基于信息技术的蛋白质识别研究[1]

陈益强[1],高文[1,2],付岩[1],李德泉[1],陈翔[1]
1 (中国科学院 计算技术研究所，北京 100080)
2 (哈尔滨工业大学 计算机科学与工程系，哈尔滨 150001)

摘要：随着质谱技术以及 NMR 等仪器测定技术的飞速发展，利用大规模质谱技术可以得到海量蛋白质的质谱数据，同时利用 NMR 技术也可获得蛋白质的精确三维结构数据。这些数据以及面向这些数据的分析方法使得对蛋白质组的研究发展十分迅速，不论是基础理论、技术方法，都在不断地进步和完善。利用信息工程技术辅助生物学家开展面向质谱技术的高通量蛋白识别，面向 NMR 的蛋白质三维结构识别以及功能发现研究，已经成为蛋白质组中的生命科学与信息科学交叉研究的热点和挑战问题。本文侧重从信息技术的角度对上述问题进行综述，并提出我们解决问题的思路和方法。

关键词：蛋白质，质谱，三维结构，信息技术
中图分类号：TP391 **文献识别码**：A

Research on Protein Recognition base on Information Technology

Yiqiang Chen[1], Wen Gao[1,2]， Yan Fu[1], Dequan Li[1], Xiang Chen[1]
1 (Institute of computing technology, Chinese academy of science, Beijing, 100080)
2 (Department of computer science, Harbin industry university, Harbin, 150001)

Abstract: With the development of mass spectrum (MS) and NMR technology, there are huge data sets obtained from protein MS data as well as protein 3D structure data. These drive the fast development on research of proteome, not only in fundamental research, but also in new applied technology. Utilizing information technology and assisted by biologists to research on MS based high through protein identification and NMR based protein 3D structure recognition and it's function discovery, it becomes hot topics and challenge in crossed research of biology and information field. This paper firstly surveys the research situation on above topics from information technology point of view, and then give the approach we are working on to solve the issues.

Keywords: Protein, Mass Spectrum, 3D structure, Information technology

1. 引言

中国是一个人口大国，提高人民生存健康水平和人口质量，是国民经济可持续发展和社会进步，繁荣昌盛的迫切需求。因此一个首要目标就是要减轻、消除重大疾病如肝癌、心脑血管疾病、肿瘤、糖尿病和老年病等对中国人民健康的影响。对中国人重大疾病和稀有但有科学研究价值疾病的相关

[1]作者简介：陈益强（1973- ），男，博士，数据挖掘，智能人机交互，生物信息学。高文（1956- ），男，教授，博士生导师，生物特征识别，智能人机交互，音视频编码，虚拟现实等。付岩（1977– ），男，博士生，生物信息学。李德泉（1976- ），男，博士生，生物信息学。陈翔（1975- ），男，博士士，生物信息学。

蛋白质的研究具有重要的理论价值和应用意义。由于对致病细胞实验产生的蛋白质是海量的，如何在已知蛋白质库中利用少量的实验信息高通量辨别未知蛋白一直是致病相关蛋白研究的核心问题。这个问题的解决应用到医学临床方面，就能重点实验出控制致病相关蛋白的药物，为人类控制疾病以及治疗疾病打下基础。同时在这个问题的基础还可以实现基于蛋白质识别的新基因发现系统。当然，创新药物的发现仅仅从 DNA 水平上是不可能完成的，必须基于蛋白质三维结构进行，对致病相关蛋白研究的另一个挑战问题就是蛋白质结构识别和功能发现。蛋白质的三维结构对于其功能而言特别重要，结构相似的两个蛋白质可能就有相同的功能。因此，通过对已知蛋白质结构的分类整理和功能发现来识别未知蛋白质的功能成了重要的研究主题。同时从大量已知蛋白质结构中发现相同子结构（蛋白质的捆绑点）也是生物中的一个重要问题，这主要由于这些子结构可能有某种功能并与其他蛋白或 DNA 进行作用，试图发现这些子结构可以让我们发现已有蛋白的新功能。

随着质谱技术以及 NMR 等仪器测定技术的飞速发展，利用大规模质谱技术可以得到海量蛋白质的质谱数据，同时利用 NMR 技术也可获得蛋白质的精确三维结构。这些数据以及面向这些数据的分析方法使得对蛋白质组的研究发展十分迅速，不论是基础理论、技术方法，都在不断地进步和完善。国际上很多大药厂和公司在巨大财力的支持下也纷纷加入蛋白质组的研究阵容。目前面向质谱与 NMR 的蛋白质识别技术研究成为蛋白质组中的生命科学与信息科学交叉研究的热点和挑战问题，我们在蛋白质组研究的背景下，开展针对疾病相关蛋白的识别及结构功能的研究，目的是利用信息工程技术，辅助生物学家开展面向质谱技术的高通量蛋白识别以及面向 NMR 的蛋白质三维结构识别技术平台建设以及功能发现研究等，最终为疾病治疗以及改善中国人口结构和健康状况打下坚实基础。如果说高通量的蛋白质识别方法是从相对较宏观的角度研究致病相关蛋白，那么从三维结构去识别蛋白质以及发现其功能就是从相对微观的角度研究致病相关蛋白。本文侧重从信息技术的角度对这两方面问题进行研究综述并提出我们解决问题的思路和方法。

