human protein databases.

表 3 列出了不同数据库间的覆盖率比较。因为没有一个总的蛋白质数据库,因此直接用数据库间相互比较度量覆盖率的差别〔其中双 95% 是模拟 IPI 构建时 95% 的整合标准制定的一种衡量序列间相似性的指标,指 BLAST 结果中匹配上的区域占查询序列总长的 95% 以上,匹配区域的一致性(Identities)在 95% 以上;双 50% 的含义类推。表 3 中所列的百分数是指在双 95% 或双 50% 以下的蛋白质数所占该数据库总蛋白质数的百分数[1)]。结果表明,nr 数据库的覆盖率几乎是最大的,但所用于比较的几个数据库都存在着或多或少的蛋白质是 nr 库中不存在的。而相对较小、非冗余的 IPI 和 UniRef100 数据库中,IPI 数据库的覆盖率更大。

NCBI 的 nr 数据库有大量的序列是其他数据库中不存在的,将其与 UniRef100 比较发现其相似性和匹配率在双 50% 以下的大概有 90% 是免疫分子,这些可能是 TrEMBL 中不收录的免疫分子的多态等位基因等。

综上所述,现有的人类蛋白质数据库一般而言数据量越大,覆盖率也越大,但冗余度也随之增大。在众多数据库中,至今没有一个数据库能覆盖所有已经报道的人类蛋白质。相对而言,IPI 和 UniRef100 是非冗余的,覆盖率也较大。这些比较研究对随后的搜索策略的制定有非常大的参考价值。

表 3　不同数据库蛋白质覆盖率比较

Table 3　Coverage comparison of different databases

Database	Double 95(%)	Double 50(%)
Entrez_protein→UniRef100	39.2	12.4
Entrez_ protein →IPI	39.0	—
UniRef100→ Entrez_ protein	1.3	0.85
IPI→ Entrez_ protein	8.0	—
Hinv DB→ Entrez_ protein	64.7	26.1
IPI→UniRef100	26.0	—
UniRef100→IPI	7.2	—

2.2.3　人类蛋白质数据库的实际搜索效果比较

为了检测不同数据库的实际搜索效果,用 Q-TOF 产生的 53 个串联质谱结果的 Peak list(pkl)文件对不同数据库(nr,nr+Hinv DB,UniRef100,IPI)进行搜索测试,并且对不同数据库的实际搜索结果(取鉴定的分值大于等于 30 的肽段)进行比较。

用于测试的 53 个 pkl 文件在 nr 数据库中共鉴定了 393 条分值大于等于 30 的肽段,总体而言,凡在 UniRef100 中搜索到的肽段,在 nr 和 nr+Hinv 复合数据库中都能被检索到。但在 nr 和 nr+Hinv 中搜索到的肽段有 4 条在 UniRef100 中未搜索到。而用 IPI 数据库搜索的结果除了与 UniRef100 一样比由 nr 鉴定的肽段少 4 个外,还检索到一条额外肽段(nr 数据库中没有,而 Ensembl 收录的预测蛋白)。

由以上结果可知,用复合数据库搜索的确能鉴定到某些小数据库中没有的蛋白质,但和 UniRef100 或 IPI 比较所增加的贡献只有 1% 不到,但数据量增大了 2 倍以上,而且可能存在不少可靠性不高的序列。因此,诸如 IPI 或 UniRef100 等小数据库在基本鉴定中已经大体够用了,但如果需要较全面挖掘质谱的结果,则可能需要额外的数据库。

2.3　蛋白质鉴定的分步搜索策略

根据上述分析论证,我们提出了一种蛋白质鉴定的分步策略:在质谱数据的蛋白质鉴定中,先用一

1) IPI 对非 Swiss-Prot 或 RefSeq 的 NP 部分用了 95% 的标准聚类。此处设定双 95% 是模拟 IPI 数据库的构建方式设定的。此外,为了获得更低的相似性蛋白质的关系,根据我们的经验,大体设定了 50% 的标准。

个相对较小、质量相对较高、覆盖率较全的非冗余数据库作为基本鉴定数据库,以便得到质量较高的鉴定结果;而由于现有的这类数据库还不够完善,因此需要进一步用其他数据库进行补充鉴定(有些已知蛋白质也可能在所选择的较小的数据库中没有收录)以及新蛋白质挖掘(指新的预测蛋白质,或核酸序列,可能质量不高)。

