基于生物质谱数据的互联网数据库检索*

魏滢 岳贵花 宋浩威 杨芃原**
(复旦大学化学系 上海 200433)

〔摘要〕利用互联网上生物数据库是进行质谱数据的分析及处理的一项很重要的途径。本文简要介绍了网上蛋白数据库知识以及它在生物质谱分析中的一些应用。

关键词: 生物质谱 数据库检索

前言

随着全球范围内互联网技术的迅猛发展, 使得信息的传播和利用得到质的飞跃。利用互联网进行生物信息的快速、全面搜索已经成为生命科学工作者必不可少的一项工作。生物学界应运而生出新的名词生物信息学(Bioinformatics), 用于巨量生物信息资源的收集、存储、处理、搜索、共享、研究和开发。可以说, 互联网技术的发展创造了一个新的生物时代。我国生物质谱的研究日趋活跃, 拥有生物质谱的院校、企业单位也越来越多。在工作中质谱工作者均会遇到根据质谱数据通过互联网到公共生物质谱数据库检索的问题, 为了帮助国内学者掌握这一技术, 本文对这方面的知识作了一些简要的介绍。

1 数据库介绍

1.1 主要的蛋白质数据库

SWISS-PROT
(http://www.expasy.ch.sprot)

蛋白质序列库, 隶属于Amos Bairoch, 日内瓦大学(University of Geneva), 是获得蛋白质序列信息的首选数据库。这个数据库的优点是非冗余(nonredundant)每个蛋白只有一个搜索的进入条目(entry); 同时有着强大的链连接功能, 能快速从其它数据库调得相关信息。内容丰富, 提示详尽。缺点是由于其高准确性, 并不是所有的蛋白序列都能从SWISS-PROT中查得。

PIR (Protein Identification Resource)
(http://www.gdb.org/Dan/proteins/Pir.html)

建立较早, 内容全面。包括蛋白和核酸数据库, 鉴定和分析蛋白的软件, 提供在线的各种信息服务。力求最大限度兼顾冗余和非冗余数据库。特别是: 1) PIR1 是非冗余(non-

* 2000-01-09 收
** 国家自然科学基金资助项目(H29927002)和国家教育部博士点基金资助项目
*** 通讯联系人

redundant)的数据库。2) PIR1+ PIR2+ PIR3 是冗余(redundant)的数据库。有多个重复的搜索进入条目,较为全面。

PDB (the Protien Data Bank)

(http://www.pdb.bnl.gov/cgi-bin/browse)

从属于美国Brookhaven 国家实验室开发,是目前最为全面的蛋白质结构数据库。查寻PDB 可以获知未知蛋白与哪些已知3D 结构的蛋白相近。PDB 也是非冗余(non-redundant)的数据库。

OWL

(http://www.gdb.org/Dan/protins/owl.html)

也是一个蛋白非冗余数据库。

NRL-3D

(http://www.gdb.org/Dan/proteins/nr13d.html)

是蛋白的序列-结构数据库。

1.2 进入数据库的途径

进入数据库有各种各样的途径,大致可以归类为以下三个途径:

(1) E-mail 服务器

使用E-mail 服务器十分简单。只要收发通过E-mail 就可以获得所需的蛋白序列或其他结果。

(2) Internet 服务器

现在有越来越多的服务器允许通过internet 获取蛋白序列或其他生物信息。比较典型的有Entrez(http://www,ncbi.nlm.gov/Entrez) (DNA/RNA + 蛋白+ 结构+ Medline 分支)从属于NCBI,Entrez 综合了蛋白序列和核苷酸序列数据库的信息以及Medline(美国规模最大,权威最高的文献数据库)的文献信息。它的主要的优势是将各种各样的数据库相互交联起来,用户可以很便捷地通过一个序列链接到另外一个相关的序列。Entrez 的一个核心概念是毗邻(neighboring)即将序列和相关信息按照计算的相似值大小来分类。对于每个序列都给出了类似的序列或相关的结构、功能和文献信息。