2. 面向质谱（MS）的高通量蛋白质识别

2.1 问题与方法

最初，蛋白质序列测定主要采用手工的埃德曼降解-环甲基化方法，效率很低。质谱（MS）技术的发展为蛋白质序列测定开辟了新的途径。使用质谱技术识别蛋白质是基于这样的事实：我们所能测量的只是分子的质量，而想要的是分子的氨基酸序列。质谱分析的基本过程为：利用二维电泳（2-D gel）过程使蛋白质混合物（比如一个细胞）在分子量和等电点两个方向上分散开来，从中取其中一个点进行质谱实验。利用选定的酶对蛋白质样品进行水解，形成多肽。不同质量的多肽被质谱仪检测出来，得到质谱(MS)。同时肽可以进一步被打碎，并测得碎片质量分布，即串联质谱（MS/MS）。用质谱识别蛋白质的方法有三类：1 基于 MS 的蛋白质识别；即通过搜索已知蛋白质数据库，用指定的酶对蛋白质进行模拟水解，得到理论 MS。理论 MS 与实验 MS 进行比较，结果按照匹配的程度排序。这样的系统有 MOWSE[1]，Mascot[2]，ProFound[3]，PeptIdent[4]，MS-Fit[5]等等。用于匹配的打分算法是这种方法的核心。MOWSE 和 Mascot 对不同质量的肽的出现频率做了统计，设计了基于概率的打分算法。ProFound 则建立了贝叶斯打分算法，形式化地考虑了更多的先验知识。基于 MS 的蛋白质识别适用于蛋白质样品包含一种蛋白质或简单的混合物的情况。其缺点是由于蛋白质混合物和污染物、部分酶解、残基修饰、质量精度等因素的影响，往往误差较大，导致错误搜索结果。2 是基于 MS/MS 的蛋白质识别；通过搜索已知蛋白质数据库，对蛋白质模拟酶解，再对肽模拟碰撞，得到理论串联质谱。理论串联质谱与串联实验质谱进行比较，符合的做为结果。这类方法可以用于识别复杂的蛋白质混合物或者验证肽质量指纹搜索的结果，是目前最常用和有效

的方法，本文将详细介绍。3 是称作 de Novo 测序的方法，它直接解释 MS/MS 数据进行识别肽序列，而不是与数据库中的序列进行比较。这类系统和算法有 Lutefisk[6][7]，SHERENGA[8]，PEAKS[9]，AuDeNS[10-13]]等等。当数据库中没有目标序列，样品蛋白质出现未知修饰，以及出现未知的或非特异性的裂解生成的肽时，搜索数据库的方法就无能为力了，所以不得不使用 de Novo 的方法。但是该方法的难点在于对 MS/MS 数据的预处理以及如何重新标定质谱，要求质谱数据有较高质量，肽断裂情况良好，所以目前尚未得到广泛的实际应用。不过 de Novo 方法即使不能完全测出肽序列，也能提供重要的肽序列标记（几个氨基酸组成的短肽），供数据库搜索参考。

2.2 基于 MS/MS 的高通量蛋白质识别

用MS/MS搜索据库识别蛋白质的基本流程如图1所示：

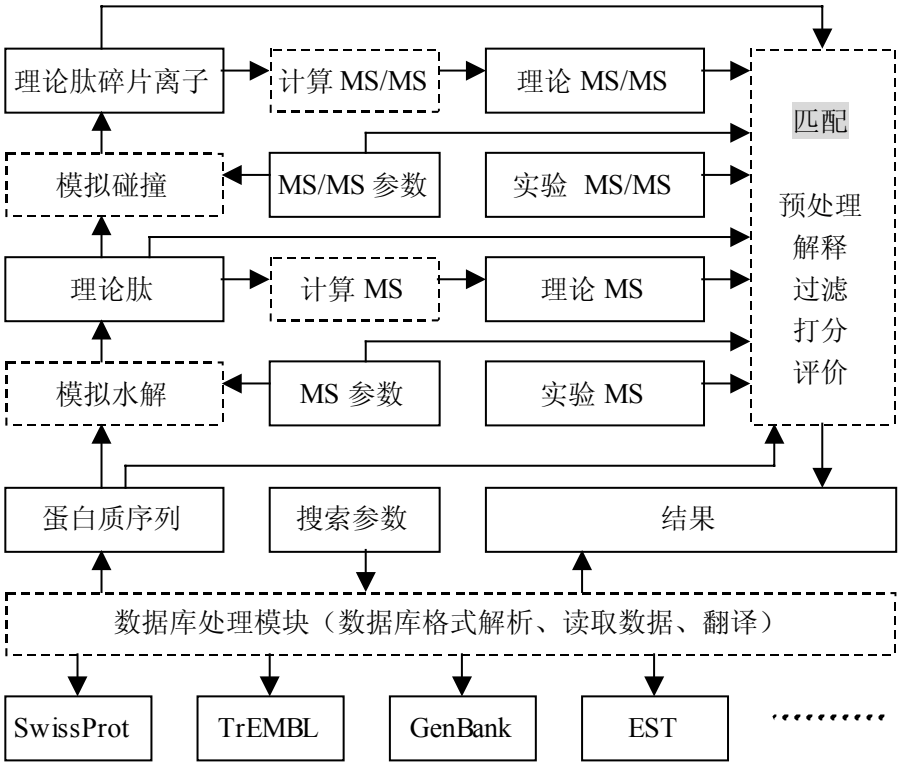

图1： 基于MS/MS的蛋白质识别框架(实线框表示数据，虚线框表示计算模块)

