摘 要

随着生物大分子数据库中蛋白质序列数目的增多，发展有效的方法，从氨基酸序列提取结构信息成为后基因组时代的重要研究课题。越来越多的证据表明，天然蛋白质的折叠类型在总数上的是有限的，一般认为只有数百到数千种，远小于蛋白质所具有的自由度数。Anfinsen 原理指出蛋白质的结构在很大程度上由其序列决定，当蛋白质结构数据库趋于完备以后，结构的解析问题就可以转化为折叠识别问题，即使用计算方法，找到与待预测蛋白质序列在三维结构上最匹配的已知折叠类型。对自然界存在的数百到数千种折叠类型进行系统研究，有助于揭示蛋白质的折叠规律，可为大型的蛋白质数据库提供结构注释，或者为蛋白质的精确结构预测提供参考。

目前的蛋白质折叠类型识别基本上都是靠专家来完成的，不同的库分类颇不相同。SCOP 通过观察将蛋白质按照同源性进行分类，但在 SCOP 的一些折叠子中，二级结构及其走向存在很大差异，为折叠识别的模型构建造成困难；CATH 以序列、结构比对的相似性打分为依据进行“Topology”的分类，并没有直接反映出蛋白质之间二级结构及其空间排布上的相似性。实际上，蛋白质的折叠类型反映了蛋白核心结构的拓扑结构模式，包括蛋白质分子空间结构的三个主要方面：二级结构单元、二级结构单元的相对排布位置以及蛋白质多肽链的整个路由关系（即肽链走向），我们在蛋白质折叠结构研究的基础上，以结构域的拓扑不变性为依据，结合二级结构片段的空间排列、取向特征和连接关系，进行蛋白质折叠类型分类，建立了 LIFCA 数据库，为蛋白质折叠识别奠定了基础。

折叠识别的一个重要方面是建立折叠识别算法。目前，折叠识别的方法大体上可以分为三类：氨基酸序列的两两比较，如使用 Blast 和 Fasta 判断序列之间的相似性；多序列建模，如 Profile HMM 方法；分类器，如神经网络，支持向量机等。与两两比较方法相比，HMM 建立了统一的模型，可以抓住一组同源序列的公共核心，因而对于那些在已知数据库没有高相似度模板的未知序列有更好的识别效果；与 SVM 等分类机器相比，虽然后者可能获得较高的准确率，但是 Profile HMM 有较为统一的构架，保留了位点信息，有详细的序列概形，与 SVM 相比更有助于对特定折叠类型进行进一步的分析和研究。

在本文中，我们在前期折叠分类的基础上，对 α 类、β 类及 α/β 类中样本数量较多的 74 种折叠类型进行研究，利用结构比对得到多序列比对，继而产

生 Profile HMM 进行识别，研究工作主要包括以下几个方面：

1. 蛋白质折叠类型分类数据库 LIFCA 的建立

选取序列同一性低于 25%的 2406 个蛋白质，包含了全α、全β，α/β三个结构类的所有代表性序列，在蛋白质折叠结构研究的基础上，以结构域的拓扑不变性为依据，结合二级结构片段的空间排列、取向特征和连接关系，进行蛋白质折叠类型分类，对于结构域的序列、二级结构等信息，提供了详细的注释。为蛋白质折叠类型识别奠定基础。

2. 折叠类型的结构比对研究

使用结构比对算法对 LIFCA 进行了同种折叠类型在结构上的差异性研究，以结构比对为基础得到了折叠类型的多序列比对结果，用于建立折叠类型的模型。

3. HMM 模型数据库的建立

对成员数目大于等于 4 并且结构比对效果较为显著的 74 个折叠类型分别建立 HMM 模型，组成隐马尔科夫模型库。使用非冗余的 Astral1.65 序列库进行识别检验，识别精度 74.5%，并保持了很低的假阳性率，识别效果比目前报导的一些方法识别效果均要好。

本文从数据集筛选及算法两个方面对蛋白质的折叠类型识别方法进行了改进，建立的隐马尔科夫模型库覆盖范围较广，识别准确率高，为折叠类型识别提供了一种新思路，对于相关的研究工作有参考价值，同时为进一步的研究提供了基础。

关键词 蛋白质； 折叠类型识别； 折叠类型分类； 隐马尔科夫； 结构比对

ABSTRACT

With the increasing of protein sequence in the bio-macromolecule database, development of new methods to extract structural information from amino acid sequences becomes an important research topic in the post-genome era. More and more evidences shows that the number of natural protein folds is limited, usually from hundreds to thousands, which is much less than the number of DOF obtained by proteins. The Anfinsen's principle suggests that protein's structure is mostly determined by its sequence. While the structural database tending to completeness, the problem of structural analysis becomes the one of fold recognition, which is, finding the best-matching three dimensional structural fold. Systematic research of those folds is meaningful to uncover the principle of protein folding, to provide structural annotation for large protein database, or helping for precise protein structural prediction.

Currently, protein fold recognition mostly depends on experts, and different database has different principle. By means of observation, SCOP classifies proteins based on homology, however, for some fold, it is difficult to construct fold recognition model since their secondary structure and its strike direction does exist difference. The classification of topology in CATH is based on the similarity score of sequence and structure alignment, which dose not directly show the similarity existing in protein secondary structure and its space assignment. In fact, protein fold type reflects the topology of protein core, which contains three aspects of protein space structure: element of secondary structure, relative assignment of SSE in sequence and entire route relationship of polypeptide chains (means direction of polypeptide chains). Based on modern protein fold research and the conservative of protein domain topology, we reclassify protein domains from three aspacts: the assignment, the direction characteristics, and the connection relationship of protein SSE. Finaly, a database named LIFCA was built, which formed the base of protein fold recognition.

A significant aspect of fold recognition is to develop new algorithm. For modern research, there are mainly three kinds: pair comparison between amino acid sequences (e.g. checking the sequences similarity by means of Blast and Fasta), model construction based on multiple sequences alignment (e.g. Profile HMM method) and classification machine (e.g. NN, SVM). Compared with pair comparison, HMM could construct uniform model and extract the core of multiple homologous sequences, thus

it has better recognition result to these sequences which do not exist high similar template in those known databases. In addition, although classification machine such as SVM could obtain higher accuracy, profile HMM has some unreplaceable merits, such as more uniform framework, keeping informations of conservative locus, detailed statistical analysis of amino acids in sequences, etc. Also for profile HMM, sequence model could be simply obtained by a multiple alignment, which is more suitable to further analysis and research. The main work of this paper includes the following:

1. Establish LIFCA database based on the topologies of folding cores

Choose 2,406 protein sequences from Astral with sequence identity 25% or low. The mainly α, mainly β, and α/β structure class are included in. Then reclassify those protein domains based on the study of protein folding, which means the SSE contents, their arrangement orientation and connections. This work laid the foundation for further research.

2 Structure-based sequence alignments within topologies

For each topology, a structure-based sequence alignment are conducted, the difference within each topology is researched too. The multiple alignment results from this step are used for model building.

3. Profile HMM library

There are 74 representative topologies which contain no less than 4 members in LIFCA, so totally 74 Profile HMM model are establishing. Using Astral1.65 100% identity sequence database for test datasets, the classification accuracy is 74.5%, still maintain a low false positive rate than other identification methods, the Profile HMM library performance better in most topologies.

In this paper, data sets and algorithm has been improved both, the hidden Markov model library based on this method gets a broader coverage and a good accuracy rate. For related research work, it's valuable.

Keywords: Protein, topology recognition, topology, Hidden Markov Model, structure alignment

独 创 性 声 明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京工业大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签名：[signature] 日期：2008.6.2

关于论文使用授权的说明

本人完全了解北京工业大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

（保密的论文在解密后应遵守此规定）

签名：[signature] 导师签名：[signature] 日期：2008.6.2

第 1 章 绪 论

1.1 蛋白质结构及其研究意义

几乎一切生命现象都通过蛋白质的功能而表现出来，而蛋白质的结构又被认为是功能行使的基础，如果只了解蛋白质的一级结构而不了解蛋白质的空间结构，那么就很难阐明生物大分子之间的关系。例如胰凝乳蛋白酶就通过三个一级结构不相连的氨基酸共同催化底物化学反应，只有通过多肽链的盘旋、折叠形成空间结构才能解释这一现象[1]。揭示每种蛋白质的空间结构，对于蛋白质结构与功能关系的研究、蛋白质工程改造等都是非常必要的。

蛋白质的结构是有层次的，分为一级结构，二级结构，超二级结构，结构域，三级结构，较大的蛋白质分子还具有四级结构，不同层次的研究具有不同的意义，一级序列是高级结构的基础，可以通过实验手段测定；二级结构只有 α 、β 、无规卷曲等少数的几种，二级结构的预测一直是近年来研究工作的热点，目前准确度可以达到 70%左右；与功能关系最为直接的是结构域和三级结构及四级结构，本文研究工作的重点正是结构域的分类和识别，我们希望能通过对结构域的深入研究来帮助促进肽链折叠、蛋白质功能分类的研究工作[2-6]。

一般情况下，一段天然的氨基酸链与一个天然态的蛋白质相对应，目前由基因测序等手段得到的序列数据已经非常多，为了了解这些序列如何行使生物学功能，对于大规模蛋白质结构测定的要求是非常迫切的。但是由于蛋白质难以结晶等问题，使得实验手段测定蛋白质结构异常困难，序列数据的迅速扩张和结构数据的匮乏形成明显的对比，因此理想的方法是在有目的的实验测定的基础上，最终通过计算方法对所有蛋白质的结构加以解决。在目前研究中，蛋白质结构的解析可以分为两个部分：一方面，用 X－ray，NMR 等试验手段测定更多蛋白质的空间结构保存到数据库中，提供大量高精度的结构模板；另一方面要发展利用这些实验模板的计算方法，使用比较建模、串线、动力学计算等手段预测尽可能多的蛋白质空间结构，从而建立一些模式生物蛋白质表达谱和重要疾病相关蛋白质表达谱，解释复杂的生物学过程[4-5]。

由于蛋白质折叠问题本身的复杂性，通过计算方法来解析蛋白质的精细结构仍面临诸多难题。但是在一些关于蛋白质的进化、功能分类等研究层次上，对于

蛋白质精细结构的要求不是非常严格，三维拓扑粗粒化的预测就可以作为较为可信的参考，那些序列同一度很低，而三维拓扑走向相同的蛋白质在功能上往往是相近的，这可能是收敛进化的结果。本文的研究工作并不是以计算精确的三维结构为目的，而是针对蛋白质结构域层次上的拓扑走向分类预测。与精细结构预测相比，这样的预测容易实现，结果较为可信，计算成本低，同时，以此为基础可以进行进化、功能等相关方面的研究，也可以为精细结构预测的模板选择提供有价值的参考。

近年来，蛋白质结构的研究进展使人们对生命过程和本质的认识已大大提高，但是随着许多问题的解决，更多深刻的问题又被提出来。相信在新的世纪，在后基因组研究计划的推动下，蛋白质的相关研究会获得更大的发展，提升我们对生命活动本质的认识。

1.2 蛋白质结构的层次

蛋白质的结构是分层次的，一级结构即序列，由 20 种氨基酸组成，在此基础上形成了蛋白质的二级结构，根据一些折叠理论，这些二级结构通过延伸、碰撞等过程形成了三级结构，进而行使复杂的功能。

1.2.1 蛋白质的基本组成单位

氨基酸是蛋白质的基本组成单位，氨基酸是带有氨基的有机酸，它有一个氨基、一个羧基、一个氢原子和一个 R 基团组成（如图 1-1）。每一个从细菌到人类的所有物种中，一切蛋白质都是由 20 种氨基酸构成的。氨基酸侧链 R 的大小、形状、电荷、形成氢键的能力和化学活性方面都存在着差异。蛋白质的功能范围之所以如此之广，就是由于这 20 种氨基酸的差异，以及它们的各种组合的变化结果，表 1-1 是 20 种氨基酸的简写和符号。

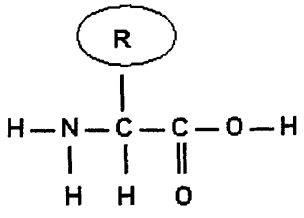


图 1-1 氨基酸结构示意图

Fig.1-1 the schematic drawing of structure of amino acids

表 1-1 20 种氨基酸的符号

Table 1-1 Symbols of 20 amino acids

氨基酸	英文符号	氨基酸	英文符号
丙氨酸	A(Aln)	甲硫氨酸	M(Aln)
半胱丙氨酸	C(Cys)	天冬酰胺	N(Asn)
天冬氨酸	D(Asp)	脯氨酸	P(Pro)
谷氨酸	E(Glu)	谷氨酰胺	Q(Gln)
苯丙氨酸	F(Phe)	精氨酸	R(Arg)
甘氨酸	G(Gly)	丝氨酸	S(Ser)
组氨酸	H(His)	苏氨酸	T(Thr)
异亮氨酸	I(Ile)	缬氨酸	V(Val)
赖氨酸	K(Lys)	色氨酸	W(Trp)
亮氨酸	L(Leu)	酪氨酸	Y(Tyr)

1.2.2 蛋白质的结构层次

由于蛋白质的立体结构的形成是分阶段的，同时在已知立体结构的蛋白质中也看到了不同类型的规则的有序结构，因此在这个基础上提出了蛋白质的结构是立体的多层次的学说。通常把蛋白质结构层次分为一级结构、二级结构、超二级结构、结构域、三级结构以及四级结构（图 1-2）。

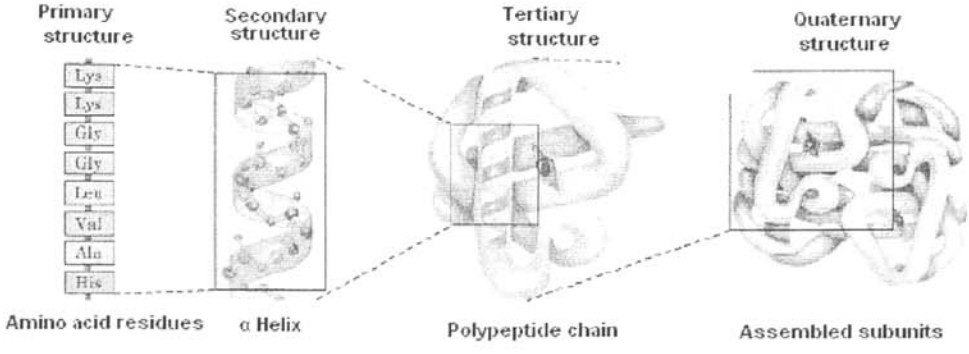


图 1-2 蛋白质的结构层次

Fig.1-2 the structural levels of protein

（1）一级结构

蛋白质一级结构是指多肽链的氨基酸残基的排列顺序，也是蛋白质最基本的结构。它是由基因上遗传密码的排列顺序决定的，各种氨基酸按遗传密码的顺序通过肽键连接起来。图 1-3 是溶菌酶的一级结构。

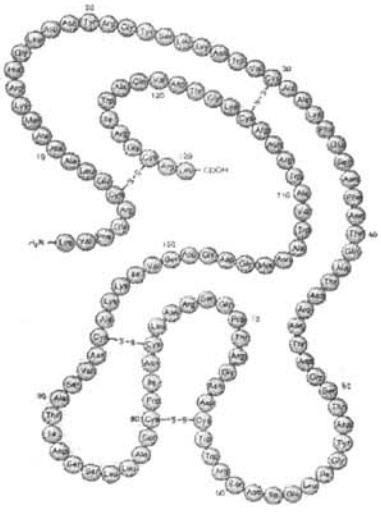


图 1-3 溶菌酶的一级结构

Fig.1-3 the primary structure of lysozyme

（2）二级结构

蛋白质二级结构是指多肽链借助于氢键沿一维方向排列成具有周期性的结构的构象，是多肽链局部的空间结构（构象），主要有 α 螺旋、β 折叠、β 转角、无规卷曲等几种形式，它们是构成蛋白质高级结构的基本要素。图 1-4 是几种二级结构的示意图。