2.3.1 基本鉴定

大规模蛋白质组表达谱研究不可能对每个鉴定结果均追溯其结果的可靠性(包括质谱结果和数据库记录的可靠性),因此如果直接选择高质量的蛋白质数据库,能有效地提高鉴定质量。此外,冗余的数据除了增加搜索时间外,并不能对搜索结果有任何额外的贡献(同一个蛋白质的不同变异结果在蛋白质鉴定时是没有多大意义的)。因此选择一个非冗余的数据库是必要的。然而如果数据库中不存在相应的蛋白质或蛋白质的某种特定剪接形式(或酶解形式),可能会导致该蛋白质无法被鉴定出来。所以这就要求用于搜索的数据库尽量选择高质量、非冗余、且存在不同剪接形式及相似蛋白质(但非同一个基因型的蛋白质)的数据库。

由上述比较得知,NCBI 的 nr 蛋白质数据库冗余很大,而且质量参差不齐;Swiss-Prot 数据库质量很高,但覆盖率不高,而且没有将可变剪接分开(虽然它提供了可以将其分开的工具);IPI 和 UniRef100 均为非冗余的,但与 nr 蛋白质数据库相比覆盖率也不是特别高(IPI 比 UniRef100 覆盖率稍高)。考虑到 nr 数据库中存在而 IPI 或 UniRef100 没有的序列中有 90%可能是免疫多态分子(这些分子对鉴定的贡献不大),因此 IPI 和 UniRef100 可以说已经覆盖了人类大部分质量较高的蛋白质(相对而言,IPI 数据库的覆盖率较 UniRef100 略大)。而 IPI 按数据质量等级不同进行构建的方式[5]也基本上说明了其质量相对较高,因此我们建议选择 IPI 作为基本蛋白质鉴定的数据库。

2.3.2 补充鉴定和新蛋白质挖掘

用于基本蛋白质鉴定的高质量、非冗余的数据库势必不可能覆盖现有的所有人类蛋白质。随着人类基因组测序完毕,非人类蛋白质数据库对人类蛋白质数据库的补充已经是非常有限,而人类基因组中蕴含的潜在蛋白质还可能存在很大的空间,因此我们建议只用人类的数据做相应的补充。现有的人类蛋白质数据除了 nr 数据库中收录的以外,还有不少人类基因组预测的蛋白质序列,以及其他数据库收录的部分蛋白质(图 2);此外,人类的 EST 序列可能含有部分潜在的蛋白质编码区,部分编码区可能没有被翻译成蛋白质,因此蛋白质数据库中并没有相应的记录,这些序列可能用于新蛋白质的鉴定;鉴于基因组中还有多达 20%的外显子没有被准确预测[3],因此除了上述两个部分外,人类基因组也可能是一个补充。

搜索基本数据库的鉴定结果的判读可以直接依据一般的鉴定结果判别方法。而对于其他补充鉴定和新蛋白控据的结果,则需要更高的判据以及进一步的分析验证。

此外,基于对现有的人类蛋白质数据库的分析(图 2),复合蛋白质数据库可以通过合并大量的人类蛋白质数据库(包括大量的预测蛋白质)、并对这些序列进行去冗余实现。在本实验室,人胎肝蛋白质组研究中,所合并的数据库包括 2004 年《核酸研究》杂志列出的基本蛋白质数据库[6](包括 NCBI nr, TrEMBL, GenPept, Ensembl, DoTS, Trome-TrEST-TRGEN, UCSC predicted proteins 等)和 Hinv DB 的 ORF[7]。而对于人类的 EST 和基因组数据,可以直接从 NCBI 及 Ensembl 中下载得到。

综上所述,对于大规模人类蛋白质组表达谱的蛋白质鉴定而言,理想的蛋白质鉴定策略及步骤是:首先搜索一个高质量、非冗余、高覆盖率、收集了不同蛋白质剪接形式的数据库;其次搜索尽量包括所有人类蛋白质的复合数据库;第三,搜索人类 EST 数据库;第四,搜索人类基因组数据库;如果可能的话,最后可用程序直接读取质谱数据(相当于 *De novo* 测序),以鉴定由于蛋白质数据库未收录的高质量质谱结果,以尽可能地挖掘质谱数据。

2.4 蛋白质鉴定策略在胎肝蛋白质组研究中的应用

在本实验室的人胎肝蛋白质组研究中,我们采用了以上介绍的蛋白质鉴定策略。其中,本实验室产生的 Q-TOF 数据搜索基本数据库(IPI)鉴定了 3 000多次蛋白质,搜索复合蛋白质数据库得到 272 个额外的肽段,人类 EST 得到 151 个额外肽段,人类基因组得到 21 个额外肽段(未去冗余)。这些肽段(不计用 IPI 数据库鉴定的结果)对应蛋白质中补充鉴定的蛋白质(IPI 中未收录的已知蛋白质)有 28 个,已知蛋白质的不同形式(包括变异,酶切和可变