数据库处理模块根据数据库、物种组织等搜索参数从数据库中读出蛋白质序列（如果是核酸序列，需要翻译成蛋白质序列)。根据酶、遗漏切点等MS参数模拟水解这个蛋白质，生成所有可能的理论肽。从这些理论肽可以计算出理论MS。根据质谱仪等MS/MS参数和肽的物化性质对每个理论肽进行模拟碎裂过程，生成理论肽碎片离子。从这些理论碎片离子可以计算出每个理论肽序列对应的理论MS/MS。以上大部分是对生物过程的模拟，匹配则是核心计算模块。在把理论MS/MS和实验MS/MS比较打分之前，需要对实验MS/MS进行预处理和解释。预处理是为了去除实验质谱中的噪音，提取有效数据。对MS/MS的解释可以得到序列标签等信息。用这些解释信息和用户输入的一些参数可以直接过滤掉不满足要求的蛋白质序列或肽序列。最后，将理论MS/MS与实验MS/MS进行比较，给肽和蛋白质打分。这时也可能会用到MS、序列标签等信息。得到的结果按分值排列，高通量的系统还应对分值的有效性进行评价。显示给用户的结果中可能还要包括更多的信息，需要从数据库中读取。这是所有用MS/MS搜索数据库识别蛋白质的系统总体的流程，每个系统侧重点和算法各不相同。如果完全依赖于解释就成了De novo的方法。所有搜索数据库的系统都强调打分模块，这也是本

节介绍的重点。

2.2.1 数据预处理

不管是 De novo 测序，还是数据库搜索方法，都需要对输入的实验 MS/MS 数据进行预处理，以缩减噪音，提取有效数据。对于质峰的选取，需要确定选取多少数目的质峰和哪些质峰。简单的方法是选取固定数目的丰度值最高的质峰，比如 SEQUEST[14]。而 Mascot[2]则动态的确定要选取的值峰的数目，它考虑了匹配时肽的性质。更复杂的去噪音的办法还有利用最大熵原理，比如 Micromass[15]等。但仅凭丰度值选取质峰是不够的，因为有用的质峰可能被噪音埋没。所以合理的做法是加入先验知识。Mascot 通过在每一个固定范围内(14Da)保留一个质峰的方法加强了对小丰度值质峰的选取，以获得与离子系列相对应的质峰阶梯。也有人考虑了峰之间的相关性来调整质峰的丰度，以选择最优的一组峰，如 AuDeNS[10]。我们认为，对于质峰的选取这一特殊的生物问题，充分利用关于质谱的先验知识并将其融入传统的信息处理方法中应是解决问题的关键。

2.2.2 解释与过滤

虽然完全依赖于对质谱的解释的方法是 De vovo 测序，但是搜索数据库的各个系统也或多或少的加入了解释模块，其目的是直接过滤掉不符合的序列以加快搜索速度和减少错误匹配。通常的方式是利用蛋白质或肽的质量信息进行过滤，比如 SEQUEST 和 Mascot 等。PepFrag[16]利用了氨基酸组成信息(氨基酸组成与氨基酸序列是不同的概念，氨基酸组成只是说明氨基酸的出现，没有位置信息)。有的还加入蛋白质的等电点值，比如 MS-Tag[17]等。更复杂的是从质谱中解释出部分氨基酸序列信息，称作序列标签，比如 PepSea[18]。准确的序列标签对于消除假阳性非常有效。值得注意的是，过多的和不正确的解释也可能误导搜索的过程。多大程度的解释才算合适，以得到最好的搜索效果是有待探讨的问题。无论如何，有效的解释模块对于提高系统的性能是非常重要的，这也是未来搜索数据库系统的发展趋势。

2.2.3 评分算法

打分算法是搜索数据库算法的核心。普通的打分算法包括或者简单地对匹配的质峰计数，如 MS-Tag，或者累加匹配的质峰的丰度值，或者把丰度值取对数后再累加。取对数的目的是弱化丰度的贡献。

SEQUEST 除了累加匹配的峰的丰度值外，还考虑了特定信息对分值的加权。他打分的方法如下，

$$s_p = \left(\sum i_m\right) n_i (1+\beta)(1+\rho) / n_t$$

其中，n_i 是在正负 1-u 误差内与质谱中观察到的离子相匹配的理论碎片离子的数目，i_m 是它们的丰度。若发生连续的离子系列匹配，则加入分数增长项 β（0.075）。对于质谱中的 immonium 离子（与 His，Met，Trp，Tyr，或者 Phe 等氨基酸相联系），视序列中是否出现这些氨基酸，加减分数项 ρ（0.15）。n_t 是理论碎片离子的总数。SEQUEST 还用了交叉-相关分析重新对用上述方法得出的前 500 个结果重新打分，以加强对小肽的打分。最后的分数 C_n 如下计算：

$$C_n = R_\tau(\tau = 0) - \frac{1}{149} \sum_{-75<k<75} R_{\tau=k}, \quad R_\tau = \sum_{i=0}^{n-1} x[i]y[i+\tau]$$