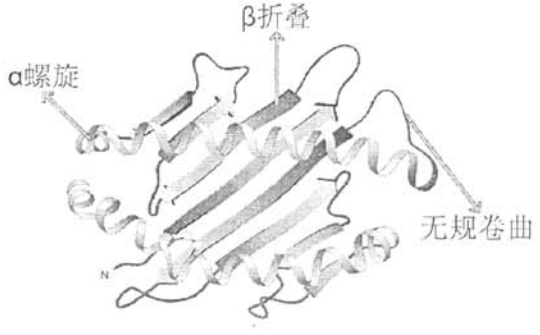


图 1-4 蛋白质的二级结构

Fig.1-4 the secondary structures of Protein

（3）超二级结构

超二级结构是介于蛋白质二级结构和三级结构之间的空间结构，指相邻的两个或者多个二级结构单元组合在一起，彼此相互作用，排列成规则的、在空间结构上能够辨认的二级结构组合体，并充当三级结构的构件，其基本形式有 αα、βαβ、βββ 等。图 1-5 是几种蛋白质超二级结构。

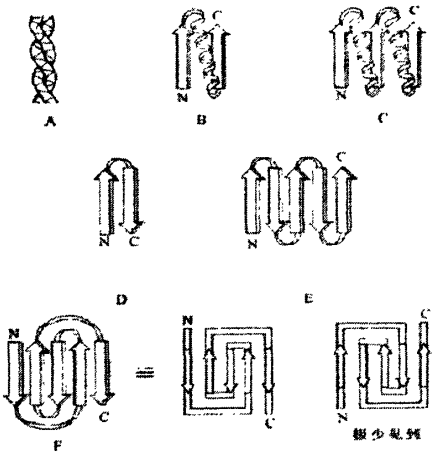


图 1-5 蛋白质超二级结构

Fig.1-5 the supersecondary structures of protein

（4）结构域

结构域和超二级结构都是在二级结构和三级结构之间公认的过度层次，结构域通常是超二级结构与二级结构或者多个超二级结构相互结合构成，一般要大于超二级结构。

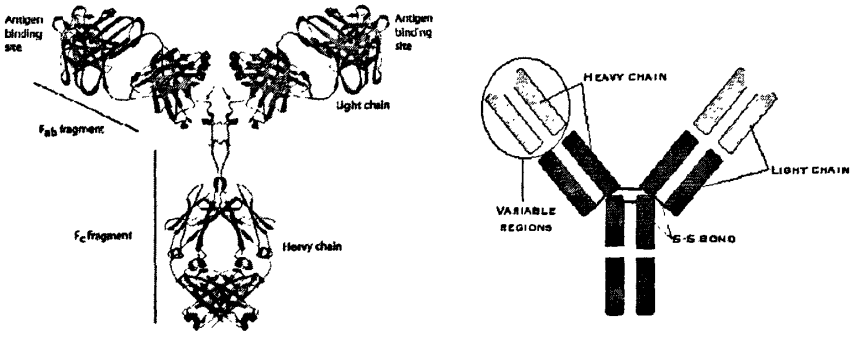


图 1-6 免疫球蛋白质的结构域划分

Fig.1-6 the domains of IgG protein

首先从结构比较和解剖的角度，注意到一些大的球状蛋白可以被拆分为若干部分，各部分之间存在较为明显的间隙。从蛋白质折叠的角度，也认识到一条很长的肽链在折叠过程中，可能先由部分二级结构构象单元或超二级结构折叠成局部的、具有三级结构的区域，而后几个区域通过一定长度的肽链相连，成为一个完整的蛋白质立体结构。

根据蛋白质化学多方面的研究和观察，提出了结构域的概念，可以理解为蛋白质构象单元组成的一些实体，它们有一定的三级结构，而且往往有特定但不

完全的生物活性，很多实验指出，这些结构域之间的连接肽段或一些连接的肽键，经过蛋白水解酶酶解断裂以后，可以分离成彼此独立的实体。

如图 1-6，免疫球蛋白 G（IgG），一级结构上，由两条轻链和两条重链组成，不论从一级序列、高级结构特征、或者蛋白酶水解实验，都可以分为 12 个结构域。每条轻链由两个结构域组成，每条重链由四个结构域组成。单独的 fab 结构域保留了部分功能，只能与抗原结合，但是不能引发免疫反应。

（5）三级结构

三级结构主要针对球状蛋白质而言，指的是整条多肽链由二级结构元件构建成的总三维结构，包括一级结构中相距远的肽段之间的几何相互关系，骨架和侧链在内的所有原子的空间排列。图 1-7 是溶菌酶分子的三级结构。

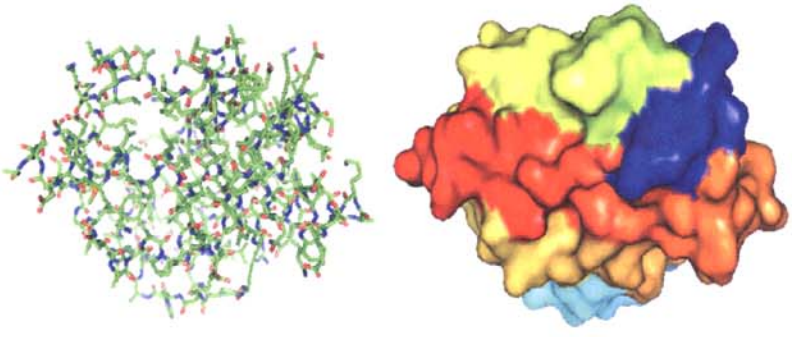


图 1-7 球蛋白的三级结构

Fig.1-7 the tertiary structure of Globins

（6）四级结构

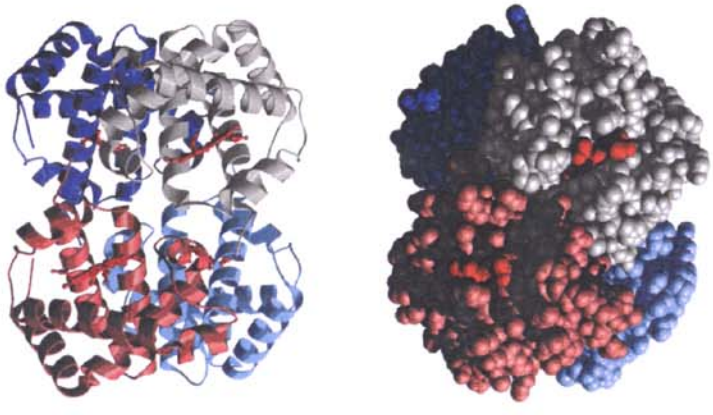


图 1-8 血红蛋白的四级结构

Fig.1-8 the quaternary structure of hemoglobin

蛋白质的四级结构是指在亚基和亚基之间通过疏水作用等次级键结合成为有序排列的特定的空间结构。四级结构的蛋白质中每个球状蛋白质称为亚基，亚基通常由一条多肽链组成，有时含两条以上的多肽链，单独存在时一般没有生物活性。图 1-8 是一个简单的四级结构示例，两个蛋白质结构域结合在一起以实现功能。

1.3 蛋白质结构分类

截至 08 年四月，PDB 数据库中包含了约 4 万条蛋白质记录，有约 4 万条较高质量的数据（见表 1-2），经过专家的细致工作，从其中可以剥离出 9 万多个结构域。由于肽链折叠的复杂性，从实验测得的一个晶体数据切割、划分不同的结构域是一个繁杂的过程，存在一条链多条结构域、一个结构域多条链、结构域之间分界不明晰等等问题， SCOP[6-7]在这方面已经做了大量基础性的工作并取得了普遍的认可，下文中所有的研究工作，若非特殊说明，折叠类型的分类和识别均是以 SCOP 专家分割过的结构域为基本单位进行。

表 1-2 PDB 保留的蛋白质数目

Table 1-2 Current Holdings of protein in PDB

Exp.Method	X-ray	NMR	Electron Microscopy	Other	Total
Records	40066	6321	119	88	46594

规范化的蛋白质结构分类以数据库的形式发布，对于结构域的分类，应用较为广泛的有 SCOP（按照同源性分类）、CATH[8-13]（根据结构、序列相似性分类）、FSSP[14]（通过结构比对分类）等数据库。这些库的分类的标准存在差异，分类的手段不尽相同。研究者们分别从结构、进化、功能等不同的角度进行分类以解决不同的问题。不同的分类在某些层次差距较大，在一定的层次上，又大致相当，后面会有相关讨论。

1.3.1 SCOP 分类及折叠子（Fold）

蛋白结构分类数据库(Structural Classification of Proteins, SCOP)是英国医学研究委员会(Medical Research Council, MRC)属下的 MRC 分子生物学实验室和蛋白质工程研究中心(MRC Laboratory of Molecular Biology and Centre for Protein Engineering)开发维护的，它是通过手工比较辅以自动计算方法，对已知结构的

蛋白质进行相似性分析和进化同源性分析得到。

SCOP 的构架是一种层状结构，把蛋白质结构域从粗到细分成 7 个层次：(1)根(Root)；(2)结构类(Class)；(3)折叠子(Fold)；(4)超家族(Superfamily)；(5)家族(Family)；(6)蛋白质(Protein)；(7)种属(Species)。

表 1-3 SCOP 分类统计（1.73 版）

Table 1-3 Scop Classification Statistics（1.73 release）

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	259	459	772
All beta proteins	165	331	679
Alpha and beta proteins (a/b)	141	232	736
Alpha and beta proteins (a+b)	334	488	897
Multi-domain proteins	53	53	74
Membrane and cell surface proteins	50	92	104
Small proteins	85	122	202
Total	1086	1777	3464

从分类程序上，SCOP 的分类先从序列家族开始，序列同一性大于 30%的氨基酸序列可以归为一个家族，当蛋白质结构功能上有非常大的相似性，也可以降低序列相似性要求，家族分类使用的是序列比对方法。

在家族的层次之上划分超家族：当家族之间结构比对相似性非常显著，或者有功能上的研究证据表明他们可能来自共同的进化祖先，则归在同一个超家族，超家族划分的方法是结构比对和参阅相关的文献。

SCOP 的折叠子是在超家族的基础上，按照二级结构及其空间分布及拓扑连接进行分类。在物理和化学的角度上讲，在蛋白质折叠过程中，出于对结构打包方式的要求，不同源的蛋白质可能有结构上的相似性，因此这些蛋白质可以从结构的角度综合考虑，它们之间的功能可能是没有关系的，也有可能是有不明显的进化关系及未被发现的关系，目前在 SCOP 中，这一层的分类只能依靠专家的经验人工完成[6-7]。

折叠类是按照二级结构含量分的，即：(1)全 α 类(All alpha proteins)；(2)全 β 类(All beta proteins)；(3) α/β 类(Alpha and beta proteins(α/β))；(3) α+β 结构类(Alpha and beta proteins(α+β))；(5)多结构域 αβ 混合蛋白质(Multi-domain proteins (alpha and beta))；(6)膜蛋白与细胞表面蛋白质和肽(Membrane and cell surface proteins and peptides)；(7)小蛋白(Small proteins)。

SCOP 的最新统计数据及分类情况见表 1-3。

1.3.2 CATH 分类及拓扑（Topology）

CATH 数据库蛋白质分类的构架有 5 个层次：(1)结构类(Class)；(2)框架(Architecture)；(3)拓扑(Topology)；(4)同源(Homology)；(5)序列(Sequence)。CATH 的名称就来源于 4 个层次的英文缩写。

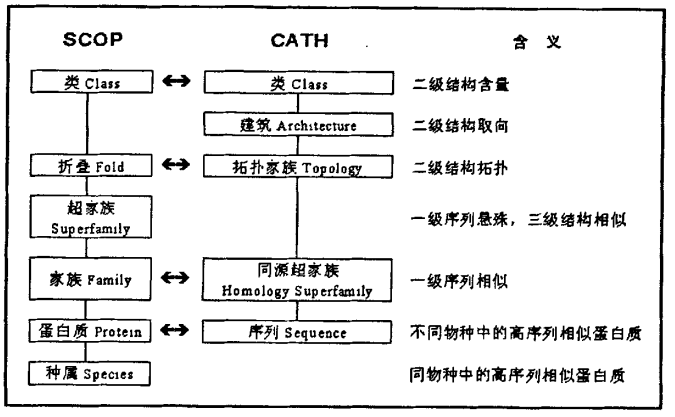


图 1-9 SCOP 与 CATH 数据库结构比较示意图

Fig.1-9 Schematic representation for comparison between SCOP and CATH databases

表 1-4 CATH 分类统计（3.1.0 版）

Table 1-4 CATH Classification Statistics（3.1.0 release）

C	A	T	H	S	O	L	I	D
Mainly Alpha	5	305	652	1850	2329	3001	5587	19729
Mainly Beta	20	191	415	1860	2531	3846	6503	25537
Alpha Beta	14	496	922	3922	5303	6659	12998	47193
Few Secondary Structures	1	92	102	162	200	275	403	1426
Total	40	1084	2091	7794	10363	13781	25491	93885

第一层次主要是按蛋白质二级结构的含量来划分。目前分为 9 个“类”：(1) α 类(Mainly Alpha Class)；(2) β 类(Mainly Beta Class)；(3)混合 α β 类(Mixed Alpha-Beta Class)；(4)无二级结构类(Few Secondary Structures Class)；(5)多结构域类(Multi domain Class)；(6)预分配的单结构域类(Preliminary single domain assignments Class)；(7)Psi-Blast 序列家族类(Psi-Blast sequence families Class)；(8)大于 35% 序列家族类(CATH-35 Sequence families)；(9)多链结构域片断类(Fragments from multi-chain domains)。其中，前 4 类是蛋白质主要的类别，第 3 类包含了 α+β 类和 α/β 类。

第二个层次是 CATH 的特色之处，它反映了二级结构的空间取向，而不考虑二级结构之间的连接方式，故称之为“框架”。形象地说，建筑刻画了蛋白质空间的结构模体，可直观地区分蛋白质 3D 结构的框架。例如，3 层三明治(3-layer sandwich)，桶(barrel)，马蹄(horseshoe)，螺旋桨(propeller) 等，人工完成分类。

第三层次为“拓扑”，与 SCOP 的第二层次“折叠子”相似，它是根据二级结构单元(secondary structural element)的拓扑连接对蛋白质结构进行分类的，其识别方法是基于识别结构相似性的经验算法 SSAP[12]，那些 SSAP 程序打分大于 70，且二者中大结构 60%以上与小结构匹配的蛋白质被归入同一拓扑家族。

第四层次为“同源”，与 SCOP 的第三层次“家族”类似，CATH 在此把共同祖先(具有相似结构或功能)的蛋白质放入同一个同源超家族中，但 CATH 更注重结构本身的相似性。

第五层次被称为“序列”，它具有比同源更强的氨基酸序列相似性，即氨基酸序列全同性大于 35%，且较大蛋白质至少有 60% 与较小蛋白质相匹配，包括具有相似的结构和功能。CATH 与 SCOP 的异同如图 1-9 所示，CATH V3.1.0 的分类见表 1-4。

1.3.3 FSSP 自动分类

FSSP 是结构相似蛋白质家族(Families of Structurally Similar Proteins)的缩写，它们的蛋白质来源于 PDB。目前 FSSP 数据库中约有 330 种具有代表性的蛋白结构家族，收录了 3 242 个序列家族，30 624 蛋白质，收录蛋白质的标准为：彼此结构同源性范围为 30～70%，小于 30%被认为同源性较小，大于 70% 则结构差别不大。

FSSP 的作用在于：(1)可用于研究蛋白质折叠进化中保守性与多样性；(2)研究结构相似蛋白质之间的关系；(3)确定蛋白质结构的核心部分，以便进行模建或者蛋白质改造；(4)检测同源性分析结果的可靠性；(5)蛋白质结构统计分析。

SCOP 对于折叠子的分类完全由人工完成，而 CATH 的 Topolgy 则是通过 SSAP 算法打分估计其同源性来得到，在更大的层次上实现了聚类算法自动化[14]。

1.4 蛋白质序列自动分类、识别方法概述

从以上 1.1-1.3 节可以看出：在折叠类（CLASS）层次，各个数据库对于蛋白质结构域的分类大致相同，针对结构类层次的预测也比较容易实现，可以通过

对二级结构组份的估计等手段得到。

在序列同一性较高的超家族、家族等层次，分类和预测也较为简单，序列比对方法对于这些高同一性的数据集是有效的，结合 Profile、SVM 等其它手段往往能得到非常高的准确率。