(3) 本地数据库

用户也可以建立自己的数据库,这对于经常使用该数据库的用户很有好处。大多数的数据库可以从匿名FTP 服务器上获得,许多也可以通过CD ROM 获取。缺点是信息不能及时更新。

1.3 检索算法

BLAST (Basic Local Alignment Search Tool)

(http://www.ncbi.nlm.nlh.gov/BLAST)

局部比对基本检索工具,是最为常用的数据库中搜寻生物序列的同源性检索软件。它是由NCBI 开发维护。其主要特别如下:

(1) 力求局部的最大相似性而不是整个序列的最大相似性。

(2) 依据可靠的理论基础。

(3) 快速的检索速度。

(4) 包括四个特色组成: BLASTP, BLASTN, TBLASTN, BLASTX。其中 BLASTP 用于在蛋白质序列库中检索蛋白序列; BLASTN 用于在核酸序列库中检索核酸序列。

FASTA

是另一种常用的蛋白及 DNA 同源性检索工具。可以用 FASTA 与 BLATA 对同一系列进行检索, 以增加找到同源性的可能性。

1.4　综合性的强大的生物信息网站

nr(NCBI)

(http://www.ncbi.nlm.nih.gov)

由美国生物技术研究中心(NCBI)组建, 作为其 BLAST 搜寻的目标数据库。它包括了 SwissProt, SwissProt updates, PIR, PDB 等数据库。

EMBL (European Molecular Biology Laboratory)

(http://www.embl-heidelberg.de)

欧洲生物信息搜索服务中心规模大、信息量广, 它的数据库包括核酸, 蛋白质以及大分子结构等, 是生物工作者首选浏览的站点之一。其中 Mann 蛋白质与肽段研究组(http://www.mann.embl-heldelberg.de)主要从事蛋白质(包括质谱分析)以及基因组学研究。在这个主页上有蛋白鉴定分析的工具软件、各种质谱信息以及广泛的链连接功能, 还可以链连到相关的生物信息网站、研究所和生物仪器公司。

SRS(Sequence Retrieval System)

(http://www.ebi.ac.uk/srs)

将 EMBL, SwissProt, PIR, PDB, Prosite 等大规模的站点链接为一个有机的整体。对于种类敏多的分子生物学数据库, SRS 是功能强大的检索系统, 它能检索 45 个数据库, 而且这些数据库通过超链接整合在一起, 通过这些超链接能方便地在不同数据库间跳跃测览。

2　基于生物质谱数据的检索

2.1　质谱数据和数据库中序列的相关分析原理:

数据库中蛋白质序列的快速增长大大促进了多肽串级质谱谱图的分析, 通过计算机算法, 多肽的 CID 谱图的信息可以用来做蛋白质和核酸数据库搜索。Yates 等发展了一种相关数据学的方法, 用质谱谱图中的碎片信息来做蛋白质和核酸数据库的搜索从而识别出谱图所代表的氨基酸序列, 进一步鉴别出蛋白质。利用谱图中的碎片和多肽分子量信息, 对于大多数计算机算法来说, 两个或两个以上的具有较好的信噪比的谱图匹配到同一蛋白质序列的可能性很小, 因此可以用一个质谱谱图来鉴别蛋白质。蛋白质质谱谱图的专一性和特殊性使我们可以鉴别混合物中的单个蛋白质, 包括蛋白质污染物。如, 角蛋白、抗体和蛋白酶自解产物等。

多肽的质谱谱图包含了三个水平的信息。最基本的是多肽的质量, 如结合用来产生多肽的酶的特殊性, 只通过多肽的质量信息就可以把可能的序列信息减少到很小的数量(10～1000), 但如果酶的特殊性不知道, 并且质量误差较大时, 在一个大的数据库中单靠质量信息会得到很多的可能的多肽, 如 Yates 等研究了一个质量为 1480.7u(±3u)的多肽, 在一个核酸数据库中搜索到了 6.5×10^6 个可能的氨基酸序列。为了鉴别一个测得质量的特定的氨基