其中，x_i 是理论质谱，y_i 是实验质谱。最后，分数被归一化到 1.0。

SCOPE[19]是 Celera 公司设计的打分算法。它利用贝叶斯模型进行打分，对于给定的质谱求每个序列的后验概率作为分值，即

$$p^* = \arg\max \psi(S|p), \quad \psi(S|p) = \sum_F \psi(S|F,p)\Pr(F|p)$$

其中 $\psi(S|p)$ 是给定肽 p 生成质谱 S 的概率密度函数，F 是碎裂模式。

Mascot 是基于概率的打分算法，提出了 p-值的概念，即，计算匹配为随机的概率，而分数为 -log(p)。Sonar MS/MS[20]把质谱转化为向量，用向量的点积作为分值。

这些打分算法复杂程度各异，也各有优缺点。交叉相关分析来源于信号处理理论，用于比较两个信号是否一致，但未必适合于质谱这一特殊信号。如果我们简单的把质谱当作普通信号进行处理，不加入先验知识，那么必将丢失很多信息。Mascot 的 p-值应该说是对匹配很好的一个测量，文献中没有给出计算方法。但这样的 p-值必定是不可计算的，所以只能是个估计值，仍要做很多近似和利用经验信息。SCOPE 打分算法的理论模型完备，但为了能够计算做了很多假设和近似。我们认为，融入更多的专家经验，充分利用对质谱的解释信息，考虑更多的因素比好看的理论模型更重要。打分方法仍是个尚未完美解决的问题。

2.2.4 有效性评价

高通量的蛋白质识别应尽量排除人的干预。所以，在打分之后应对分值的有效性进行评价，也就是要确定匹配是完全随机的，还是真实匹配。SEQUEST 在进行交叉相关分析之后，经验地发现如果排在第一位和第二位的结果的分值相差在 0.1 以上，那么排在第一位的结果通常是真实匹配。这只是个特殊情况。

一般的做法有三种。一是使用模拟的方法，首次在 MS 搜索中提出[21]。在模拟中，随机的构造理论质谱，再用理论质谱搜索数据，最高的分数被保存。反复进行这一过程，就可以得到一个随机识别的分数分布。从这分布就可以计算实际识别的蛋白质或肽为随机的概率。另一种方法是直接计算随机匹配的概率。比如 Mascot 认为 p-值在 0.05 以下的匹配才是真实的。然而，这些直接计算的方法不如模拟的方法可靠，因为计算的过程非常复杂，不得不做近似处理。第三种方法是在每次搜索中统计分数的分布。比如，Sonar MS/MS 提出了期望值的概念，期望值在 1 以下的匹配才被认为是有效的。一个统一的评价标准对于比较各个系统的性能是必需的。

目前，我们已经研究和开发的一个基于 MS/MS 的高通量蛋白质识别系统 pFind。pFind 对用户提交的 MS/MS 数据进行了有效的缩减和去噪音，并通过对质峰重新归一化的方法加强了对阶梯质峰系列的提取。pFind 的打分算法使用核函数技术计算实验质谱和理论质谱之间的相似性，从而形式化地考虑了离子之间的相关性。这有效地减少了假阳性搜索结果的产生。pfind 的核心程序由 C 语言写成，用户可以通过窗口界面指定数据库、酶、遗漏酶切位点、固定和可变修饰、肽碎片离子类型等参数，以及提交各种格式的 MS/MS 数据。pFind 系统还提供了其他质谱处理工具，比如 PeakCatcher 是一个阅读、修改和以图形方式显示质谱数据的方便工具。

3. 基于 NMR 的蛋白质三维结构识别

3.1 问题和方法

据统计，目前根据 NMR 或其他各种方式测定的蛋白质已经数以万计，如公用蛋白质数据库

PDB[22]，截至 2002 年 11 月 26 日，PDB 库中共有 19,311 个蛋白质的结构。同时每个 PDB 文件都提供了蛋白质分子的质量、温度、电荷数以及蛋白质分子的一级、二级、三级结构信息，包括残基的类型、个数、各种旋转、折叠、转角的信息和各个原子的三维坐标等。在这个世界公用蛋白质结构库的基础上，对蛋白质三维结构识别研究一般可分为两个方面：一是给定一个用 NMR 或其他方式测定的新的蛋白质三维结构，在 PDB 库中找出与之结构类似的蛋白质，以便于获取未知蛋白质的初步信息。二是通过已知结构和功能的蛋白质数据集进行全局和局部的蛋白质结构分析，获取关于结构与功能对应上的提示信息，为进-步创新药物设计奠定基础。

3.2 面向 PDB 的蛋白质三维结构识别

整个框架如图 2 所示，一般分为三个步骤，首先是预处理：即将 PDB 库中的蛋白质结构进行中心对准和主轴变换，使各个蛋白质结构在空间上达到最佳的比对位置，在此基础上抽取、选择特征值。然后在得到 PDB 库整体的特征后分别计算出各个蛋白质结构的特征值。然后是相似性度量：即找到需要作为参考蛋白质的三维结构信息，同样经过上述变换后得到特征值，并经过相似性度量公式得到 PDB 库中各个蛋白质与这个参考蛋白质的相似性程度,可理解为评分算法。最后是结果比较：即将评分结果与蛋白质家族库（FSSP 库）或与其它蛋白质比对软件（如 DALI）进行比较，得到实验结果并在此基础上进行实验分析。