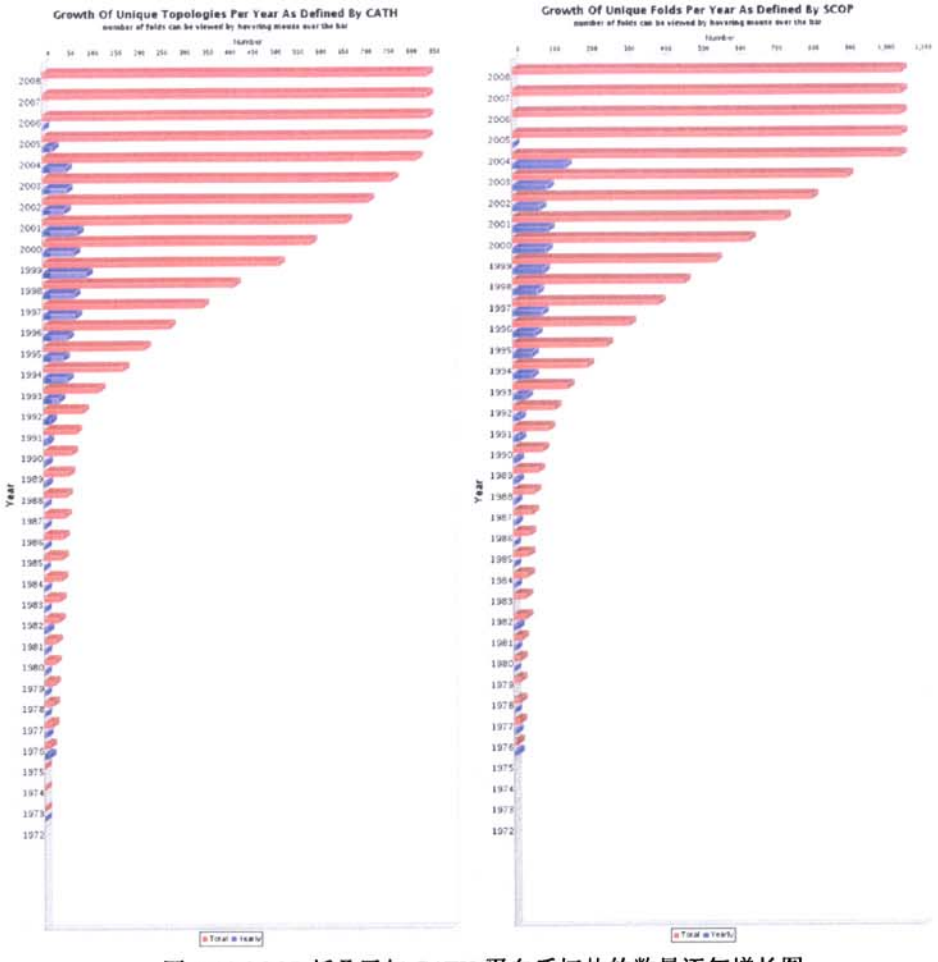


图 1-10 SCOP 折叠子与 CATH 蛋白质拓扑的数量逐年增长图

Fig.1-10 Growth of unique SCOP fold and CATH topology per year

在折叠类型层次，分类识别是一个较为复杂的问题，然而目前的数据表明，蛋白质折叠类型是有限的，从 SCOP 的折叠子或者 CATH 的 Topolgy 统计，均验证了这一点，如图 1-9。在折叠类型有限的条件下，蛋白质折叠类型预测实际上是分类预测问题，目前的应用到折叠子、拓扑层次的分类识别方法有许多：两两比较算法如 Smith-Waterman 序列比对，即通过氨基酸序列之间局部比对的打分来衡量其相似性，在两两比对基础上发展的基于多序列比对的各种 Profile 方法，人工神经网络，基于氨基酸组份、二级结构及其它参数的 SVM 分类机器等，本

章综合介绍与算法相关领域的研究现状和发展趋势。

1.4.1 序列比较

对于一条待预测序列，通过比较它和已知结构的数据库中每一条记录在序列上的相似性，来确定的两条氨基酸序列是否属于同一种折叠类型的方法均属于两两比较的算法，其中最常用的算法就是 Smith-Waterman 序列比对（序列排比、序列联配），使用这种算法的工具有 Blast、Fasta 等，当序列的同一度大于 60% 时，有理由相信它们是同一种结构，但是当序列之间的同一度较小时，识别准确度随着同一度的降低急剧降低，当序列同一度小于 30%，序列、结构、功能之间的关系非常不明显，因此序列比对的适用范围有很大的制约[17-20, 62]。

1.4.2 Profile 方法

由于序列比对等方法均是采用“一对一”的比较方法，对一组序列的共有特征不能有效利用，因此发展了“多对一”的方法：即从一个折叠类型的所有序列中提取氨基酸在位点上的分布信息，与未知序列比较。一种较为有效的方法是通过多序列比对，产生一个针对某种折叠类型的统计基序 Profile，通过一个 Profile 综合该折叠类型的氨基酸序列在所有位点上可能发生突变的概率，这种方法对识别的敏感度有较为显著的改进。

Cons	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z	Gap	Len
I	8	3	-2	5	4	5	5	-4	24	0	15	13	1	1	1	-7	2	22	21	-18	-6	4	100	100
T	13	19	-5	24	18	-18	19	7	1	7	-7	-4	14	11	10	-1	9	29	3	-28	-14	15	100	100
L	5	5	-5	3	4	13	4	2	8	-4	14	12	8	-5	0	-10	0	10	10	-1	5	2	22	22
S	17	14	17	13	10	-12	29	-5	-5	6	-14	-9	12	10	0	-2	34	19	1	-8	-15	4	100	100
T	15	3	22	0	-1	-5	12	-2	7	-3	-8	-6	5	7	-8	-7	16	29	9	-22	6	-4	100	100
T	8	-1	12	-2	0	5	6	-4	19	-4	8	5	-1	2	-8	-8	7	22	19	-15	4	-3	100	100
C	17	0	24	-1	-3	11	8	-1	7	-10	1	-2	1	-3	-8	-14	8	5	9	-5	14	-7	100	100
V	11	0	18	-1	-2	2	14	-10	26	-4	9	7	-3	7	-7	-7	21	10	31	-19	-5	-5	100	100
C	10	-8	15	-11	-11	6	8	-7	11	-10	4	3	-7	0	-11	-4	11	5	15	-22	14	-11	100	100
V	7	7	-3	8	8	-3	11	1	20	-1	14	10	4	2	8	-5	0	5	26	-24	-6	8	100	100

图 1-11 热动蛋白 70 的 Profile

Fig.1-11 the Profile of Hsp70

Profile 是对一组序列进行多重序列比对时发现的：考虑残基替代对于不同位点的特殊性，通过对多序列比对的每一列进行统计，计算每个位点上 20 种氨基酸出现的概率，求对数值就得到一个最基本的序列 Profile。如图 1-11 应用 Profile 方法鉴定序列模式的例子，使用 GCG 工具包中的 PILEUP 对来自不同有机体的

热动蛋白 70(Hsp70)进行排列，应用 GCG 的 Profile make 程序生成了序列 Profile，纵向表示位点，横向是 20 种氨基酸及插入、扩展空格打分。Profile 中的数值为对数几率，即氨基酸在目标序列中出现的概率与随机概率比值的对数值。左边一列为该位点概率最大的残基。Profile 方法建立了一组序列的正则表达式，保留了每个位点上的残基概率分布，有效提高了识别成功率，在序列相似性较高的家族层面，使用 ferredoxin 家族检验，准确率 96%，但是在同一个折叠类型中建立一个有效的序列 Profile 用于识别仍然比较困难[21]，目前 Profile 与 HMM 结合产生了 Profile HMM 方法，Profile 还可以与 SVM 方法结合，作为 SVM 的核函数使用，能有效提高 SVM 的分类能力。

1.4.3 NN 方法

NN 方法是一种在许多领域使用的成熟分类方法，基本的神经网络由输入层、隐元层、输出层等组成，每层中包含若干神经元，每层中的权重相连构成了权重矩阵。如图 1-12 所示。

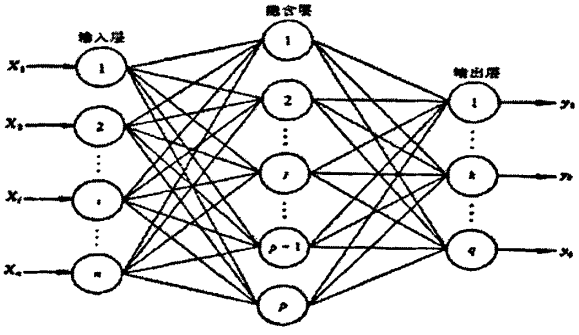


图 1-12 神经网络示意图

Fig.1-12 Schematic representation for NN

神经网络的计算结果是由训练过程中的迭代得到权重矩阵体现的，有 BP、RBF、GRNN、PNN 等种类，应用在蛋白质结构研究中，可以有效的预测蛋白质二级结构，但在蛋白质的折叠类型预测中，识别效果比 SVM 略低[22]。

1.4.4 SVM 方法

支持向量机（Support Vector Machine，SVM）是 Vapnik 等人根据统计学习理论提出的一种机器学习方法，由于其出色的学习性能，已经成为当前研究热点，

并成功应用于生物信息学：基因微阵列表达模式、转录起始点、蛋白质家族、蛋白质亚细胞定位、蛋白质折叠子识别和蛋白质四级结构等方面，在折叠类型识别方面，支持向量机算法有较好的效果[22-25]。

SVM 最大特点就是泛化能力比较强，即由有限的训练集样本得到的小误差仍能够保证对独立的测试集的小误差。另外，由于支持向量机算法是一个凸优化问题，因此局部最优解即为全局最优解，可防止过学习。这些特点是其它学习算法，如神经网络学习算法所不及的。应用支持向量机进行分类的基本思想可简述为：首先将输入空间的样本通过某种非线性函数关系映射到一个特征空间中(维数可能较高)，在此特征空间中构造最优分类超平面使两类样本(可推广到多类样本)在此特征空间中可分。映射函数仅与输入空间的低维输入向量和特征空间的内积有关，此内积可用输入空间的核函数来替代，这样就可以构造出特征空间的分类超平面，而不必清晰地描述特征空间，同时还可避免“维数灾难”，从而解决高维特征问题[22, 26-28]。

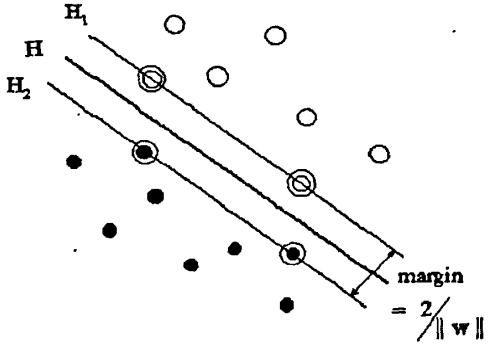


图 1-13 支持向量机示意图

Fig.1-13 Schematic representation for SVM

对于线性可分情况，支持向量机的基本思想可用图 1-13 说明。图 1-13 中，实心点和空心点代表两类样本，H 为分类线，H_1、H_2 分别为过各类中离分类线最近的样本且平行于分类线的直线，它们之间的距离叫做分类间隔(margin)。所谓最优分类线就是要求分类线不但能将两类样本正确分开，而且使分类间隔最大。推广到高维空间，最优分类线就成为最优分类面[26-28]。

在蛋白质的结构域分类方面，支持向量机及其相关研究非常多，不同的支持向量机使用的核函数不同，可以是组份、二级结构等，也可以是序列比对信息，也可以将多种属性综合考虑，不同的特征向量提取对于识别效果有非常大的影响，在折叠子层次，SVM 及其融合网络分类精度一般不大于 70%[22, 26-31]。

第 2 章 Profile HMM 折叠类型识别

2.1 引言

隐马尔可夫模型(Hidden Markov Models,HMM)是一种概率论模型，这种方法已经成功地应用于多个领域，如语音识别、光学字符识别等。HMM 在生物信息学领域中也有着重要的应用，如序列分析、基因识别等。在序列 HMM 的基础上，引入多序列比对在位点上的氨基酸概率分布，即构成 Profile HMM。因为序列比对方法对于序列同一性的要求较高，而同一折叠类型的蛋白质在序列上的差异较大，因此 Profile HMM 在蛋白质的家族、超家族识别上有较高的准确率，但在折叠类型方面，应用尚且不多[36-38]。

2.2 Profile HMM 概述

隐马尔可夫模型是马尔可夫模型的进一步发展，马尔可夫模型是马尔可夫过程的模型化，它把一个总随机过程看成一系列状态的不断转移。

考虑一个具有多个状态的系统 S，$S=\{s_1,s_2,\ldots,s_{|S|}\}$，令 S^0、S^1、…、S^t 为一系列在各个时刻系统状态的变量，即状态链。对于每个从 1 到|S|的整数，它们分别与状态链中的一个状态相联系，并且，在任何时刻，这条链都处于一种特殊的可观察状态 X_t。当且仅当对于任何 t 有 $P(S^{t+1}|S^0, S^1 \ldots S^t)=P(S^{t+1}|S^t)$，则 S 形成一条隐马尔可夫链。简单地说，就是系统未来的状态仅依赖于当前状态，通过观测 S，可以得到一条可观察序列 X，X_t 称为在时刻 t 系统链的状态。一条马尔可夫链完全决定于初始分布 $P(S_0)$和转换概率 $P_t=P(S^{t+1}|S^t)$。我们仅讨论确定模型的马尔可夫链，即状态转换概率不随时间变化的马尔可夫链。令状态转换矩阵为 $F=(f_{ij})$，f_{ij} 代表从状态 X_i 跳转到状态 X_j 的概率。

定义一个 HMM 模型是一个三元组 M=(Σ，S，Θ)，其中：

(1) Σ是字母表的集合，对于氨基酸序列，Σ是 20 种氨基酸字母的集合，即 Σ={G,A,L,M,F,W,K,S,N,D,P,V,I,C,Y,H,R,T,Q,E}；

(2) $S=\{S_1, S_2, \cdots, S_N\}$ 为状态集合，N=|S|是状态个数，对于蛋白质及 DNA，S

包含匹配、插入、删除等状态，对应于各种突变；

(3) Θ为概率集合，包括两个部分：一是状态转换概率 f_{kl}（k、l∈S），表示从状态 k 转换到状态 l 的概率；二是字符释放概率，记为 $e_k(b)$ (k∈S, b∈Σ)，表示在状态 k 下释放出字符 b 的概率。令路径 $\Pi=(\pi_1, \pi_2, \cdots, \pi_L)$ 是模型 M 的一个相继状态序列，$X=(x_1, x_2, \cdots, x_L)$ 是一个字符序列，按下述方式定义状态转换概率和字符释放概率：

$$f_{kl}=p(\pi_i=l|\pi_{i-1}=k)$$

$$e_k(b)=p(x_i=b|\pi_i=k)$$

对于给定的路径Π，可以按下面的公式计算产生序列 X 的概率：

$$P(x|\Pi)=f_{\pi_0,\pi_1}\prod_{i=1}^{L}e_{\pi_i}(x_i)\times f_{\pi_i,\pi_{i+1}}$$

这里，令 π_0 为起始状态，π_{L+1} 为终止状态。

对于氨基酸序列的隐马尔科夫模型，状态转换概率是指从序列在插入突变、删除突变、匹配等状态之间互相转换的概率，而字符释放概率则是在某个位点上 20 个氨基酸的分布概率，即位点的 Profile。

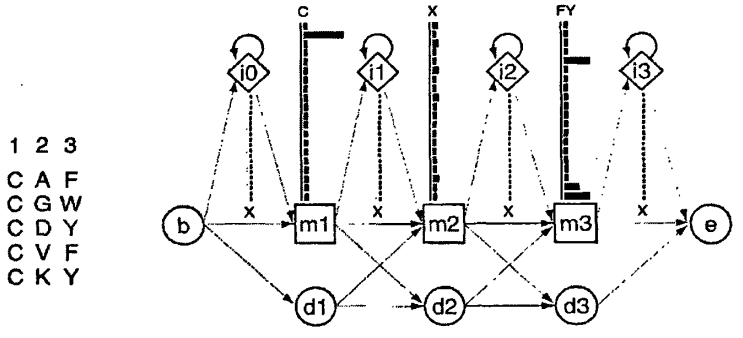


图 2-1 HMM 示意图

Fig.2-1 Schematic representation for the architecture of HMM

以 Krogh 等人于 1994 年引入的生物大分子序列 HMM 构架为例，有开始(b)、结束（e)、匹配（m)、插入（i)、删除（d）五种状态，HMM 模型从代表开始的 b（begin）状态开始，以 e（end）状态结束，如图 2-1，三个匹配态表明此模型由三列匹配位点（m1、m2、m3）组成，匹配态还包含了此位点上 20 种氨基酸的概率分布 Profile（m 态上方柱状图），即 HMM 的字符释放概率，如 m1 上残基 C 出现的概率远远大于其他残基，这是由多序列比对的统计得到的。在匹配态的前后有插入态（i0、i1、i2、i3），表明在各个位点之间可能存在插入空位的情况；而在不相邻的匹配态及 b、e 之间可以通过删除态（d1、d2、d3）连接，