酸序列, 必须利用第二层次的信息- 串级质谱谱图所提供的碎片模式和序列离子的信息, 这些信息对于一个给定的序列来说是专一性的。通过把符合给定质量的每个序列的碎片离子和实际的串级质谱图相关联, 可得到每个序列的符合程序。同一研究小组设计了一个方法, 利用从数据库中得到的氨基酸序列, 构造理论的串级质谱谱图, 再和实际的谱图作互相关分析, 可以得到理论谱图和实验所得谱图的相似性, 并通过互相关得分的大小来识别哪一个序列是最可能的。串级质谱谱图中包含的第三层次的信息是多肽的实际序列, 通过实际谱图中的一小段并不足以识别特定多肽的序列, 结合整个多肽的质量信息和碎片离子的质量信息, 就可以得到这一多肽的较专属的特性, 在数据库搜索中就可以得到较少的可能序列。对于不同的质量分析器的串级质谱谱图都有相应的数据库搜索方法, 但利用串级质谱谱图进行数据库搜索存在着一个潜在的缺点: 相似或相关的蛋白质存在着相同的序列, 因此只用一个谱图不能唯一地鉴别, 必须有另外的串级质谱谱图来区分这些相同的蛋白质。

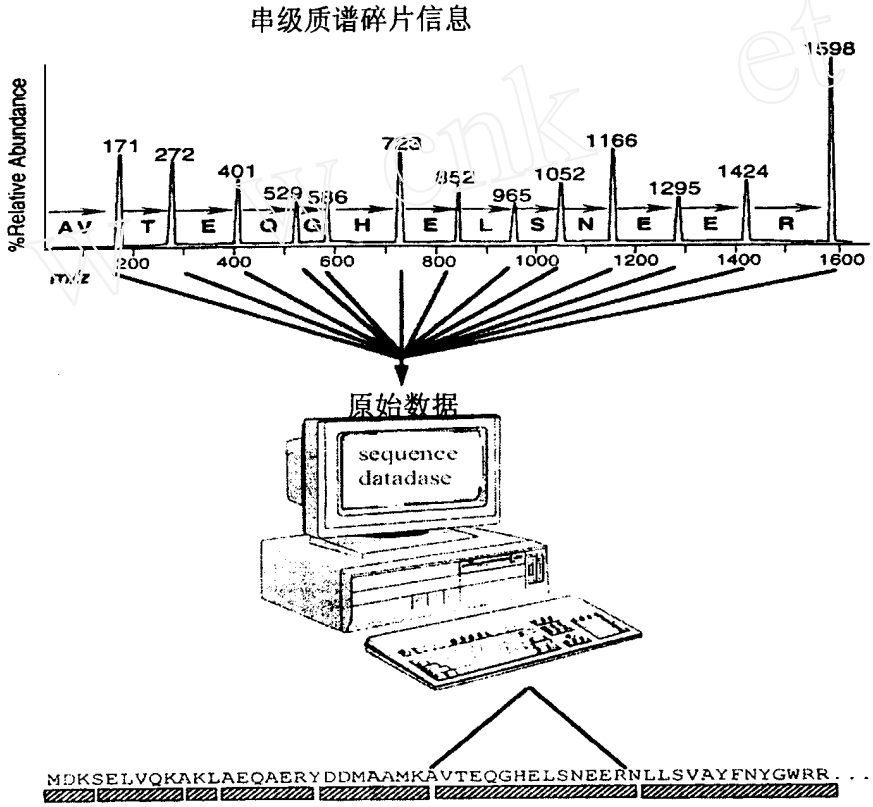


图 1 蛋白序列库中检索示意图

2.2　数据库在蛋白质谱分析中的应用

蛋白序列分析的准确度的影响因素主要有如下两个: (1) 检索的数据库, 一般选择非冗余(non-redundant)的数据库。 (2) 搜索的算法一般选择 BLA STA , FA STA。

质谱分析中主要通过三个途径获得蛋白质信息: (1) 肽段质量检索。 (2) 肽段序列标记(peptide sequence tag) 检索。(3) 通过氨基酸序列检索。通过检索软件, 可以将蛋白质谱数据(肽段质量或碎片离子质量等) 与数据库中的同源性蛋白关联起来, 获得相应的蛋白质各种信息, 如图 1 所示。

下面简要介绍几个比较常用的蛋白质谱数据分析软件:

(1) 利用肽段质量(质谱信息) 进行检索

PeptideSearch (EMBL) 比较快速, 且较全面, 将数据输入, 如图所示:

EMBL Protein & Peptide Group

PeptideSearch/WWW

Protein identification by peptide mass data

Fill out the form and press Start PeptideSearch to perform a database search.