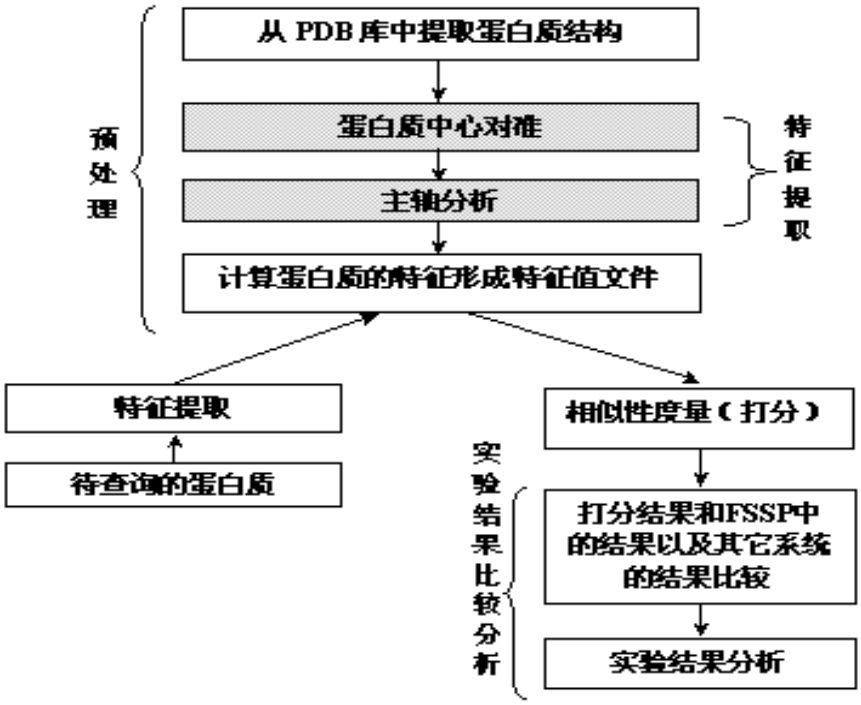

图 2：蛋白质三维结构识别框架

3.2.1 特征提取

如何从蛋白质三维结构中抽取最能反映其结构性质的特征是首先要考虑的问题。现在最基本的思路都是考虑蛋白质分子的一级到三级结构（如：不同类型的残基个数、不同类型的旋转、折叠和转角个数以及原子的三维坐标信息）[23][24]。DALI[25]提取的特征是蛋白质内部氨基酸之间相连的碳原子之间的距离，并用二维距离矩阵组来表示。STRUCTAL[26]提取的特征是蛋白质骨架上所

有肽基原子之间的距离：它首先计算两个结构中所有的肽基原子之间的的距离，这样对于参与比对的蛋白质结构来说，它中间的每一个肽基碳原子到被比对的蛋白质结构中的所有肽基碳原子之间的距离都可被计算出来。然后这个距离矩阵被转化为一个打分矩阵。VAST[27]提取的特征是参与比较的蛋白质上具有相同类型的二级结构片断之间的距离，这些二级结构片断在图中被表示为一对顶点。如果这两个相对应的二级结构之间的距离在某个阀值范围内，就在它们之间用线段相连。这样就形成了一个表征蛋白质二级结构之间类型、方向、连接属性关系的连通图。LOCK[28]提取的特征是参与比较的蛋白质的二级结构信息，它的算法核心是找到合适的比对残基，并使比对的残基之间的均方根差(RMSD)最小。3DSEARCH[29]提取的特征是所有目标蛋白质的所有二级结构向量，并在此基础上构建一个高冗余度的 HASH 表。而现在出现的一些更新的蛋白质匹配（识别）算法[30][31]更是同时考虑了一级到三级结构信息，并将这些信息附以一定的权值后加入打分算法。

随着对蛋白质结构-功能认识的深入，蛋白质结构识别算法需要抽取的特征信息将会越来越多。提取的描述蛋白质三维结构的特征应该满足旋转、平移与尺度不变的要求，同时要保证基于这些三维特征的给出的相似性度量与专家的经验感知结果基本一致。然而，生物学家的领域知识和经验不能完全固化到各式各样的特征提取算法中。因此，我们考虑可以引进交互式学习机制--相关反馈，它使系统在与用户的交互过程中不断的精确查询和动态的调整检索模型，从而适应不同用户的查询和不同的应用，实现符合用户需要的检索功能。