这表明某些位点在实际的序列中可能缺失。各个状态之间由箭头连接，表明跳转，每个箭头都有其跳转概率。

即使一个只有少数几个位点的模型，若连续在插入态中跳转，仍然可能产生任意长度的任意序列，但不同的序列产生的概率显然是不同的，一个特定的模型总是与一类特定的序列相对应，产生该类序列的概率远大于其他序列，下面介绍模型的评估、解码、学习问题。

2.3 算法基础

给定 HMM，将其应用到实际须解决以下三个基本问题:

问题 1：已知观察序列 X=(x1,x2,...,xL)和模型 M=(Σ，S，Θ)，如何有效地计算在给定模型 M 条件下产生观察序列 X 的条件概率即 P(X|M)，即评估问题。

问题 2：即给定一个隐马尔可夫模型 M=(Σ，S，Θ)和一个观测序列 X=(x1,x2,...,xL)，在 M 中为 X 寻找一条最优路径Π^*，该路径从起始状态出发，结束于终止状态，在路径中的每一个状态都选择释放一个字符，要求使得 P(X|Π^*)最大，记为：

$$\Pi^*=\arg\max_{\Pi}\{P(X|\Pi)\}$$

即如何选择相应的在某种意义上最佳(能最好解释观测序列)的状态序列,即解码问题。

问题 3 如何调整模型 M=(Σ，S，Θ)的参数，以使条件概率 P(X|M)最大，即学习问题。

对应于 HMM 三个问题的求解，产生 HMM 在应用中的三个算法:前向-后向算法，Viterbi 算法，期望最大(EM)算法。

2.3.1 向前-向后算法

给定模型 M 和一个字符序列 X，假设产生 X 的对应路径未知，要求计算模型 M 产生 X 的概率 P(X|M)。由于一个 HMM 模型中可能的路径非常多，穷举每条路径显然是不合适的。下面介绍解决该问题的前向算法(forward algorithm)与后向算法(backward algorithm)。算法的根本任务是对于每个 $1\leqslant i\leqslant L$ 及 $k\in S$，计算概率 $P(\pi_i=k|X, M)$。给定一个序列 $X=(x_1,x_2,...,x_L)$，令 $\alpha_k(i)$为前向概率，即释放前缀$(x_1,x_2,...,x_i)$后到达状态 $\pi_i=k$ 的概率。

(1)初始化：

$$\alpha_{begin}(0)=1$$

$$\forall_{k\neq begin}\alpha_k(0)=0$$

(2)对于每个 i=0,…,L-1 及每个 l∈S，计算过程和结果为：

$$\alpha_l(i+1)=e_l(x_{i+1})\times\sum_{k\square}\alpha_k(i)\times f_{kl}$$

$$P(X|M)=\sum_{k\square}\alpha_k(L)\times f_{k,end}$$

与前向算法相对应，给定一个序列 $X=(x_1,x_2,\ldots,x_L)$，令 $\beta_k(i)$为给定状态 $\pi_i=k$ 后缀$(x_{i+1},x_{i+2},\ldots,x_L)$的产生概率。后向算法如下：

(1)初始化：

$$\beta_{end}(L+1)=1$$

$$\forall_{k\in s}\beta_k(L)=f_{k,end}$$

(2)对于每个 i=0,…,L-1 及每个 l∈S，计算过程和结果为：

$$\beta_k(i)=\sum_{l\square}f_{kl}\times e_l(x_{i+1})\times\beta_k(i+1)$$

$$P(X|M)=\sum_{l\square}f_{begin,l}\times e_l(x_1)\times\beta_l(l)$$

利用前向和后向概率，可以计算出 $P(\pi_i=k|X)$。由于 HMM 的阶数为 1，当前的状态仅依赖于前一个状态，则：

$$\begin{aligned}P(X,\pi_i=k)&=P(x_1,\ldots,x_i,\pi_i=k)\times P(x_{i+1},\ldots,x_L|x_1,\ldots,x_i,\pi_i=k)\\&=P(x_1,\ldots,x_i,\pi_i=k)\times P(x_{i+1},\ldots,x_L|\pi_i=k)\\&=\alpha_k(i)\times\beta_k(i)\end{aligned}$$

根据条件概率的定义，得到解：

$$P(\pi_i=k|X,M)=\frac{P(X,\pi_i=k)}{P(X|M)}$$

$$=\frac{\alpha_k(i)\times\beta_k(i)}{P(X|M)}$$

2.3.2 Viterbi 算法

给定一个字符序列 $X=(x_1,x_2,\ldots,x_L)$，以 $V_k(i)$代表序列前缀$(x_1,x_2,\ldots,x_i)$终止于状态 $k(k\in S,1\leqslant i\leqslant L)$的最可能路径的概率。求解过程如下：

(1)初始化：

$$V_{begin}(0)=1$$

$$\forall_{k\neq begin},V_k(0)=0$$

(2)对于每个 i=0,…,L-1 及每个 l∈S，按下式进行迭代计算：

$$V_l(i+1)=e_l(X_{i+1})\max_{k\in s}\{V_k(i)\times f_{kl}\}$$

(3)最后，计算序列 X 终止于状态“end”最可能的路径概率，即 $P(X|\Pi^*)$ 的值：

$$P(X|\Pi^*)=\max_{k\in s}\{V_k(L)\times f_{k,end}\}$$

在前向的递归计算过程中，保持向前推进的后向指针，这样，在前向计算完成后，根据后向指针重构最优路径Π^*。算法的时间复杂度为 $O(L|S|^2)$，空间复杂度为 O(L|S|)，其中，|S|代表状态集合的大小。

在实际应用中计算机在处理大量的乘法运算时会产生误差，可以使用对数值解决这个问题。因此，以 $v_k(i)$代表序列前缀$(x_1,x_2,\ldots,x_i)$终止于状态 $k(k\in S,1\leqslant i\leqslant L)$的最可能路径的对数得分值，则初值按如下方式设置：

$$V_{begin}(0)=0$$

$$\forall_{k\neq begin},V_k(0)=-\infty$$

迭代计算及最终得分计算改为：

$$V_l(i+1)=\log e_l(x_{i+1})+\max_{k\in s}\{V_k(i)+\log(f_{kl})\}$$

$$\text{Score}(X|\Pi^*)=\max_{k\in s}\{V_k(i)+\log(f_{k,end})\}$$

2.3.3 期望最大(EM)算法

学习问题是三类问题中更基本的，因为前面介绍的两类问题的算法，都是假设有一个 HMM 模型，其中的状态转换概率和字符释放概率都是已知的。然而在实际中，所知道的可能仅仅是 HMM 模型 M(A，S，Θ)产生的序列即 n 个字符序列 $X^{(1)},X^{(2)},\cdots,X^{(n)}$的集合，而不知道 M 中的各个概率值。问题是要根据给定的 n 个字符序列重构 M，使得 M 产生的这 n 个字符序列具有最大的概率。由于各个字符串是独立产生的，则：

$$P(x^{(1)},x^{(2)},\ldots,x^{(n)}|\Theta)=\prod_{i=1}^{n}P(x^{(i)}|\Theta)$$

若使用对数表示，则目标就是寻找一个Θ^*，使得：

$$\Theta^*=\arg\max_{\Theta}\{\text{Score}(x^{(1)},x^{(2)},\ldots,x^{(n)}|\Theta)\}$$

其中：

$$\text{Score}(x^{(1)},x^{(2)},\ldots,x^{(n)}|\Theta)=\sum_{i=1}^{n}\log(P(X^{(i)}|\Theta))$$

这里的 n 个字符序列 $X^{(1)}, X^{(2)},\cdots,X^{(n)}$被称为“训练序列”。期望最大(EM)算法的基本步骤为：

(1)初始化，给Θ中的参数赋予初值；

(2)计算从状态 k 到状态 l 转换的期望次数，使用与计算 $P(X,\pi_i=k)$时相同的参数，则：

$$P(\pi_i=k,\pi_{i+1}=l|X,\Theta)=\frac{\alpha_k(i)\times f_{kl}\times e_l(x_{i+1})\times\beta_l(i+1)}{P(X)}$$

对所有训练序列 $X^{(j)}(j=1,\cdots,n)$的所有位置 i(i=1,…,L(j)，L(j)为序列 $X^{(j)}$的长度)进行求和运算，按下式计算期望值 F_{kl}：

$$F_{kl}=\sum_{j=1}^{n}\frac{1}{P(X^{(j)})}\times\sum_{i=1}^{L^{(j)}}\alpha_k^{(j)}(i)\times f_{kl}\times e_l(x_{i+1}^{(j)})\times\beta_l^{(j)}(i+1)$$

其中 $\alpha_k^{(j)}(i)$是针对序列 $X^{(j)}$的前向计算结果，$\beta_k^{(j)}(i+1)$是后向计算结果。接下来计算在状态 k 释放字符 b 的期望次数：

$$E_k(b)=\sum_{j=1}^{n}\frac{1}{P(X^{(j)})}\times\sum_{\{i|x_i^{(j)}=b\}}^{L^{(j)}}\alpha_k^{(j)}(i)\times\beta_k^{(j)}(i)$$

(3)重新计算Θ的参数值 F_{kl} 和 $E_k(b)$，则关于Θ最大似然估计值为：

$$f_{kl}=\frac{F_{kl}}{\sum_{s\in S}F_{ks}}$$

$$e_k(b)=\frac{E_k(b)}{\sum_{c\in A}E_k(c)}$$

为了避免零概率，当处理数量较少的样本时，需要对 F_{kl} 和 $E_k(b)$进行修正：

$$F'_{kl}=F_{kl}+r_{kl}$$

$$E'_k(b)=E_k(b)+r_k(b)$$

r_{kl}、$r_k(b)$为拉普拉斯修正项，通常情况下为 1。

(4)反复执行步骤(2)、(3)，直到 $Score(X^{(1)},X^{(2)},\cdots,X^{(n)}|\Theta)$的增量小于给定的一个值很小的参数 ε 为止。

期望最大算法保证目标函数 $Score(X^{(1)},X^{(2)},\cdots,X^{(n)}|\Theta)$单调增加，并且概率的对数值接近于 0，保证算法收敛。需要注意，但是即使目标函数变化缓慢，Θ的参数值可能波动较大，这意味着算法所得到的结果不稳定。

本节详细内容可参阅相关文献[36-43]。

2.4 Profile HMM 在家族和超家族识别中的应用

由于 Profile HMM 方法的优越性，在家族和超家族的识别中，已经得到了较为成功的应用。具有代表性的是 SUPERFAMILY[44-47]数据库和 pfam[48-53]数据库。SUPERFAMILY 是 SCOP 数据库的一个衍生数据库，它以 SCOP 超家族分类为基

础，使用 SAM 软件包的 HMM 框架，见图 2-2，在已知的每个超家族内手工提取多个代表性序列分别作为种子（seed），在序列数据库 ASTRAL 中使用 Blast 搜索每个种子的高同源序列，将搜索得到的一系列结果进行序列比对作为一个训练集产生一个特定的隐马尔科夫模型；pfam 的方法和 SCOP 基本一致，但使用的 HMM 构架为 HMMER-plan7 见图 2-3，和 SUPERFAMILY 略有差别，使用的工具为 HMMER，有文献表明 HMMER 和 SAM 可以进行一个大致的转换，且二者的预测准确率和覆盖率均非常高，SUPERFAMILY 在 2001 年可以为 45%的细菌基因组蛋白质提供超家族预测，而当前的 pfam 可以为高达 74%的所有已知蛋白质序列提供超家族、家族预测[52-55]。随着数据的增多，超家族、家族分类会越来越完备规整，训练集会越来越丰富，可以预见到使用 HMM 预测蛋白质家族、超家族是一种很有前景的研究方法。

这两个代表性数据库中，一个超家族包含有多个种子，每一个种子对应与一个 HMM，因此每个超家族需要多个 HMM 来表示，如 Globin 超家族，在 pfam 中包含 107 个独立模型，在 SUPERFAMILY 中包含 101 个独立模型，这样极大的保证了预测的准确度，但忽略了同一个超家族中的序列之间可能存在的联系，同时存在的多个 HMM 也在一定程度上使运算资源可能被耗费。

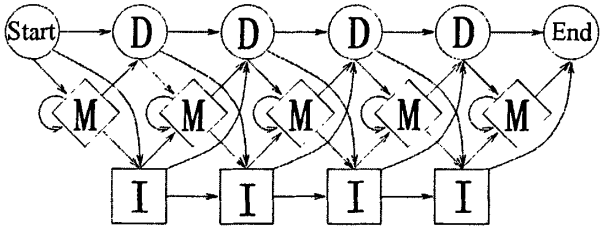


图 2-2 SAM 的 HMM 拓扑图

Fig.2-2 Schematic representation for HMM Architecture of SAM

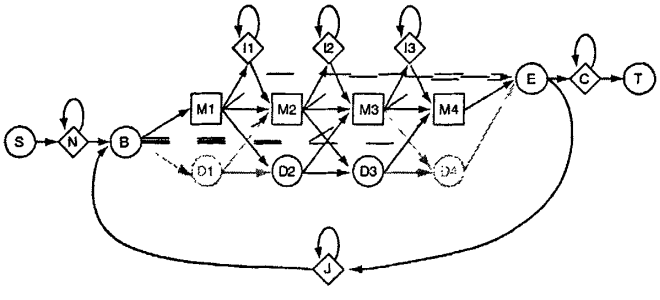


图 2-3 HMMER 的 plan-7 拓扑图

Fig.2-3 Schematic representation for Plan-7 HMM Architecture of HMMER

2.5 本章小结

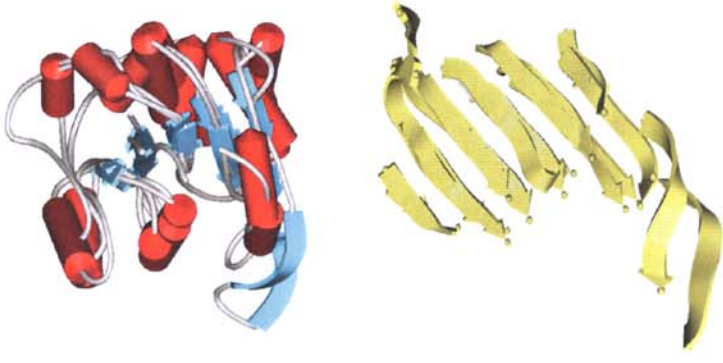
本章重点介绍了 HMM Profile 方法的算法及适用范围等相关信息。从 Profile HMM 在家族和超家族识别中的成功表明，作为一种较为成熟的方法，以氨基酸的多序列排布为基础进行序列、结构、功能关系的预测是可行的。

第 3 章　LIFCA：非冗余蛋白折叠核心分类数据库

3.1 引言

SCOP、CATH等数据库的蛋白质结构分类有不同的侧重方面，但是它们均不是以折叠识别为目的建立的，为了取得更好的识别效果，我们依据在结构相似的蛋白质中普遍存在的二级结构保守核心对现有的非冗余结构域数据库进行了重新分类。

有研究表明，蛋白质的二级结构对蛋白质立体构象的形成至关重要，蛋白质的二级结构决定了蛋白质的立体构象[56]，蛋白质二级结构的组成排布和走向与蛋白质三维结构之间有重要的联系。蛋白质折叠类型，实际上是指构成蛋白质空间形状的基本模式，包括了蛋白质分子空间结构组成的三个方面，即二级结构单元（如螺旋、折叠等）、二级结构单元的相对排布位置关系、蛋白质多肽链的整个路由关系，也即蛋白质的二级结构的连接方式[57]。在进化的变异中折叠类型很可能是二级排列中保守的部分，因此按照折叠类型进行的分类能够有效的发现某种折叠类型的所有可能的序列特征。



a）　　b）

图 3-1 d1ceqa1，d1g5qa_ and d1kjna 叠合图.　a）模型示意图 b）β 核心叠合图

Fig.3-1 structure superposition of d1ceqa1，d1g5qa_ and d1kjna. (a)carton (b) β sheet only