剪接等)有 131 个,蛋白质序列已知功能未知的蛋白质有 12 个,蛋白质序列未知的序列有 15 个,此外还有 3 个可能是不在真实的编码区相位中的假阳性结果(结果尚未发表)。这结果通过对总体可靠性较高的 IPI 数据库进行搜索,获得了大部分的鉴定结果,随后对其他数据库进行进一步搜索,获得了额外的补充鉴定结果和候选新蛋白鉴定结果。这些结果证明了本文所介绍的蛋白质鉴定的分步搜索策略是行之有效的。

3　讨　论

由上述比较可知,现有的人类蛋白质数据还不够完善,国际上至今还没有一个供蛋白质组研究中蛋白质鉴定用的理想数据库。现有的蛋白质组研究一般只是挑选一个相对较可靠且覆盖率较全的数据库作为蛋白质鉴定的搜索数据库。也有部分研究开始考虑用其他数据库做补充鉴定[8]以及新蛋白质挖掘(HPPP 采用的方法)。但已有的工作都没有对数据库进行系统地比较,也未能提出一个较妥当、系统的蛋白质鉴定数据库选择的策略。

本研究对现有的蛋白质数据库进行了调研和比较,结果表明,人类蛋白质数据库很不完善,但其他物种的蛋白质数据对人类蛋白质数据的补充非常有限,而人类基因组中还存在着对当前人类蛋白质数据库的补充空间。

基于这些比较研究,我们提出了蛋白质鉴定的分步搜索策略。这种策略首先用一个质量相对较高,覆盖率较大的数据库进行基本的蛋白质鉴定,以便获得大部分较可靠的鉴定结果;随后搜索其他蛋白质数据库,并用较高的阈值,以进行补充鉴定或新蛋白质挖掘。这样做的优点一方面可以实现大部分搜索结果的可靠性,另外一方面可以实现较完全的蛋白质鉴定。

对现有数据库的比较研究的结果表明,IPI 数据库具有数据质量较高、覆盖率较大等诸多优点,比较符合基本蛋白质鉴定的需求。而由于 IPI 自身的不完全性,为了尽可能利用质谱数据,需要进行额外的补充鉴定及新蛋白质的挖掘。而对于搜索其他数据库,由于数据量增大,因此需要较高的标准确认此结果,同时对于预测蛋白质或没有蛋白质序列报道的核酸序列,还需要进行进一步的确认。

蛋白质鉴定的分步搜索策略适用于大规模的蛋白质组表达谱研究。如果只是为了在某个实验中鉴定个别蛋白质,可以参考这种策略并根据具体情况进行相应的调整。此外,对 PMF 类型的数据,所获得的是蛋白质酶切肽段谱,而非串联质谱数据的肽段测序,故而一般只进行基本的蛋白质鉴定,如果也采取补充鉴定和新蛋白质挖掘的策略,其结果需要特别谨慎(因为新蛋白质可能只有部分肽段是可靠的,但很难实现全长都是可靠的,而 PMF 是基于全长蛋白质或大部分序列鉴定的)。

对复合数据库搜索结果分析时发现不少已知蛋白质的变异形式,包括氨基酸变异、非特异性切割、可变剪接等。由于 IPI 中没有收录这些序列,无法鉴定这些结果,因此这种策略也能被用于鉴定已知蛋白质的其他形式。这些蛋白质的其他形式中有些是信号肽被切除后的蛋白质(属于有规律的切割产物),这些问题可能在随后的工作中从数据库的构建或搜索软件的设计上解决。

参 考 文 献(References):

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

[2] Liska A J, Shevchenko A. Expanding the organismal scope of proteomics: cross-species protein identification by mass spectrometry and its implications. *Proteomics*, 2003, 3(1): 19～28.

[3] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 2001, 409(6822): 860～921.

[4] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 2002, 420(6915): 520～562.

[5] Kersey P J, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R. The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, 2004, 4(7): 1985～1988.

[6] Galperin M Y. The Molecular Biology Database Collection: 2004 update. *Nucleic Acids Res*, 2004, 32: Database issue: D3～22.

[7] Imanishi T *et al*. Integrative annotation of 21 037 human genes validated by full-length cDNA clones. *Plos Biol*, 2004, 2: 856～875.

[8] Kristiansen T Z, Bunkenborg J, Gronborg M, Molina H, Thuluvath P J, Argani P, Goggins M G, Maitra A, Pandey A. A proteomic analysis of human bile. *Mol Cell Proteomics*, 2004, 3(7): 715～728.