3.2.2 相似匹配的数学模型

由于对蛋白质三维结构的识别涉及到三维空间的搜索，因此如何使目标程序能够进行更有效率（包括准确度和速度）的计算也是一个重要的问题。DALI 用二位距离矩阵组来表示一个蛋白质内部氨基酸之间相连的碳原子之间的距离, 同时为了减小复杂度，DALI 采用了分支定界法（branch-and-bound algorithm）来得到近似解。由于 DALI 为了搜索全局最优的排列而采用了分支定界技术，DALI 算法的搜索空间要比其他几种匹配算法的搜索空间大。STRUCTAL 采用迭代动态规划算法（iterative dynamic programming）来使两个蛋白质骨架之间的均方根差（RMSD）达到最小。并且在将距离矩阵（代表肽基原子之间的距离）转化为打分矩阵后，通过动态规划算法得到蛋白质的最佳比对位置（包含从六种不同的初始状态开始的对齐操作）。STRUCTAL 系统在一些方面（如 immunoglobulin）的准确性要 DALI 和 LOCK 差很多。这可能是由于该算法在搜索空间上的不完备造成的（这个算法中采用的六个初始状态对于类似于 immunoglobulin 这样具有很多结构变化的结构来说是不够的）。VAST 算法是采用图形学中的算法来对二级结构进行比对。通过对连通图进行搜索找到其中的极大强连通子图，以此来找到同源性。LOCK 使用了迭代技术来选择比对的残基并且使它们之间的均方根差（RMSD）最小。这个迭代技术主要用来对用来比对的空间进行搜索，具体分为三个主要步骤：一、寻找局部二级结构的最佳重叠(Local Secondary Structure Superposition)，算法将参与比较的两个蛋白质的二级结构表示为向量，并且用动态规划方法来寻找这些向量的最佳的局部一致性。二、寻找原子的最佳重叠(Atomic Superposition)。三、寻找最佳的核心重叠（Core Superposition）,通过分析两个蛋白质结构上的对应原子之间的关系找出符合最佳重叠的最大的原子子集。在实际功能上，LOCK 系统在某些结构（如 myoglobin、TIM）上表现良好，甚至要好于 DALI（比如在 TIM 上）。但在有些类型（如 immunoglobulin）上表现不如 DALI，一个主要的原因是 LOCK 在它的比对算法中加入的刚体限制（rigid-body constraint）。由于 LOCK 使用分子内部距离来进行最后的比对计算，这个算法有时对于一些积累偏差处理上会很困难，而 DALI 由于采用比较灵活的打分策略，在这方面要好于 LOCK.。另外，由于 LOCK 的刚体限制，会有一些β折叠被排除在外，这样会影响最终得到的匹配结构中的原子个数。这两个方面都降低了 LOCK 的比对效果。在速度方面 LOCK 在氨基酸级上的比对速度是最快的（要比 DALI 快一个数量级）。3dSEARCH 采用一个高度冗余的 HASH 表（或索引表）来表示所有目标蛋白质的所有二级结构向量。在实际功能上，3dSEARCH

在某些结构上的表现相对较差（比如 immunoglobulin），这主要是由于 3dSEARCH 仅仅对二级结构作相似性比较造成的。另一方面，由于 3dSEARCH 主要基于比较二级结构向量的相对方向，它很难对那些只含有β折叠的蛋白质结构进行比较（因为这样的结构中的向量方向都是相似的）。同时，由于 3dSEARCH 的打分仅仅依赖于参与比对的结构向量的数量，它的准确度随着目标结构中向量数量的减少而显著降低。3dSEARCH 最大的优点是它的速度，它是所有算法中速度最快的，它比速度排在第二位的 LOCK 还要快将近两个数量级。

目前通常使用的是对于距离（残基之间的距离、原子之间的距离）的相似性度量。这种度量的计算方式有多种，比如一种简单的方式是 $\phi^R(i,j)=\theta^R-\left|d_{ij}^A-d_{ij}^B\right|$，这里的 A、B 分别代表相比较的两个蛋白质，$d_{ij}^A$和$d_{ij}^B$代表$A$、$B$的距离矩阵中对应的元 素，$\theta^R$代表相似性度量中的零值。

在打分方式上，DALI 服务器上采用的是灵活的相似性度量打分策略，它的计算公式为：

$$\phi_{(i,j)}^E=\begin{cases}(\theta^E-\dfrac{\left|d_{ij}^A-d_{ij}^B\right|}{d_{ij}^*})w(d_{ij}^*), i\neq j\\ \theta^E, i=j\end{cases}$$

最后的相似性分数为 $S=\sum_{i=1}^{L}\sum_{j=1}^{L}\phi(i,j)$ ，其中 L 代表进行比较的元素的个数。

DALI 系统从总体上说具有非常好的表现，这与它所采用的灵活的相似性打分方法有很大关系。

在 STRUCTAL 算法中采用的打分公式为 $score(i,j)=\dfrac{M}{1+\dfrac{\sqrt{dis\tan ce(i,j)}}{d_0}}$

此外，还有采用欧式距的其他一些打分方式。

$S=\dfrac{1}{N}\sum_{k=1}^{N}\dfrac{1}{1+(F_R[k]-F_Q[k])^2}$，其中 F_R和F_Q 分别表示参照和被参照的特征值。

关于结构匹配算法（全局、子结构），我们认为还可以从几个方面加以扩展：1）不同原子的质量可以被看作是不同的，现在一般的算法认为都是相同的；2）可以考虑在算法中采用更高维的矩，矩具有旋转，平移不变性；3）可加入更多生物相关的知识以完善识别算法，如结合位点（binding site）与几种离子、化学键的关系等等；4)在对各个特征值的权重取值上加入专家的反馈结果。