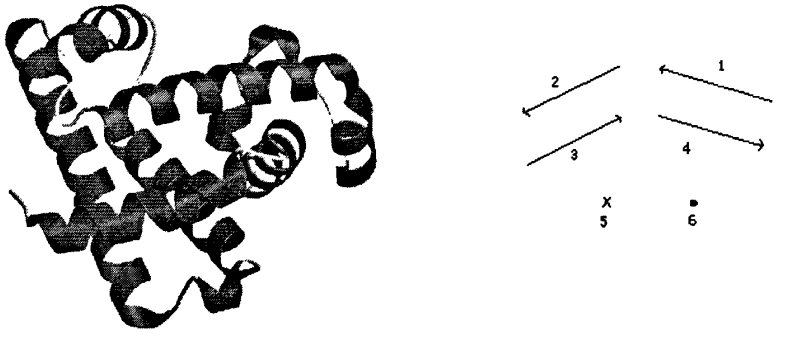
折叠类型是折叠内核的外在表现，折叠内核不包括冗余的结构片段，因此去

除外围冗余的折叠内核有着相对致密的结构。SCOP 和 CATH 分别从进化相关和结构相似角度对蛋白质进行了分类，但是都不是以折叠核心为依据，比如 Astral 编号为 d1ceqa1，d1g5qa_，d1kjna 的三个结构域在 SCOP 中分类分别属于 c.2.1.5、c.34.1.1、c.115.1.1 家族，在不同的折叠子中，但是它们的折叠核心拓扑完全一致，空间结构比较类似，如图 3-1，CATH 也存在类似的情况。

为了研究蛋白质折叠类型及其分类问题，我们认为这样的蛋白质应该归于同一个类别，本文构建了 LIFCA（Low Identical Protein Fold Cores and Annotation Database）数据库，拟从折叠核心的角度对蛋白质进行分类，希望找出同一折叠类型最本质的特征。

3.2 蛋白质折叠核心结构(Protein Fold Core Structures)

3.2.1 蛋白质的二级结构序列



a) b)

图 3-2 Oxygen Transport 蛋白（PDB代码：1a6m）的天然结构和折叠类型描述

a) 1a6m 的天然结构

b）→ 代表此折叠类型的 α-helix 单元及其空间走向；

数字序号代表 α-helix 序列的紧邻关系；

• 表示 α-helix 垂直纸面向里，× 表示 α-helix 垂直纸面向外

Fig.3-2 The structure of Oxygen Transport Protein（PDB-ID：1a6m）

a) The natural structure of 1a6m;

b) → represents a α-helix participating in fold core structure;

The number means the sequence relationship between α-helix;

• means down-towards, × means up-towards

蛋白质链可以简单化为一个由 α 螺旋、β 折叠片和无规卷曲这三种“元素”排列而成的二级结构链。其中，α 螺旋单元和 β 折叠片是规整的二级结构单元(表示为 R_i)，而无规卷曲是无规则构象和氢键模式的二级结构 (表示为 c_i)。一个蛋白质的二级结构序列可表示为：

$$C_0-R_1-c_1-R_2-c_2-R_3-\ldots -R_{n-1}-c_{n-1}-R_n-c_n,$$

其中 n 是二级结构单元数。忽略所有的无规卷曲，将规则的二级结构序列写为：

$$R_1-R_2-R_3-\ldots -R_{n-1}-R_n \ (R_i\text{是}\alpha\text{ 螺旋或}\beta\text{结构}),$$

例如，Oxygen Transport蛋白（其 PDB 代码为 1a6m （如图 3-2））的规则二级结构序列表示为：

$$\alpha_1-\alpha_2-\alpha_3-\alpha_4-\alpha_5-\alpha_6$$

3.2.2 蛋白质折叠的核心区

蛋白质折叠核心模式，简称折叠类型，反映了蛋白质核心区的拓扑模式，它包含三个要素：二级结构为基本单元，二级结构单元的序列排布，二级结构的空间走向。

在天然蛋白质 3D 结构中，相隔较远的二级结构单元可能会被分子弱相互作用(氢键、二硫桥、静电相互作用、范德华力、疏水作用、堆积力等) 聚拢到一起，形成折叠结构。蛋白质核心折叠结构，简称核心区，是由规则的α螺旋和β折叠组成，再折叠成一个三维结构。在这些二级结构中，空间相邻的氨基酸可能会形成氢健，而氨基酸间有相似的 Φ 角和 ψ 角。这些结构的形成会使每个氨基酸上的极性基团中性化。在疏水环境中，二级结构牢固地填塞在蛋白质核心区。在结构预测中，核心区被解释为在进化变异中很可能是二级排列中保守的部分，这也是我们分类的依据。

3.3 蛋白质折叠核心注释数据库的构建

3.3.1 材料与方法

数据库的原始序列、结构来自 ASTRAL1.65 数据库[53-55]，ASTRAL 是一个在

SCOP 基础上发展的数据库，如图 3-3，其中的结构域序列均有较高质量的三级结构和 SCOP 分类信息，提供的非冗余子集是按照 Blast 序列比对后序列同一性为标准进行筛选的。

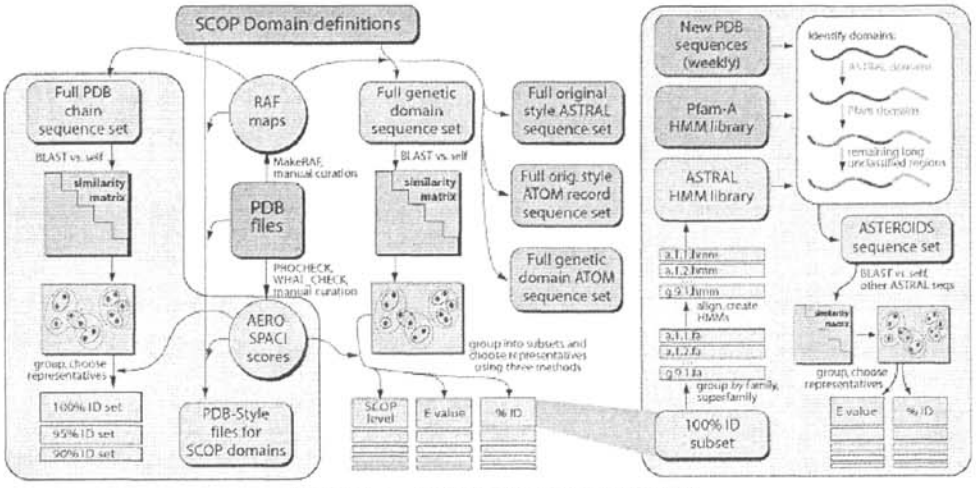


图 3-3 Astral 数据库的数据组织流程图

Fig.3-3 Data flows in ASTRAL database

构建方法如下：从 ASTRAL-1.65 数据库中选取分辨率小于 2.5Å 的、序列同一性小于 25%的非冗余子集，其中 α、β、α / β 类蛋白共 2406 个。以 ASTRAL-1.65 为基础，利用 PDB[42]数据库的资料建立蛋白质的氨基酸序列及二级结构资料数据，并利用 DSSP[43]数据库资料对 PDB 数据库中的错误信息进行纠正。

利用 PDB 数据库的原子坐标信息和图形分析软件 RasMol 分析蛋白质的空间结构特征，确定每一个蛋白的折叠类型，给出每一种折叠类型的图形化描述和空间方位信息标定文件。标定文件包含蛋白质参与此折叠类型形成的二级结构片段的空间位置取向、排布关系（如平行、垂直），二级结构片段对应的氨基酸序列信息等。最终形成了蛋白核心折叠注释数据库LIFCA (Low Identical Protein Fold Cores and Annotation Database)，目前数据库已经连接到Web，可以访问 http://bioinfo.bjut.edu.cn/LIFCA。

3.3.2 空间方位信息及标定文件

该文件主要是对每一个蛋白中参与折叠类型形成的二级结构片段进行序列位置标定和空间方位信息标定，生成相应的标定文件，如表 3-1。

其中，$ss 为二级结构单元，其中 H 代表螺旋，E 代表折叠，C 代表无规卷曲；site 为二级结构单元在蛋白质二级结构序列中的位置，3 种二级结构单元独

立排序；A 列代表二级结构单元是否参与蛋白质折叠类型框架结构形成，Y-参与，N-未参与；B 列为该二级结构单元在伸展成平面的蛋白质折叠类框架结构中的位置，分别用 0-1~9-A-B-C-D-E-F 等表示，相邻数字代表空间紧邻，其中 1 为参考片段且顺时针方向为正方向；C 为该二级结构单元相对于参考片段的取向，其中 P 代表平行、N 代表反平行、V 代表垂直；D 为折叠类型框架结构是片状、筒状，此信息只在参考片段位置标定，其中 Y 代表筒状，N 代表片状；len 为二级结构蛋白的长度（含残基数量）；start 为二级结构单元在蛋白质氨基酸序列中的起始位点；aas 为二级结构单元对应的氨基酸序列片段。

表 3-1 蛋白质折叠核心结构的空间方位信息标定文件举例

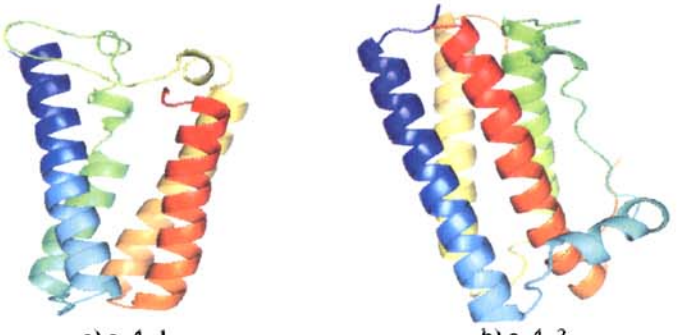
Table 3-1An example of annotation files about location & direction information of protein fold core

$ss	site	A-B-C-D	len	start	aas
"C"	1	"N-0-N-N"	4	1	"AGLS"
"H"	1	"Y-1-P-Y"	27	5	"PEEQIETRQAGYEFMGWNMGKIKANLE"
"C"	2	"N-0-N-N"	4	32	"GEYN"
"H"	2	"Y-2-N-N"	15	36	"AAQVEAAANVIAAIA"
"C"	3	"N-0-N-N"	30	51	"NSGMGALYGPGTDKNVGDVKTRVKPEFFQN"
"H"	3	"Y-3-P-N"	22	81	"MEDVGKIAREFVGAANTLAEVA"
"C"	4	"N-0-N-N"	4	103	"ATGE"
"H"	4	"Y-4-N-N"	22	107	"AEAVKTAFGDVGAACKSCHEKY"
"C"	5	"N-0-N-N"	3	129	"RAK"

3.4 分类与命名

LIFCA 对于折叠核心的命名有较为简单的原则，形如“c_6_3”，“a_4_6”等等，由两个下划线分割为三段，其中第一段可能是 a、b、c 三个字母中的任意一个，分别表示该折叠类型在折叠类层次上属于全 α、全 β 或者 α/β，第二段的数字表明该核型所包含的二级结构数目，对于全 α 类蛋白，通常是折叠核心处螺旋的数目，全 β、α/β 通常是核心处 β 片的数目，第三段的数字没有特定的意义，是在折叠类和核心片段数目一定的情况下，按照走向的不同依次分配的。例如图 3-4，a_4_1 和 a_4_3 均属于全 α 类，是两个折叠类型，并且折叠核心均由 4 个螺旋构成，只是四个螺旋的顺序连接关系不同，a_4_1 从红色的链端开始，按照逆时针方向空间排布的二级结构，在序列上依次是 1234，而 a_4_1 从链端

开始依次则是 1324，每一个 LIFCA 命名都和一定的二级结构数量、空间序列排布相对应，详细的数据可参考相关文献[32-35]。



a) a_4_1　　b) a_4_3

图 3-4 LIFCA a_4_1 与 a_4_3 示意图

Fig.3-4 Carton representations of a_4_1 and a_4_4 in LIFCA

表 3-2 代表性的列出了成员数量较多的 74 个 LIFCA 折叠类型及其分类信息。

表 3-2 部分代表性 LIFCA 折叠类型

Table 3-2Some representative LIFCA topologies

LIFCA	数量	走向及连接顺序	SCOP 分布
a_12_1	11	PVPVPVPVPVPV	a.102 一部分
a_12_2	6	PVPVPVPVPVPV	a.103 一部分
a_4_14	22	PNPV\1234	a.45
a_4_6	11	PNVV\1234	a.162.1.1 ,a.121，a.39.1 一部分
a_5_4	5	PPPPP\13245	a.26.1.3
a_5_7	6	PPPPP\15432	a.127
a_6_3	24	PVVPVV\123456	a.1
a_7_2	14	PPVPVVP\123456	a.123
b_10_1	6	PNPNPNPNPN\23496785	b.22
b_10_6	6	PPNPNPNPN\198765432	b.60.1.2
b_10_7	5	PPNPNPNPN\192745638	b.82.1.9 ,b.82.2.1 ,b.82.2.6
b_10_8	5	PNPNPNPNP\129476583	b.82.1.2
b_12_3	24	PNPNPNPPNPNP\12B4567A98C3	b.29
b_3_1	9	PNP\123	b.73,b.56,b.72,b.34.13.2
b_4_1	8	PNPN\1234	b.40.6,b.92.1.6
b_5_6	13	PPNPN\12345	b.37,b.36.1.1
b_5_7	4	PNPNP\12345	b.36.1.3 ,b.36.1.4
b_5_9	71	PNPNP\12354	特殊 b.40 去除 b.40.6.
b_6_11	10	PNPNPN\135426	b.52
b_6_14	21	PNPNPN\165432	b.55
b_6_16	5	PNPNPN\143256	b.84.1
b_6_2	5	PNPNP\14563	b.112,b.84.2.1
b_6_5	28	PNPNPN\125436	b.43,b.49

b_6_7	4	PNPNPN\165234	b.45,b.106
b_7_2	7	PNPNPNP\1567432	b.3
b_8_1	40	PNPNPNPN\12856743	b.6
b_8_14	23	PNPNPNPN\18765432	b.60.1.1 ,b.61
b_8_2	6	PNPNPNPN\16785432	b.115,b.7.1.1
b_8_3	7	PNPNPNPN\12543678	b.7.除去 b.7.1.1
b_9_1	33	PNPNPNPNP\127654389	b.1.1.1
b_9_2	23	PNPNPNPNP\127456389	b.2
b_9_3	4	PNPNPNPNP\195674238	b.113
c_3_2	4	PPP\213	c.58.1.2
c_4_10	4	PPPA\2134	c.84
c_4_13	28	PPPP\2134	c.15,c.16,c.18 等
c_4_7	12	PPPP\1423	c.45
c_4_8	4	NPPP\1234	c.52 一部分
c_5_1	4	PNPPP\12345	c.52 一部分
c_5_12	32	PPPPP\23145	c.37.1.1
c_5_13	4	PPPPN-21354	c.56.1,c.56.4
c_5_19	7	PPPPP-15423	c.46
c_5_20	7	PNPPN\12345	c.52 一部分
c_5_21	5	PPPPN\32145	c.49
c_5_22	5	PNPPP\12435	c.58.1.1
c_5_23	17	PPPPP\23415	c.37.1.12 ,c.37.1.20
c_5_25	14	PPPNP\21345	c.51
c_5_3	34	PPPPP\43251	c.3.1
c_5_31	4	NPPPP\13245	c.48
c_5_5	17	PPPPP\12345	c.14
c_5_7	7	PPPNP\32145	c.47.1.10
c_5_9	5	PPPPN\21345	c.53
c_6_13	11	PPPPPP\342561	c.37.1 一部分
c_6_14	16	PPPPPP\213465	c.36
c_6_16	8	PPPPPP\324516	c.37.1.11
c_6_2	4	PPNPPP\432156	c.55.3
c_6_20	7	PPPPNP\324156	c.60
c_6_27	4	PPPPPP\321546	c.116
c_6_29	6	NPPPPP\321456	c.62.1.1
c_6_35	5	PPPPPP\432156	c.100,c.25.1.3 ,c.25.1.5
c_6_6	4	PNPPPN\165423	c.43.1.1
c_7_13	34	PPPPPNP\3245671	c.67
c_7_14	7	PPPPPPP\2314567	c.41.1
c_7_16	6	PPPPPPP\3425617	c.37.1.19
c_7_18	6	PNPPPPP\4321567	c.2.1.7 ,c.62.1.1
c_7_22	26	PPPPPNP\3214576	特殊计算 c.66.去除 c.66.1.6 ,c.66.1.7
c_7_23	9	PNPPPPN\2341576	c.56.2,c.56.3
c_7_4	6	PPPPPPP\3214567	c.2.1.4
c_7_6	6	PPPPPPP\3241567	c.37.1.10

c_7_7	7	PPPPPNP\3214567	c.66.1.7 ,c.65.1.1 ,c.31.1.3
c_7_8	13	PPPPNPP\3214657	c.68
c_8_11	8	PPPPPPNP\43251687	c.71
c_8_6	7	PNPPPPNN\12435867	c.56.5
c_8_9	6	PPPPPPNN\32145678	c.2.1.6
c_9_1	5	PPPPPPPNP\321456789	c.72.1.1 ,c.72.1.4