Protein mass range [kDa]: 0 < Mr < 300

Cleavage agent: Trypsin

Number of peptides required for protein match: 5

Number of missed cleavage sites per peptide: 1

Peptide mass accuracy: 0.1 Da using Monoisotopic mass

Peptide charge state: Protonated (MH+)

☐ **Oxidized Methionine**

Cysteine is: Carbamidomethyl-Cys

Peptide masses:

745.4311
752.4110
797.3966
839.1514
929.8979
982.4935
1091.6398
1116.7359
1199.7238
1232.6318
1260.6841

On each result page show this number of matches: 30

Start PeptideSearch

Reset form

[Peptide sequence tag search] [Amino acid sequence search]

EMBL Protein & Peptide Group

PeptideSearch results

Date: 1999-12-27
Time: 14:02:29+01:00
Database: nrdb.index(20Oct99)
202.120.224.4
197 matches

Search parameters

Protein mass range	0-300 kDa
Cleavage agent	Trypsin
Peptide mass accuracy	0.1 Da
Methionine is	Native
Cysteine is	Carbamidomethyl-Cys
Peptide charge state	Protonated
Number of peptides required for match	5
Number of uncleavaged sites	1
Number of peptides used in search	23

Peptide masses used in the search

Masses are monoisotopic

745.4311	752.411	797.3966	839.3514	929.8979	982.4935	1091.6398	1116.7359
1199.7238	1232.6318	1256.6841	1278.8252	1307.6915	1320.7267	1345.6921	1513.7327
1627.9873	1633.8895	1716.2595	1829.0047	1887.8467	1951.9813	2023.3943	

Search result

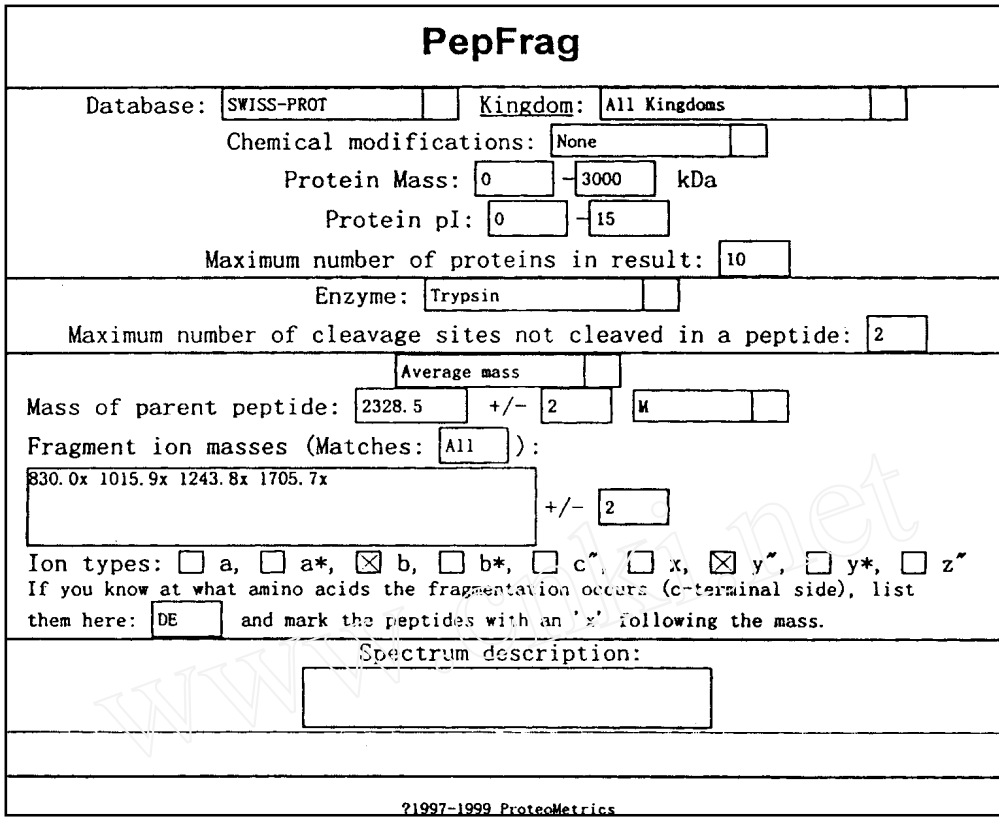
197 matches were found. Showing matches 1 through 30.