3.2.3 结果评价

目前的评价方式有两种，一种是与一些蛋白质的二级数据库，如 FSSP 库(FAMILIES OF STRUCTURALLY SIMILAR PROTEINS)进行比较。一种是与前面所说的几种国际上比较著名的蛋白质匹配软件进行比较。由于 FSSP 库中的数据都是经过具体的生物实验验证过的，因而与 FSSP 库的比较可以衡量出蛋白质比对软件在生物学意义上的实际效率；另一方面，与其他著名蛋白质匹配软件（如 DALI）的比较则可以在总结实验结果的基础上针对匹配软件（算法）的不足之处进行改进，可使匹配软件（算法）日臻完善。

4. 基于信息整合技术的公用数据库建立

随着人类基因组和蛋白质科研信息迅速膨胀，在基因组、蛋白质组水平上的生物学数据的有机整合、新的生物信息数据管理技术的成熟与否，成为生物信息技术发展的关键因素。将众多的异构生物学数据库从技术上解译并且在物理上移植到一个单独的结构一致的数据库管理体系之中，形成统一的用户使用界面，实现生物信息数据整合、智能化的多重、复合和交叉检索和基于高性能计算网格的数据共享，将为我国功能基因组研究、药物研发提供基本的信息技术支撑。目前国内外大部分数据库体系在生物信息数据的整合上采取的方法都是建立分散在不同的异构数据库中的数据之间的简单链接，数据库并没有有机地整合在一起，不能实现智能化的多重、复合和交叉检索。如何能充分利用互联网资源实现以上目标是研究的重点问题。各国研究人员已经开始着手解决数据库体系分散问题，美国和欧洲有关机构最近就宣布，将把全球三大主要蛋白质数据库资源集中起来，建设一个全新的蛋白质数据库。

结合蛋白质质识别研究工作的需要，我们考虑可建立如下的数据库：1）公用蛋白质组数据库，主要在 Swiss-Port、TrEMBL 和 TrEMBL-new 所有的非冗余的蛋白质信息数据的基础上建立一个公用蛋白质组数据库。2）多肽数据库，将产生的各种不同的多肽片断放入数据库，并与相应的酶、蛋白序列关联，记下产生的条件以及相应的位置等信息，并计算每一条多肽片断的分子量。3）整合其他资源所得的数据库，一是建立非冗余的 Cluster dbEST 以及 TC-蛋白质数据库。二是建立候选蛋白质数据库。具体来说，首先从 dbEST 以及 dbTc 出发，建立非冗余的 Cluster dbEST。其次，由于 EST 是核酸序列中与蛋白质序列关系最为密切的信息，必须充分利用这部分信息。

另外，在对大量的新鉴定的蛋白质在蛋白质组水平上进行功能与结构分析时，整合的蛋白质组数据仓库和联邦数据库体系是研究最基本的出发点，因此，可将在已有的国际上通用的 XML[32]、和 Relational 数据库技术工作基础上，进一步采用合适的 JAVA 中间件技术，结合 Gene Ontology 标准，构建数据模型，形成一套跨平台的、可整合大量主要的异构生物信息数据库，并可根据注释内容扩展的数据管理、搜索和可视化的新的数据仓库系统和联邦数据库系统，针对大批量大白质组数据的导入、冗余信息的剔除等，重要关键信息的整合，开发新的技术，使数据库体系更加完善。

5. 结束语与展望

利用信息工程技术，辅助生物学家开展面向质谱技术的高通量蛋白识别以及面向 NMR 的蛋白质三维结构识别以及功能发现研究成为蛋白质组中的生命科学与信息科学交叉研究的热点和挑战问题，本文侧重从信息技术的角度对这两方面问题进行研究综述并对可能开展研究的问题和方法进行了归纳并提出我们解决问题的思路和方法。总之，利用信息技术可以辅助完成致病相关蛋白识别及有关蛋白结构功能的研究，这项研究成果可辅助生物学家用于理解生命现象的本质，并找到提高人类生命质量的办法，这是今后生命科学研究所面临的一个主要课题，也是基因组计划的必然发展方向。它不仅是一个重大的关系全局的基础理论性的问题，关系到人口与健康领域的方方面面，同时也是一个将给生物技术发展和应用带来彻底革命的关键性研究课题，并将会毫无疑问地大大推动生物医药等产业的发展。

6. 参考文献

[1] Pappin, D.J.C., Hojrup, R, Bleasby, A.J., Rapid identification of proteins by peptide-mass finger printing. Current Biology, 1993, 3:327-332.

[2] 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, 20:3551-3567.

[3] Zhang, W, Chait, B.T., ProFound — an expert system for protein identification using mass spectrometric peptide mapping information. Analytical Chemistry, 2000, 72:2482-2489.

[4] Wilkins, M.R., Gasteiger, E, Bairoch, A, Sanchez, J.C., Williams, K.L., Appel R.D., Hochstrasser D.F., Protein identification and analysis tools in the ExPASy server. Methods Molecule Biology, 1999, 112:531-552.

[5] Clauser, K.R., Baker, P, Burlingame, A.L., Role of accurate mass measurement (± 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. Analytical Chemistry, 1999, 71:2871-2882.