3.5 LIFCA 与 SCOP 比较

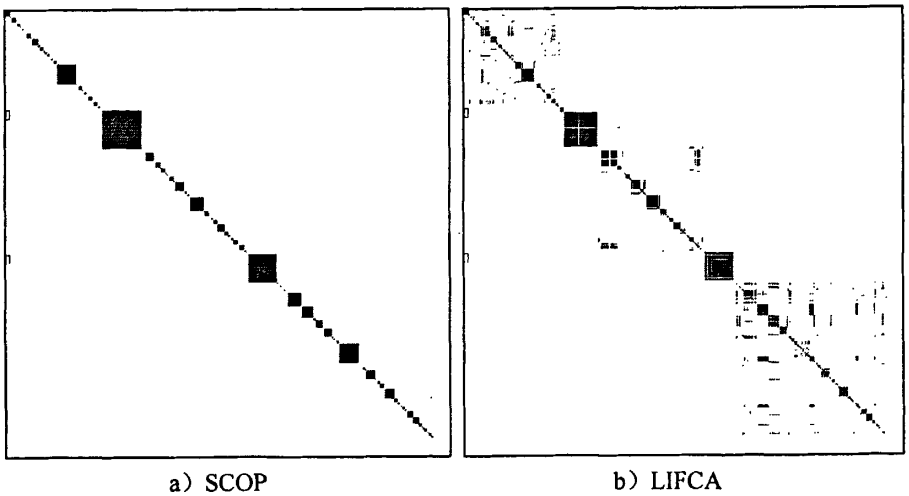
LIFCA 从折叠核心的角度对结构域非冗余数据库进行了分类，包含 2406 个蛋白质，这 2406 个蛋白质在 SCOP 折叠子的层次与 LIFCA 比较表 3-3:

表 3-3 SCOP 和 LIFCA 在折叠类中分类统计

Table 3-3 Class Statistics f SCOP and LIFCA

CLASS	α	β	α /β	Total
SCOP	84	74	86	244
LIFCA	44	70	145	259

根据分类对于某两个结构域是否同类的判断，可以绘制分类点图：横纵座标轴表示 2406 个蛋白质，按照 SCOP 家族排列，当横纵座标对应的两个蛋白质属于同一个 fold（折叠类型）时在该点画点，否则该点空白，SCOP 点图如图 3-5(a)，lifca 点图如图 3-5(b)。



a）SCOP　　b）LIFCA

图 3-5 SCOP 与 LIFCA 分类示意图

Fig.3-5 Classify representations of SCOP and LIFCA

按照统计表及点图可以看出，总体上 LIFCA 和 SCOP 的相符程度较高，但与 SCOP 相比，LIFCA 在α中 较多的表现为合并， 在α /β中较多的表现为拆分，这些差异正是分类思想不同的体现，而在全β蛋白质中，两种分类大体上是可以相互对应的。

3.6 本章小结

本章介绍了依据蛋白质折叠核心的分类数据库 LIFCA，从分类的思想上看，这种分类比 SCOP 更有利于折叠类型识别。在 LIFCA 分类的基础上，可以进行下一步的结构比对和序列建模。

第 4 章　基于结构比对的蛋白质 Profile HMM 识别

4.1 引言

Profile HMM 在家族和超家族识别中能够得到广泛应用，一个重要条件就是在这些数据集中序列的同一性较高，可以通过序列比对得到一个可靠的多序列比对结果，在此基础上可以训练 Profile HMM 模型进行预测；而在折叠类型、折叠子层次上序列的同一性一般很低，限制了 Profile HMM 的应用，在本文中，引入多结构比对方法得到序列排列，依此来训练得到适应于 LIFCA 折叠类型的 74 个 Profile HMM，并对模型的识别效果进行了统计检验，结果表明，当折叠类型内的蛋白质之间序列相似度低，可以通过引入结构比对的方法建立 HMM 有效的识别，同时也表明，在同一折叠类型内，序列之间确实存在一定的关联性。

与现在已有的一些研究相比，本文的研究工作有以下几个特点：首先，使用 LIFCA 折叠类型作为识别的依据，并作为检验结果的金标准；其次，将结构比对和 Profile HMM 方法结合，解决了低同一性序列训练 Profile HMM 的问题；最后，从预测覆盖的数据集来看，本文建立的 Profile HMM 模型库覆盖到了更大的范围，更有实用及普适性。

方法的实施主要包含结构比对、模型训练等步骤，流程如图 4-1。

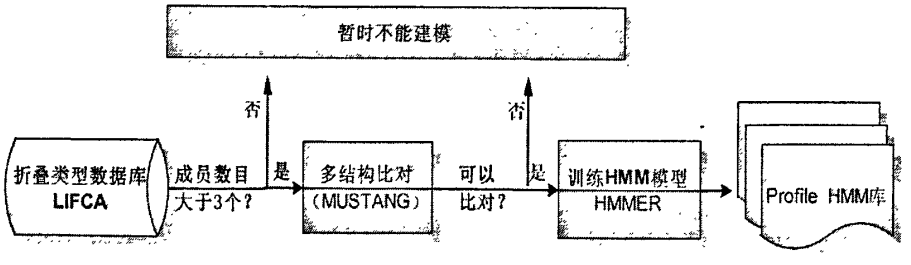


图 4-1Profile HMM 建库流程图

Fig.4-1 Flow chart of Profile HMM Library

4.2 数据集筛选

LIFCA 包含 2406 个蛋白质，覆盖了 SCOP 中 244 个折叠子，按照折叠核心分为 254 个折叠类型。由于现有的蛋白质结构数据库有极大的冗余，大部分的折叠类型只有极少数成员，而少数的折叠类型包含了大量的成员，为了建立有效的 Profile HMM 模型，使位点的统计意义更明确，成员数目小于 4 的 148 个折叠类型不予建模。剩余的 106 个折叠类型在后续的结构比对中，有 32 个折叠类型受到结构比对算法的制约不能得到比对结果，因此，最终在 α 类、β 类、α / β 类中可建模的折叠类型分别有 8、25、41 个，因此最终可以建立模型的为 74 个，包含的蛋白质数目为 893 个。

4.3 结构比对方法

从 LIFCA 中得到的的数据集，序列同一度均小于 25%，而这个区段被认为是序列比对的昏暗区，比对结果没有显著的意义，由此得到的 HMM 模型准确率会受到限制。

许多研究已经表明，结构比序列更保守，进化中即使氨基酸序列发生了巨大变化，蛋白质的三维结构仍然保守，在超家族中，蛋白质序列相似度很低却往往有类似的空间拓扑结构和二级结构走向，因此通过结构比对可以将空间上处于相似环境和位置的残基对齐。由于多结构比对技术目前仍然不成熟，在结构相似度不高的情况下难以控制，因此在本文中我们采用多结构比对算法 MUSTANG（A Multiple Structural Alignment Algorithm）对齐，MUSTANG 是 Lesk 等人在 Dali 双结构比对程序获得成功的基础上于 2006 年发展的一种多结构比对方法，对于空间拓扑、残基接触模式有较强的识别能力，双结构比对结果与 DALI 相当，多结构比对结果与其他一些现有的工具相比相当或者更好[59-61]。对于给定的一组结构，用 MUSTANG 可以得到结构叠合文件及结构比对基础上的序列比对，如 LIFCA 的 b_8_3 折叠类型包含 7 个成员，比对结果如图 4-2、4-3 所示。

但是，结构比对算法的使用范围也是有限制的，对于一些结构差异较大的折叠类型，MUSTANG 仍然不能给出比对结果，在 148 个成员数目大于 4 的折叠类型中，有 32 个折叠类型使用 MUSTANG 算法不能进行比对，这样的折叠类型，超出了可以用结构比对结合 Profile HMM 进行预测的范围，在本文中，将这些折叠类型归入暂时不能建模的数据集中。

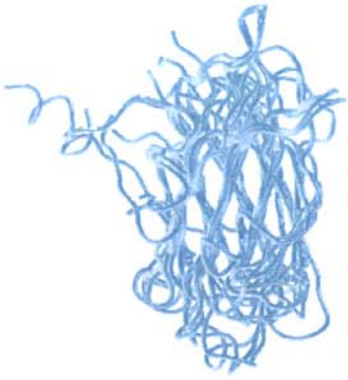


图 4-2 b_8_3 叠合图

Fig.4-2 structure superposition of b_8_3

```
d1dcea2.ent     1  -------HD-------V---------LCCVHVSRE--EA-----CLSVCF-SRP---------L   24
d1dsya_.ent     1  ---TEK-R-------G---------RIYLKAEVT--DE-----KLHVTV-RDAKNLIPMDPNG   35
d1l4ia2.ent     1  ----------VIPP-EQA--------PGKLEFTRE----------LTLFN-PTP---------Y   25
d1p5va2.ent     1  --------KGTP-IQF--------AENLSWKVD--GG-----KLIAEN-PSP---------F   28
d1qpxa2.ent     1  ------------NEVW--------QDQLILNKV--SG-----GYRIEN-PTP---------Y   25
d1rsy__.ent     1  GG---------GI-LDSMVEKLGK-LQYSLDYDFQNN-----QLLVGI-IQAAELPALDMGG   45
d1who__.ent     1  --V-----------------------P-KVTFTVE--EGSNEKHLAVLVKYE----------   24
                                                                1

d1dcea2.ent    25  TVGSRMGTLLLMVD--------EA--------PLSVEWRTPDGRNRPS-----HVWLCDLP--   64
d1dsya_.ent    36  LS---DPYVKLKLI--------PDPKNESKQKTKTIRS---------TLNPQWNESFTFK---   75
d1l4ia2.ent    26  YL---TV-TDLKAG--------N--------KSLENTMVP---------PQGKVTVN---   53
d1p5va2.ent    29  YM---NI-GELTFG--------GK--------SIPSHYIP---------PKSTWAFD---   56
d1qpxa2.ent    26  YV---TV-IGLGGS---EKQAEE--------GEFETVMLS---------PRSEQTVK---   58
d1rsy__.ent    46  TS---DPYVKVFLL--------PD---KKKKFETKVHKKT---------LNPVFNEQFTFK---   83
d1who__.ent    25  GD---TM-AEVELREH------GS-----D-EWVAMTK-G---------EGGVWTFDSE   57

d1dcea2.ent    65  AA------SLNDQLPQHTFRVIWTG----------SDS-QKECV---LL---------KDRP   97
d1dsya_.ent    76  LK-PSDK--------DRRLSVEIWDWDRTTR---NDFMG-SLSFG---VSELM-KMPA---SG  118
d1l4ia2.ent    54  I-------------GGDITYKTI---NDY-----GA-LTEQVRGV---------V----------   77
d1p5va2.ent    57  LP--------------NVSWRIIN------DQGGLDR-LYSKN---V-----------------   79
d1qpxa2.ent    59  SAN-------------YNTPYLSYI---NDY-----GG-KP-VLSFI---------C-----N---GS   86
d1rsy__.ent    84  VPYSELG--------GKTLVMAVYDFDRFSK---HDIIG-EFKVP-----MNTVDFGHV---TE  127
d1who__.ent    58  EPL-------------QGPFNFRFL---TEK-----GM-KN-VF-DDVVPEKY--------T---IG   89

d1dcea2.ent    98  ECWC-RDSATDEQ----------  109
d1dsya_.ent   119  WYKL-LN-QEEGEYYNVPIPE  137
d1l4ia2.ent        ----------------------
d1p5va2.ent        ----------------------
d1qpxa2.ent    87  RCSVE-----------------   91
d1rsy__.ent   128  EWRDLQS-A-------------  135
d1who__.ent    90  ATYAP-----------------   94
```

图 4-3 b_8_3 基于结构的序列比对

Fig.4-3 Sstructure based sequence alignment of b_8_3

4.4 训练 Profile HMM

通过第三章的论述，可以看到一个超家族的隐马尔科夫拓扑包含一下几个要

素：M 状态的数目，也就是超家族 HMM 的长度，一般和超家族成员的大致氨基酸序列长度相当；不同状态之间的跳转概率，对应与示意图中的每个箭头，表明从一个状态跳到另一个状态的可能性；每个 M 态及 I 态，都包含二十种氨基酸出现的概率，这些概率实质上是一个 Profile。

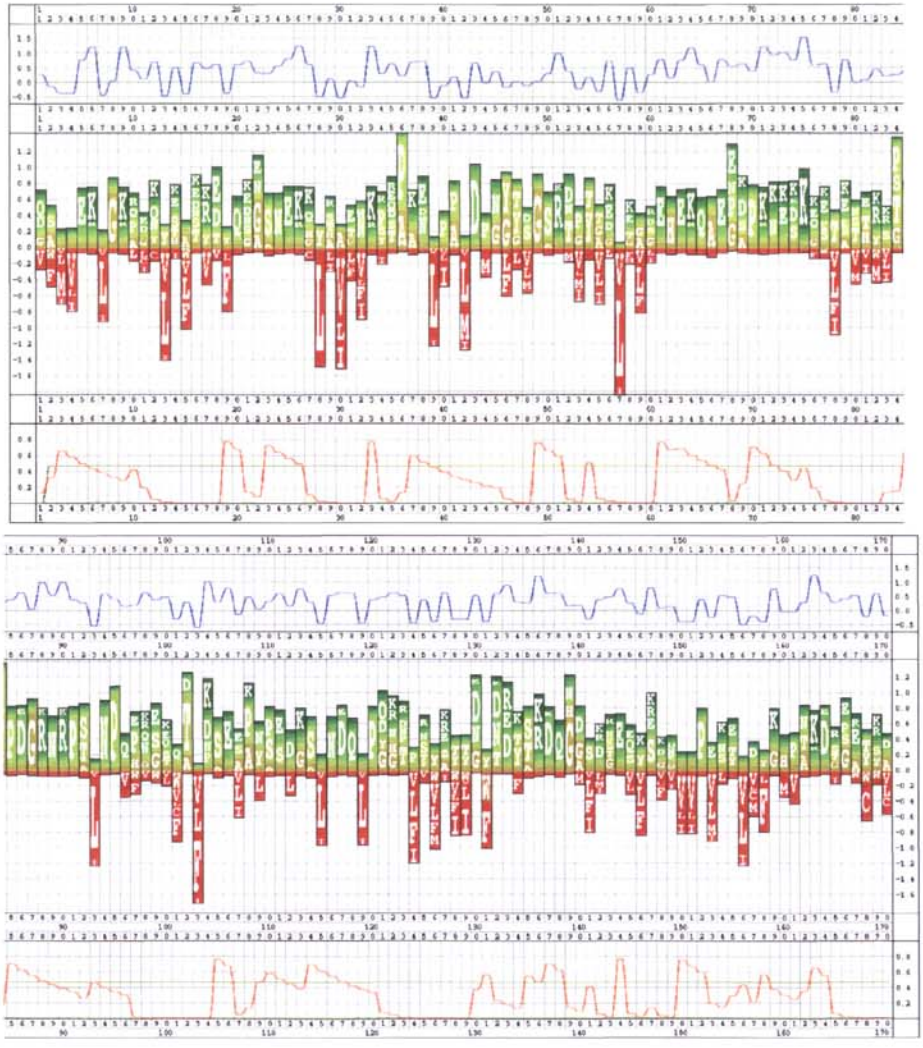


图 4-3 b_8_3 Profile HMM 模型图示

Fig.4-3 A graphical representation of the b_8_3 Profile HMM model

模型对所有位点的所有状态都用概率来描述，非常灵活，它包含了所有可能的氨基酸序列，但是决定模型三个要素的参数都是通过训练得到的。因此对于一个 HMM 模型，决定性作用的事实上是训练的样本。一个好的模型，应该有较为明显的保守位点，较为适合的匹配态数量，如图 4-4 中 b_8_3 的模型示意图[46]，顶部蓝色曲线是疏水值的数学期望；中间的柱状图表明了位点处于 M 态时出现