Peptides matched sort	Mass [kDa] sort	Database accession sort	Protein Name sort	Digest	2nd pass search
10	74.47	swissprot:P16474	GR78_YEAST 78 KD GLUCOSE-REGULA		
8	183.33	sptrembl:O93058	D78608|TM180KP_1 product: "180K		
8	183.40	trembl:AF165190	AF165190_1 product: "183kDa pro		

其他相似的软件有MS-FIT 由加州大学旧金山医学院(UCSF)开发。将质谱数据(母体质量等)与序列数据库中的最合适的已知蛋白相关,获得同源性信息。这个软件的特点是信息量大,更为详尽准确。另外还有 Profound (PROWL),MOWSE (Daresburg Lab U)也各有特色。

(2) 利用蛋白质谱序列标记进行蛋白序列分析检索

ProFrag(PROWL)软件便是进行这项任务。将数据输入,如图:



查询结果如下:

PepFrag Search Results

```
Database: sprot
Maximum number of proteins in result: 10
Protein mass: 0.0-3000.0 kDa
Protein pI: 0.0-15.0
Enzyme: Trypsin, # of incompletes: 2

Mass of parent peptide: 2328.5 +/- 2.0
Fragments: 830.0x, 1015.9x, 1243.8x, 1705.7x, Error: 2.0, Matches: 4
Ion types: b, y"
Fragments marked with 'x' cleaved at C-terminal side of D or E

Searching: sprot
. .

HS77_YEAST MITOCHONDRIAL HEAT SHOCK PROTEIN SSC1 PRECURSOR (ENDONUCLEASE
SCE1 75 KD SUBUNIT) - SACCHAROMYCES CEREVISIAE (BAKER'S YEAST) mass = 70627.
9 Da, pI = 5.5

   PPAPKGVTQEVTFDIDADGLLNVSARDKATN mass = 2329.6

      830.00 +/- 2.00 Da: y"8 (829.98 Da)
      1015.90 +/- 2.00 Da: y"10 (1016.14 Da)
      1243.80 +/- 2.00 Da: y"12 (1244.39 Da)
      1705.70 +/- 2.00 Da: y"16 (1706.89 Da)

. . . .
Search time = 7 s
```

相似软件有 PeptideSearch (EMBL), MS-TAG 等, 也是由 UCSF 开发, 内容更详些。

(3) 通过氨基酸序列查询同源性蛋白

如 PROWL 的 Proteininfo 软件, EMBL, ProteinProstector 上也有相应的服务。

3 互联网上的生物质谱资源

3.1 讨论组

网上有很多以某一研究领域为主题的讨论组, 在讨论组上不仅可以查询到大量信息, 还可以与相关人士进行交流, 公布实验中的经验和创新技术等。规模比较大的生物学家电子论坛有 BIONET (http://www.bio.net) 由美国科学基金会创立于 1991 年, 它根据不同专题分为几十个消息组。用户以电子邮件的形式参与论坛; BIOMOO (http://bioinformatics.weizmann.ac.il/BioMOO) 是生物学家网上聚会的场所, 即所谓的虚拟会场。此外, BioNews Bioinformatics Forum (http://bioinformatics.weizmann.ac.il/bionews.html) 以及生物质谱方面的讨论组 sci.techniques.mass.spec 等。