[6] J.A.Taylor and R.S.Johnson, Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. Rapid Comm. Mass Spectrum.11:1067-1075,1997.

[7] J.A.Taylor and R.S.Johnson,Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. Analytical Chemistry. 73:2594-2604,2001.

[8] V. Dancik, T.A.Addona, K.R.Clauser, J.E.Vath, P.A.Pevzner, De Novo Peptide Sequencing via Tandem Mass Spectrometry: A Graph-Theoretical Approach. RECOMB 99, pp 135-144,1999.

[9] Ma, B., Zhang, K., Lajoie, G., Doherty-Kirby, A., Liang, C. and Li, M. 2002. A powerful software tool for the de novo sequencing of peptides from MS/MS data. Abstracts of the 50th ASMS Conference on Mass Spectrometry and Allied Topics, June 2-6, p70.

[10] S. Baginsky, M. Cieliebak, W. Gruissem, T. Kleffmann, Z. Liptak, M. Muller, and P. Penna, AuDeNS: A Tool for Automatic De Novo Peptide Sequencing. Technical Report no. 383, ETH Zurich, Dept. of Computer Science

[11] J. Fernandez-de-Cossio, J. Gonzalez, T. Takao, Y. Shimonishi, G. Padron, and V. Besada, A Software Program for the Rapid Sequence Analysis of Unknown Peptides Involving Modifications, Based on MS/MS Data. 45th ASMS Conference on Mass Spectrometry and Allied Topics, 1997.

[12] Ting Chen, Ming-Yang Kao, Matthew Tepel, John Rush, George M.Church, A Dynamic Programming Approach to De Novo Peptide Sequencing via Tandem Mass Spectrometry. Proc. of the 11th SIAM-ACM Symposium on Discrete Algorithms (SODA 2000),pp 389-389,2000.

[13] D.R. Goodlett, A. Keller, J.D. Watts, R. Newitt, E.C. Yi, S. Purvine, J.K. Eng, P. von Haller, et al.2001. Differential stable isotope labeling of peptides for quantitation and de novo sequence derivation. Rapid Comm. Mass Spectrum. 15:1214-1221.

[14] 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 Am. Soc. Mass Spectrum, 1994, 5:976.

[15] Jonathan, C. C. and Brian, N. G., MaxEnt: An Essential Maximum Entropy Based Tool for Interpreting Multiply-Charged Electrospray Data. Micromass UK Limited. Available at http://www.micromass.co.uk/.

[16] Fenyo, D, Qin, J, Chait, B.T., Protein identification using mass spectrometric information. Electrophoresis, 1998, 19:998-1005.

[17] Clauser, K.R., Baker, P, Burlingame, A.L., Role of accurate mass measurement (± 10 ppm) in

protein identification strategies employing MS or MS/MS and database searching. Analytical Chemistry, 1999, 71:2871-2882.
[18] Mann, M, Wilm, M: Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Analytical Chemistry, 1994, 66:4390-4399.
[19] Vineet, B., Nathan, E., SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. Bioinformatics, Vol. 17 Suppl. 1 2001, pp.13-21.
[20] Helen, I.F., David, F, Ronald, C.B., RADARS: a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. Proteomics, Vol.2, Issue 1, 2002. pp. 36-47.
[21] Yates, J.R., Eng, J.K., McCormack, A.L., Schieltz D, Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Analytical Chemistry, 1995, 67:1426-1436.
[22] Liisa, H., Chris, S., Searching Protein Structure Databases Has Come of Age, Proteins, 1994, Vol.19, pp.165-173.
[23] Liisa, H., Chris, S., Mapping the Protein Universe, Science, 1996, Vol. 273, pp. 595-602.
[24] Minoru Kanehisa，Post-genome Informatics， OXFORD university press, 2001.
[25] Liisa Holm, Chris Sander, Protein Structure Comparison By Alignment of Distance Matrices, Journal of Molecule Biology, 1993, Vol.233, pp.123-138.
[26] Gerstein, M, Levitt, M, Using Iterative Dynamic Programming to Obtain Accurate Pairwise and Multiple Alignments of Protein Structures, Proceedings of the Fourth International Conference on Intelligent Systems in Molecular Biology, 59-67 (Menlo Park, CA, AAAI Press, June 12-15,1996).
[27] Gibrat, J.F., Madej, T., Bryant, S.H. Surprising similarities in structure comparison. Current Opinion in Structural Biology. 1996, Vol. 6, 377-385.
[28] Amit, P.S., Douglas, L.B., Hierarchical Protein Structure Superposition using both Secondary Structure and Atomic Representations, ISMB-97, Vol. 4, 284-293.
[29] Amit, P.S., Douglas, L.B., Protein Structure Alignment: A Comparison of Methods，1999.
[30] Shannching Chen, Tsuhan Chen, Retrieval of 3D Protein Structures, ICIP2002, Rochester, NY, U.S.A., September 2002.
[31] Shannching Chen, Tsuhan Chen, Protein Retrieval By Matching 3D Surfaces, GENSIPS 2002, Raleigh, North Carolina, USA, October, 2002.
[32] F. Achard, Vaysseix G., Barillot, E., XML, bioinformatic and data integration, Bioinformatic, 2001, Vol.17, No.2，pp.115-125.