20 种氨基酸的概率，单字母缩写按照亲疏水从上到下排列，只有概率较大的氨基酸才被显示，字母的大小与概率成正比，整个柱体的高度是此位点的氨基酸匹配状态和随机分布的差异，即保守性；图像底部深红、浅绿两条曲线分别表示插入和删除残基的概率，墨绿色曲线是连续插入空位的概率。保守位点如 28Leu、36Pro 等均在 HMM 中得到体现。

对于选定的 74 个折叠类型，用 HMMER 工具包输入经过整理的结构比对结果进行训练，将所有的模型综合在一起进行归一化处理，得到了 Profile HMM 库，以便于进行折叠类型的分类预测。

4.5 结果检验

为了检验 HMM 库的识别效果，选用 Astral1.65 结构域非冗余数据库作为检验集，检验集包含结构域序列 19095 条，检验分为单模型识别效果检验和全库识别效果检验。

4.5.1 单模型识别效果检验

单模型检验是一个两分类问题，识别的结果分为“是”和“否”两种，可以按照以下方法检验，定义统计评估量：t_p为该类正确预测为真的总数，f_p为其它类错误预测为该类的总数，t_n为其他类正确预测为非该类的总数，f_n为该类错误预测为其他类的总数，则：

识别特异性指标：

$$S_p = \frac{t_n}{t_n + f_p} \times 100\%$$

识别敏感性指标：

$$S_n = \frac{t_p}{t_p + f_n} \times 100\%$$

Matthew 相关系数（MCC）：

$$MCC = \frac{(t_p \times t_n) - (f_p \times f_n)}{\sqrt{(t_p + f_n) \times (t_n + f_p) \times (t_p + f_p) \times (t_n + f_n)}}$$

为了检验 Profile HMM 的识别效果，使用单个模型在序列库中进行搜索，使用默认的显著性阈值参数 E<10，结果见表 4-1，其中 b_9_2 和 b_5_9 的敏感性为 0，原因是训练集数目偏多，序列、结构差异性较大，在结构比对中加入了大量

的空位，HMM 训练中匹配态数目与序列的长度明显不符，可以认为是建模失败。

表 4-1 单模型统计检验

Table 4-1 statistical test of single HMM model

LIFCA	tp	fp	tn	fn	MCC	Sp(%)	Sn(%)
a_4_14	77	2	19016	0	0.99	99.99	100
a_4_6	46	15	19034	0	0.87	99.92	100
a_5_4	29	0	19065	1	0.98	100	96.67
a_5_7	44	1	19050	0	0.99	99.99	100
a_6_3	227	0	18859	9	0.98	100	96.19
a_7_2	91	0	19004	0	1.00	100	100
b_10_1	19	0	19076	0	1.00	100	100
b_10_6	34	1	19060	0	0.99	99.99	100
b_10_7	17	0	19078	0	1.00	100	100
b_10_8	31	0	19062	2	0.97	100	93.94
b_3_1	16	1	19068	10	0.76	99.99	61.54
b_4_1	17	2	19073	3	0.87	99.99	85
b_5_6	42	5	19048	0	0.95	99.97	100
b_5_7	7	19	19069	0	0.52	99.9	100
b_6_11	20	2	19069	4	0.87	99.99	83.33
b_6_14	52	0	19031	12	0.90	100	81.25
b_6_16	14	0	19081	0	1.00	100	100
b_6_2	8	2	19085	0	0.89	99.99	100
b_6_7	5	0	19083	7	0.65	100	41.67
b_8_1	153	25	18917	0	0.93	99.87	100
b_8_14	125	1	18964	5	0.98	99.99	96.15
b_8_2	15	10	19069	1	0.75	99.95	93.75
b_9_1	670	63	18362	0	0.95	99.66	100
b_9_3	5	0	19090	0	1.00	100	100
c_4_10	6	2	19082	5	0.64	99.99	54.55
c_4_7	51	0	19043	1	0.99	100	98.08
c_4_8	15	0	19080	0	1.00	100	100
c_5_1	8	0	19087	0	1.00	100	100
c_5_12	53	113	18921	8	0.52	99.41	86.89
c_5_13	7	0	19088	0	1.00	100	100
c_5_19	12	0	19083	0	1.00	100	100
c_5_20	21	0	19074	0	1.00	100	100
c_5_21	15	0	19080	0	1.00	100	100
c_5_22	20	0	19075	0	1.00	100	100
c_5_23	54	65	18969	7	0.63	99.66	88.52
c_5_25	25	1	19069	0	0.98	99.99	100
c_5_5	20	27	19034	14	0.50	99.86	58.82
c_5_7	25	0	19055	15	0.79	100	62.5
c_5_9	18	1	19076	0	0.97	99.99	100
c_6_20	24	3	19068	0	0.94	99.98	100
c_6_27	5	0	19084	6	0.67	100	45.45
c_6_29	36	0	19059	0	1.00	100	100
c_7_13	117	1	18977	0	1.00	99.99	100
c_7_14	50	3	19042	0	0.97	99.98	100
c_7_23	24	0	19071	0	1.00	100	100
c_7_6	33	12	19038	12	0.73	99.94	73.33

c_7_8	59	0	19024	12	0.91	100	83.1
c_8_11	33	3	19059	0	0.96	99.98	100
c_8_6	23	0	19070	2	0.96	100	92
c_8_9	15	17	19057	6	0.58	99.91	71.43
c_9_1	11	1	19083	0	0.96	99.99	100
b_12_3	99	0	18925	71	0.76	100	58.24
b_6_5	70	0	19000	25	0.86	100	73.68
b_7_2	64	0	19024	7	0.95	100	90.14
b_9_2	0	0	19005	90	—	100	0
c_3_2	3	4	19088	0	0.65	99.98	100
c_5_31	9	0	19086	0	1.00	100	100
c_6_14	32	0	19056	7	0.91	100	82.05
c_6_35	7	4	19084	0	0.80	99.98	100
c_4_13	76	0	19019	0	1.00	100	100
c_6_13	17	11	19067	0	0.78	99.94	100
c_6_16	23	6	19042	24	0.62	99.97	48.94
c_6_2	11	2	18976	106	0.28	99.99	9.4
c_7_16	13	17	19049	16	0.44	99.91	44.83
c_7_18	19	0	19058	18	0.72	100	51.35
c_7_4	15	7	19072	1	0.80	99.96	93.75
c_7_7	13	5	19070	7	0.68	99.97	65
c_5_3	52	88	18930	25	0.50	99.54	67.53
c_6_6	16	1	19078	0	0.97	99.99	100
a_12_	62	0	19013	20	0.87	100	75.61
c_7_22	65	23	18998	9	0.80	99.88	87.84
b_8_3	21	0	19074	0	1.00	100	100
b_5_9	0	0	18824	271	—	100	0
平均					0.86	99.96	84.83

4.5.2 全库识别效果检验

将 74 个模型放在一个 HMM 库中，将一条未知折叠类型的序列提交到 HMM 数据库中比较，寻找最为匹配的模型，即序列的折叠类型预测。若将检验集中的序列逐条提交，统计结果，可以检验 HMM 库的优劣。4.5.1 节中单模型检验是在序列数据库中搜索与模型匹配的结果，是筛选序列，而全库检验则是在 HMM 库中搜素与序列匹配的最佳模型，是折叠类型识别问题。

在检验集中，有 4083 个序列在 Profile HMM 库中有对应分类，剩余 15012 条序列无对应的模型。这是因为数据库的高冗余，及结构比对算法在适用范围上的限制。

对于 4083 条在 Profile HMM 中有分类的检验集，定义：

$$N=n_1+n_2+\cdots n_k$$

$$C=c_1+c_2+\cdots c_k$$

其中 N=4083， k=74; n_1 表示检验集中第 1 类的个数，n_2、n_3 依此类推；c_1 表示

第 1 类中被模型 1 唯一的识别出来的序列个数，c_2、c_3 依此类推。则对于单个模型定义识别精度 Q_i：

$$Q_i = \frac{c_i}{n_i}$$

对于整个模型库，各个类别在库中所占比重不同，则分类精度为单模型识别精度的权重加和：

$$Q = \sum_{i=1}^{k} W_i Q_i$$

$$W_i = \frac{n_i}{N}$$

最终得到：

$$Q = \frac{C}{N}$$

即模型库的分类精度等于分类正确的数量除以检验集中序列的总数[22]。

检验结果中，对于模型库有覆盖的 4083 条序列，3043 个被正确分类，88 个不能分类，52 个被错误分类；对于模型库未能覆盖到的 15012 条序列，预测为库中无有效匹配的 14821 个，给出匹配结果的有 91 个，则个检验量的值如表 4-2。

表 4-2 Profile HMM 模型库统计检验

Table 4-2The statistical test of Profile HMM Library

统计量	S_p(%)	S_n(%)	MCC	Q(%)
数值	82.97	75.49	0.82	74.53

4.5.3 结果比较

Ding(2001)等人使用 SVM 及 NN 方法，将氨基酸组份、极性、疏水性、二级结构作为特征向量，在 27 个折叠子之间进行分类，得到的分类精度 Q 约为 56%，相同的训练集、试验集情况下，NN 的表现略差[22]。

Chinnasamy A 等人 2005 年使用贝叶斯分类器对于上述数据集进行分类，对折叠子的分类结果，分类精度精度 Q 为 58%[64]。

Yu Chen 等人 2006 年使用自己的折叠类型分类标准，使用分类树的策略，以序列两两比较为依据，进行折叠子识别研究，Top5 准确率最高 45%，Top1 准确率约 30%[65]。

施建宇、潘泉等人 2006 年使用支持向量机融合网络(SFN)识别折叠子，分类精度 Q 的实际结果为 61.04%，理论上界为 69.35%[27]。

在数据覆盖度方面，Ding 及 Chinnasamy A、施建宇的工作涉及到 27 个 SCOP

折叠子，训练集 313 条，测试集 385 条，Yu Chen 等人的工作首先建立识别分类树，然后做 leave one out 检验，涉及到的序列总数为 379。在本文中，训练集使用了 893 个结构域，测试集 4083 个，涉及到 LIFCA 折叠类型 74 个，SCOP 折叠子 74 个（参见表 3-3）。

综合以上结果，可以认为本文使用的 LIFCA 分类方法及基于结构的 Profile HMM 分类精度的提高是显著的。

4.6 识别方法的软件实现

以本章建立的 HMM 数据库及 HMMER 工具包为基础，设计实现蛋白质折叠类型预测软件，将相关功能集成到 EASY BLAST 软件中，软件还包含 BLAST 等其他模块，可以通过 http://bioinfo.bjut.edu.cn/HMMER 网址下载该软件，本节介绍和蛋白质折叠类型识别有关的模块。

4.6.1 软件功能

EASY BLAST 的 HMM 模块实现的功能是，当用户输入 fasta 格式的氨基酸序列（可以是单条或者多条），软件调用 HMMER 工具在本文建立的 74 中折叠类型模型中寻找匹配，如果有匹配，还应给出匹配的统计显著性打分，序列和模型的序列比对等相关信息。

4.6.2 开发工具与环境

该软件算法实现主要由 Microsoft Visual C# .NET 语言编写，在 Windows 平台下编写，

开发工具—Microsoft Visual C# .NET　69514-335-0000007-18090

运行环境—Microsoft Windows XP；

硬件要求—CPU 要求 Pentium III 以上(建议 P4 2.0G)；内存 256M 以上（建议 512M）

4.6.3 界面设计

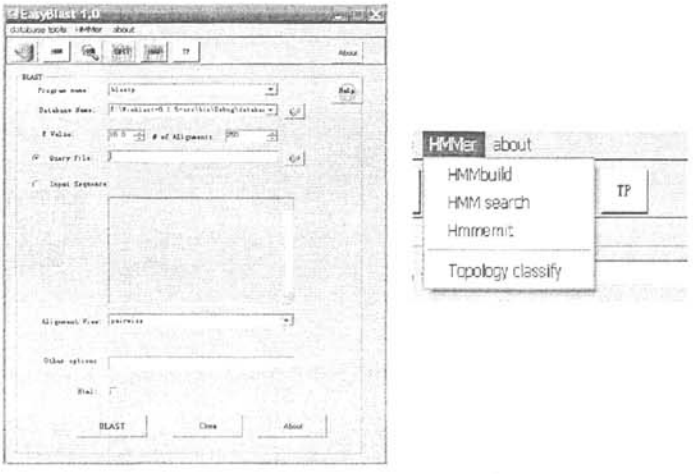


图 4-4 Easy BLAST 界面示意图

Fig.4-4 the main interface of Easy BLAST

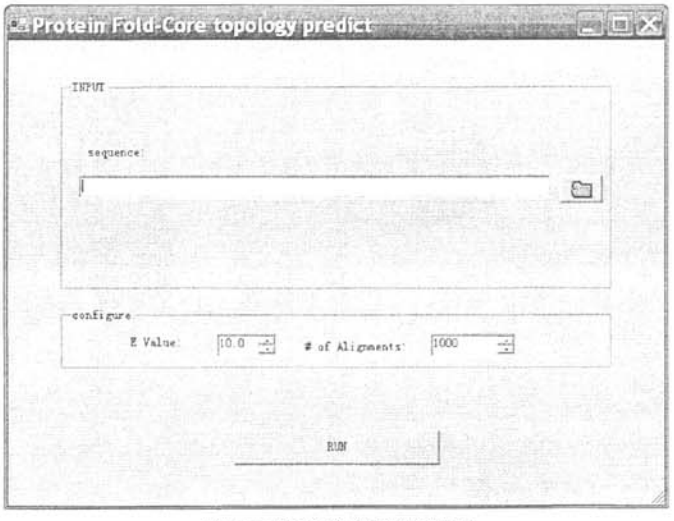


图 4-5 折叠类型预测界面

Fig.4-5The Main interface of topology recognize

程序界面如图 4-4 所示，点击 HMMER 菜单的 Topology Classify，或者点击工具栏的 TP 按钮，打开折叠类型识别对话框，如图 4-5，在 Sequence 文本框中输入要预测的蛋白质序列文件路径及名称，或者点击右侧文件夹图标使用“打开文件对话框”定位文件路径，然后在 Configure 中设定参数，默认的 E 上限阈值为 10，输出结果的最大条数为 1000，可以根据需要做相应修改，点击 RUN 按钮进行搜索，输出的结果为文本格式，一个典型的匹配的如图 4-6 所示。

```
hmmpfam - search one or more sequences against HMM database
HMMER 2.3.2 (Oct 2003)
Copyright (C) 1992-2003 HHMI/Washington University School of Medicine
Freely distributed under the GNU General Public License (GPL)
--------------------------------------
HMM file:                 E:\Winblast-0.1.5-src\bin\Debug\bin\data\LIFCA.hmm
Sequence file:            C:\res.fa
--------------------------------

Query sequence: d1jzka_
Accession:      [none]
Description:    a.1.1.2 (A:) Hemoglobin I (Ark clam (Scapharca inaequivalvis))

Scores for sequence family classification (score includes all domains):
Model    Description                                        Score    E-value  N
-------- -----------                                        -----    ------- ---
a_6_3                                                       138.8    1.2e-040   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -------- ----- -------    -----  -------
a_6_3      1/1       9   145 ..     1   170 []   138.8  1.2e-040

Alignments of top-scoring domains:
a_6_3: domain 1 of 1, from 9 to 145: score 138.8, E = 1.2e-040
                     *->LseLdaikgivktaaqkdlvkkswakiaanaeaigkealirfF ayP
                        L                 ta+ k+++++sw++t+++++++g+ ++   fF  ++
      d1jzka_      9  L----------TADVKKDLRDSWKVIGSDKKGNGVALMTTLFaDNQ 44

                     etkayFskfkgkseaeGqlasspDkvaaharkvinalreavdnldCPAGd
                     et+ yF++++++s      q+++++ k+++h++++++al+++++d+ld     +
      d1jzka_     45 ETIGYFKRLGNVS----QGMAND-KLRGHSITLMYALQNFIDQLD----N 85

                     tpgdlearllttLaekKHkdsdvlgVppenfelskalkevlashlgkef
                      p+dl++ ++++++a++ H   +   +++++++f++++   +k+vlas++     f
      d1jzka_     86 -PDDLVC-VVEKFAVN-HI-T--RKISAAEFGKFNGPIKKVLASKN---F 126

                     tpaakaAwakafdvvaaaLsskyh<-*
                     ++++++Awak+++vv+aaL
      d1jzka_    127 GDKYANAWAKLVAVVQAAL-----      145

//
```