3.2 主要搜索引擎

使用搜索引擎是在最短时间内查找所需信息的最好的方法。主要的搜索擎有主题搜索 Yahoo (http://www.yahoo.com), 关键词搜索引擎 AltaVista (http://www.altavista.com), Pedro's 生物分子搜索工具 (http://www.public.iastate.edu/~pedro/research-tools.html) 是较为全面的生物方面搜索引擎, 它包括三个部分, 即分子生物检索与分析; 生物文献及 WWW 搜索; 帮助工具。哈佛大学的生物信息搜索引擎 Biology Links (http://golgi.havard.edu) 是 Interne 上生物学的总汇, 将生物数据库、分子生物学数据库、软件等几类收录全面。

3.3 一些与生物质谱有关的站点介绍

PROWL (http://www.prowl.rockefeller.edu)

前面提到的几个重要的蛋白质谱分析软件如 ProFound, 都是由美国 Rockefeller 大学开发的。它着重于蛋白质的质谱分析。其特点是简明、针对性强, 同时也具有很强的链连接功能, 能够灵活的进行蛋白质质谱数据分析。PROWL 还有各种服务于蛋白分析的信息和资料。

CBRG (http://cbrg.inf.eth2.ch/subsection3-1-3.html)

BasePeak (http://base-peak.wiley.com)

侧重于质谱方面的信息, 对于生物质谱工作者也是值得经常浏览的站点。

ProteinProspector (http://prospector.ucsf.edu)

从属于 UCSF 大学, 也是一个很好的站点, 信息丰富。

Weizmann Institute (德国基因组和生物信息学研究所)

(http://bioinformatics.weizmann.ac.il/index.html)

经典的站点, 其 FAQ 是初学者很好的入门知识。

UCSF Ms-Fit http://rafael.ucsf.edu/Ms-Fit.html

PeptideMass Search (e-mail 方式)

http://www.mdc-berlin.de/~emu/peptide-mass.html

MOWSE (e-mail 方式) mowse@dl.ac.uk

关于非冗余(non-redundant)的解释

是比较数据库的一个重要概念,现在许多数据库都力求非冗余。所谓非冗余,即指对于每个蛋白只有一个简单确切的定义条目。生物数据的复杂性使得很难给非冗余(non-redundant)以明确的定义。因此造成每个非冗余的数据库有各自的定义。有些数据库使用自动检测,而有些数据库用手工的筛选方法达到非冗余的目的。也有数据库为使数据库更为全面,而不追求非冗余。

与有关生物大分子相关的重要站点总汇

Basepeak	Http://base-peak.wiley.com
GenBank	http://www.ncbi.nlm.nih.gov
EMBL	http://www.embl-heidelberg.de
GDB	http://www.gdb.rog
PIR	gttp://www.gdb.org/Dan/proteins/pir.html
SW-PROT	http://www.expasy.ch/sprot/sprot-top.html
NRL3D	http://www.gdb.org/Danproteins/nr13d.html
Enzyme	http://www.expasy.ch/sprot/prosite.html
SW-2D	http://www.expasy.ch/ch2d-top.html
BLAST	http://www.ncbi.nlm.nih.gov/blast
BLITZ	http://www.ebi.ac.uk/bic_sw
ExPASx	http://www.expasy.ch
PROWL	http://prowl.rockefeller.edu
MOTIF	http://www.genome.ad.jp/SIT/MOTIF.html
BLOCKS	http://www.blocks.fhcrc.org/blocks_release.html
ENTREZ	http://www.ncbi.nlm.gov/Entrez
PROSITE	http://www.expasy.ch/sprot/prosite.html
LCCG	http://arep.med.harvard.edu
TREMBL	http://www.embl-heidelberg.de/~seqanal

The Bio-Mass Spectrometry Database Search on the Web

Wei Ying, Yue Guihua, Song Haowei, Yang Pengyuan
(Dept of Chemistry, Fudan University, Shanghai 200433, China)

Received 2000-01-09

Abstract

Using biology related database on the Web is very important for the analysis of mass spectrometric data. In this paper the protein database and its application to mass spectrometry are discussed.

Key Words: Bio-Mass Spectrometry, database search, Internet