图 4-6 有匹配 HMM 时的程序输出

Fig.4-6 Output of a match Profile HMM search

```
hmmpfam - search one or more sequences against HMM database
HMMER 2.3.2 (Oct 2003)
Copyright (C) 1992-2003 HHMI/Washington University School of Medicine
Freely distributed under the GNU General Public License (GPL)
--------------------------------------
HMM file:                 E:\Winblast-0.1.5-src\bin\Debug\bin\data\LIFCA.hmm
Sequence file:            C:\no.txt
--------------------------------

Query sequence: d1dlwa
Accession:      [none]
Description:    a.1.1.1 (A:) Protozoan/bacterial hemoglobin -> {Ciliate -> (Paramecium
  -> caudatum)}

Scores for sequence family classification (score includes all domains):
Model    Description                                        Score    E-value  N
-------- -----------                                        -----    ------- ---
  -> [no hits above thresholds]

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -------- ----- -------    -----  -------
  -> [no hits above thresholds]

Alignments of top-scoring domains:
  -> [no hits above thresholds]
//
```

图 4-7 无匹配 HMM 时的程序输出

Fig.4-7 Output without of match Profile HMM searching

其中顶部是相关的说明信息，如 HMMER 的版本等，中间的“Query sequence”

段是输入的序列的信息，在“Model Description”项中包含了搜索结果中匹配的模型名称及其打分，如图 4-6 匹配的是 a_6_3 折叠类型，在“Alignments of top-scoring domains:”段中列出序列和模型的比对情况，标出了匹配显著的残基。

当未找到合适的匹配结果时，在“Model Description”项中指出：“[no hits above thresholds]，如图 4-7。

4.7 本章小结

针对 LIFCA 的折叠分类，本章就折叠类型的预测问题做了相关研究，首先将符合条件的数据集进行结构比对，得到基于结构的序列比对结果，然后建立了包含 74 个模型的折叠类型 HMM 库，对识别效果进行了统计检验，最后为了方便使用和管理，将预测过程集成到软件中，介绍了自动化的基本情况。

结论与展望

蛋白质折叠和蛋白质结构-功能关系的研究属于后基因组时代生命科学的核心问题，具有重要的科学意义，本文对于蛋白质结构域折叠核心的分类识别做了相关研究，主要内容包含以下几点：

1. 分类问题：在蛋白质结构域数据库中挑选代表性结构域，按照二级结构组成、序列连接、空间走向重新分类，建成折叠类型分呢类数据库，数据覆盖了全α、全β、α/β三个折叠类。对于数据库中结构域的序列信息、二级结构信息、走向信息进行了详细的整理注释，以此作为识别和预测的基础。
2. 结构比对问题：在分类的基础上，使用结构比对算法 MUSTANG，在符合数据条件的折叠类型内部进行了结构比对，得到基于结构信息的残基序列比对，与一般的 Profile HMM 方法相比，对于结构信息的引入能够更有效的提取共性，提供低相似度序列之间的位点保守信息。
3. Profile HMM 库：使用 HMMER 工具，为 74 个折叠类型建立了统一的序列模型，数据的覆盖度有较大的提高。对模型的识别结果进行检验表明，单模型的识别敏感度 84.83%，特异度 99.96%，全 Profile HMM 库的分类敏感度 75.49%，特异度 82.97%，识别精度 74.53%，均属于高水平。
4. 识别方法的自动化：编写程序使得方法能够较为方便的使用。

蛋白质的序列结构关系是非常复杂、困难的研究课题。本文的工作取得了初步的结果，可以从以下几点做进一步深入研究：

1. 由于结构比对方法的局限性，对于某些数据集，结构比对仍然不能有效的发现其中的共性，比对结果不显著，造成不能建模的问题，可以尝试对于同一个折叠类型建立多个模型进行识别，继续提高方法的普适性。
2. 序列的 Profile HMM，只能给出经过训练后折叠类型在残基分布上的概率特征，但是对于折叠概率中包含的物化信息没有进行有效的挖掘，可以针对这方面做更深层的研究。

参考文献

1 阎隆飞，孙之荣，蛋白质分子结构. 清华大学出版社（1999）.

2 陶慰孙，李惟，姜涌明，蛋白质分子基础. 高等教育出版社（1995）.

3 赵南明，周海梦，生物物理学. 高等教育出版社（2006）

4 Burley S K. An overview of structural genomics. Nature Struct Biol. 2000,7 (suppl): 932-934

5 David Baker, Andrej Sali. Protein Structure Prediction and Structural Genomics. Science. 2001, 294 (5540): 93-96

6 Murzin A G, Brenner S E, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 1995, (247): 536-540

7 Lo Conte L, Brenner S E, Hubbard T, Chothia C. SCOP database in 2002: refinements accommodate structural genomics. Nucl. Acid Res. 2002, 30(1): 264-267

8 Orengo C A, Michie A D, Jones S, Jones D T, Swindells M B. CATH- A Hierarchic Classification of Protein Domain Structures. J.M. Structure. 1997, 5(8): 1093-1108

9 Harrison A, Pearl F, Mott R, Thornton J, Orengo C A. Quantifying the similarities within fold space. Journal of Molecular Biology. 2002, 323(5): 909-926

10 Harrison A, Pearl F, Sillitoe I, Slidel T, Mott R, Thornton J, Orengo C. Recognizing the fold of a protein structure. Bioinformatics. 2003, 19(14): 1748-1759

11 Pearl FM, Bennett CF, Bray JE, Harrison AP, Martin N, Shepherd A, Sillitoe I, Thornton J, Orengo CA.The CATH database: an extended protein family resource for structural and functional genomics.Nucleic Acids Research. 2003, 31(1): 452-455

12 Pearl F, Todd A, Sillitoe I, et al. The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Research. 2005, 33(Database Issue): 247-25

13 Sillitoe I, Dibley M, Bray J, Addou S. and Orengo C.Assessing strategies for improved superfamily recognition. Protein Science. 2005, 14(7): 1800-1810

14 Holm L, Sander C. Mapping the protein universe. Science. 1996, 273: 595-603.

15 Chothia C. One thousand families for the molecular biologist. Nature. 1992, 357(6379): 543-544

16 David Baker. A surprising simplicity to protein folding. Nature. 2000, 405(6782): 39~42

17 Kristin K Koretke, Robert B Russell. Fold Recognition from Sequence Comparisons. PROTEINS: Structure, Function, and Genetics. 2001, 5(Suppl):68–75

18 SF Altschd, W Gish. A basic local alignment search tool. Journal of Molecular Biology,1990,215:403-410

19 SF Altschul, TL Madden, AA Schaffer. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997, 25(17): 3389-3402

20 D Eisenberg. Into the black of night. Nat. Struct. Biol. 1997, 4: 95-97

21 David W Mount. Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor

Laboratory Press. 2001: 241-279.
22 Chris H Q. Ding, Inna Dubchak. Multi-class protein folds recognition using support vector machines and neural networks. Bioinformatics. 2001, 17(4): 349-358
23 C Z CAI, L Y Han, Z L Ji, X Chen, and Y Z Chen. SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res. 2003, 31(13): 3692-3697
24 F Liang. An Effective Bayesian Neural Network Classifier with a Comparison Study to Support Vector Machine. Neural Comput. 2003, 15(8): 1959-1989
25 M T A Shamim, M Anwaruddin, H A Nagarajaram. Support Vector Machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs. Bioinformatics. 2007, 23(24): 3320-3327
26 张绍武, 潘泉, 陈润生, 张洪才.基于支持向量机的蛋白质同源寡聚体分类研究. 生物化学与生物物理进展. 2003, 30(6): 879-883
27 施建宇, 潘泉.基于支持向量机融合网络的蛋白质折叠子识别研究.生物化学与生物物理进展,2006, 33(2): 155-162
28 E Bindewald, A Cestaro, J Hesser, M Heiler, and S C E Tosatto. MANIFOLD: protein fold recognition based on secondary structure, sequence similarity and enzyme classification. Protein Eng. 2003, 16(11): 785-789.
29 N Bhardwaj, R E Langlois, G Zhao, H Lu. Kernel-based machine learning protocol for predicting DNA-binding proteins. Nucleic Acids Res. 2005, 33(20): 6486-6493.
30 H B. Shen, K C Chou. Ensemble classifier for protein folds pattern recognition. Bioinformatics. 2006, 22(14): 1717-1722.
31 S Hochreiter, M Heusel, and K Obermayer. Fast model-based protein homology detection without alignment. Bioinformatics. 2007, 23(14): 1728 - 1736
32 刘晓辉,李晓琴. 全α类蛋白质核心结构的折叠分类研究.生物物理学报,2006, 22(增刊): 370-371
33 张 炜,李晓琴. 基于二级结构片段的β类蛋白质折叠类型分类研究.生物物理学报,2006, 22(增刊): 387-388
34 刘晓辉. 全α类蛋白质核心结构的折叠分类研究. 北京工业大学硕士论文. 2007:5-23
35 张 炜.基于二级结构片段的β类蛋白质折叠类型分类研究. 北京工业大学硕士论文. 2007:16-24
36 Sean R Eddy. Profile hidden Markov models. Current Opinion in Structural Biology, 1996, 6(3): 361-365
37 SR Eddy. Profile hidden Markov models. Bioinformatics.1998, 14(9): 755-763
38 SR Eddy. Multiple alignment using hidden Markov models. Proc Int Conf Intell Syst Mol Biol. 1995, 3:114-20
39 K Karplus, C Barrett, R Hughey. Hidden Markov Models for Detecting Remote Protein Homologies. Bioinformatics, 1998, 14(10): 846-856
40 A Krogh, M Brown, I S Mian. Hidden Markov models in computational biology: applications to protein modeling. J. Mol. Biol. 1994, 235: 1501-1531.
41 R Durbin, S Eddy, A Krogh. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge: Cambridge University Press, 1998, 100-132
42 Jong Park1, Kevin Karplus, Cyrus Chothia. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. Journal of Molecular

Biology. 1998, 284(4): 1201-1210
43 Kevin Karplus, Kimmen Sjölander, Christian Barrett. Predicting protein structure using hidden Markov models. Proteins: Structure, Function, and Genetics. 29(S1): 134 - 139
44 Julian Gough1, Kevin Karplus, Richard Hughey, Cyrus Chothia. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. Journal of Molecular Biology. 2001, 313(4): 903-919
45 Gough J, Chothia C. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic Acids Res. 2002, 30: 268-272
46 Madera M, Vogel C, Chothia C. The SUPERFAMILY database in 2004: additions and improvements. Nucleic Acids Research. 2004, 32(Database issue): 235-239
47 Madera M, Gough J. A comparison of profile hidden Markov model procedures for remote homology detection. Nucleic Acids Res. 2002, 30, 4321-4328
48 Sonnhammer E L L, Eddy S R, Durbin R. Pfam: A comprehensive database of protein domain families based on seed alignments. Proteins. 1997, 28: 405-420
49 A Bateman, E Birney, R Durbin, SR Eddy. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res. 1999, 27: 260-262.
50 EL Sonnhammer, SR Eddy, E Birney, A Bateman. Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Research. 1998, 26(1): 320-322
51 Alex Bateman, Ewan Birney, Lorenzo Cerruti. The Pfam Protein Families Database. Nucleic Acids Research, 2002, 30(1):276-280
52 Robert D Finn, Jaina Mistry, Benjamin Schuster-Böckler. Pfam: clans, web tools and services. Nucleic Acids Research. 2006, 34(Database issue): 247-251
53 Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL compendium in 2004. Nucleic Acids Research. 2004, 33(D):189-192.
54 Chandonia J M, Walker N S, Lo CONTE L, Koehl P, Levitt M, Brenner SE. ASTRAL compendium enhancements. Nucleic Acids Rearch 30:260-263(2002).
55 Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for sequence and structure analysis. Nucleic Acids Research 28:254-256(2000).
56 J. Patrick, Fleming, Haipeng Gong, D.George, Rose. Secodary structure determines protein topology.Protein Sci. 2006, 15: 1829-1834
57 AV. Finkelstein, OB. Ptitsyn. Why do globular proteins fit the limited set of folding patterns. Prog. Biophys. Molec. Biol. 1987, 50: 171-190
58 Konagurthu, A.S., Whisstock, J.C., Stuckey, P.J., Lesk, A.M. MUSTANG: A multiple structural alignment algorithm. Proteins: Structure, Function, and Bioinformatics. 2006, 64(3): 559-574
59 Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004, 32:1792-1797.
60 Notredame C. Recent progress in multiple sequence alignments: A survey. Pharmaco genomics, 2002, 3:1-14.
61 Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J.Mol. Biol., 1993, 233: 123-138.
62 KristinK.Koretke, RobertB.Russell. Fold Recognition from Sequence Comparisons. PROTEINS: Structure, Function, and Genetics. 2001, 5(Supp)l: 68-75
63 R. Kuang, E. Ie, K. Wang, K. Wang, M. Siddiqi, Y. Freund and C. Leslie. Profile-based

string kernels for remote homology detection and motif extraction. Proceedings of the IEEE Computational Systems Bioninformatics. 2004: 152-160

64 Chinnasamy A, Mittal A. Protein structure and fold prediction using tree-augmented naive bayesian classifier. JBioinform Comput Biol. 2005, 3 (4): 803-820.

65 Yu Chen, Gordon M, Croppen. Fold Recognition via a Tree, Journal of Computational Biology. 2006, 13(9): 1565-1573

攻读硕士学位期间发表的学术论文

1. 任文科，徐海松，李晓琴. Globin-like 蛋白质折叠类型识别. 生物化学与生物物理进展. 2007. 35(5):548-554.
2. Hai Song Xu, Wen Ke Ren, Xiao Qin Li. ALPHASUM: a score matrix based on low identity all-alpha proteins. 2008
3. Hai Song Xu, Wen Ke Ren, Xiao Hui Liu and Xiao Qin Li. Improving sequence alignment using class-specific score matrices. 2008
4. 刘晓辉，李晓琴，任文科，徐海松, α 类蛋白折叠类型的氨基酸取代矩阵研究. 中国生物医学工程进展. 2007: 974-977
5. 张玮，李晓琴，徐海松，任文科. 蛋白质折叠类型识别方法研究. 生物物理学报. 2008，1(24)
6. 刘晓辉，李晓琴，徐海松，任文科. 构建基于 α 类蛋白折叠核心的氨基酸取代矩阵. 中国生物化学与分子生物学报. 2008，7

攻读硕士学位期间获得科研成果

轻松的局部序列比对软件. 计算机软件著作权.登记号：2008RBJ0201.

攻读硕士学位期间参加的科研项目

国家自然科学基金资助项目（30570427），蛋白质折叠信息数据库及折叠经验规律探索。

致 谢

首先，我要感谢李晓琴教授，在她的精心指导下，我对于科学研究有了更深入的了解，学习到了更多的知识，提高了学习和工作的能力；导师在生活上还给予我深切的关怀，帮助我顺利的完成了学业。李老师的教诲将使我受益终身。

同时，我要感谢实验室刘晓辉、张炜、徐海松、杨效增、刘岳、乔辉等同学在学术和生活上的关心与帮助，正是有了他们的陪伴，我的研究生生活才能丰富快乐，学有所成。

我还要感谢生命学院和北京工业大学在三年中提供的良好的学习科研环境，使我能安心学习，认真工作！

最后，感谢我的父母，他们的爱是我力量和勇气的源泉，使我克服各种困难，总以最好的精神来面对生活。

任文科

2008 年 4 月