PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2008, 35(5): 548~554
www.pibb.ac.cn

Globin-like 蛋白质折叠类型识别 *

任文科 徐海松 李晓琴 **
(北京工业大学生命科学与生物工程学院，北京 100021)

摘要 蛋白质折叠类型识别是蛋白质结构研究的重要内容．以 SCOP 中的 Globin-like 折叠为研究对象，选择其中序列同一性小于 25%的 17 个代表性蛋白质为训练集，采用机器和人工结合的办法进行结构比对，产生序列排比，经过训练得到了适合 Globin-like 折叠的概形隐马尔科夫模型(profile HMM)用于该折叠类型的识别．以 Astral1.65 中的 68 057 个结构域样本进行检验，识别敏感度为 99.64%，特异性 100%．在折叠类型水平上，与 Pfam 和 SUPERFAMILY 单纯使用序列比对构建的 HMM 相比，所用模型由多于 100 个归为一个，仍然保持了很高的识别效果．结果表明：对序列相似度很低但具有相同折叠类型的蛋白质，可以通过引入结构比对的方法建立统一的 HMM 模型，实现高准确率的折叠类型识别．

关键词 蛋白质，折叠类型识别，Globin-like，隐马尔科夫模型，结构比对
学科分类号 Q615，Q518，O629.7

蛋白质的氨基酸序列如何决定其空间结构是生命科学研究领域的核心问题之一，被称为第二遗传密码．国际上由蛋白质的氨基酸序列预测其空间结构的研究方法有：同源模建、从头预测、折叠识别[1, 2]．同源模建受序列同源性限制，从头预测受计算能力制约，介于上述两种方法之间的折叠识别被认为是最有前途的方法．

蛋白质结构的理论预测经过几十年的发展，预测结果依然不能被蛋白质结构数据库(PDB)所接收．实际上，精确的蛋白质三级结构预测是困难的，比较有效的方法是对蛋白质粗粒化的空间构象进行分类和预测．相关的工作主要集中在蛋白质结构型(6 种左右)识别、框架结构(十几种)预测、蛋白质折叠类型数目的估计和分布上，但对蛋白质折叠类型进行系统研究的工作报道却很少．

一般认为蛋白质的折叠类型只有数百到数千种[3]，远小于蛋白质所具有的自由度数．折叠所潜藏的物理的简单性使我们预期：折叠速率和机制看上去在很大程度由天然态的拓扑所决定[4]．蛋白质折叠类型包含了蛋白质空间结构形成中起关键作用的疏水内核，同时，形成蛋白质折叠类型的不同二级结构片段又通过长程相互作用实现空间紧邻．因此，抓住蛋白质折叠类型研究问题可以将疏水内核在结构形成中的关键作用与长程相互作用两者有机组合．对自然界存在的数百到数千种折叠类型进行系统研究，探索蛋白质折叠形成的经验规律，这将有助于揭示蛋白质的折叠规律、解决蛋白质折叠问题．

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[5, 6]，迫切需要一个建立在统一原理基础上的蛋白质折叠类型数据库，从而能给研究者以指导，在使用中对不同的库做出适当评价和取舍．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[7, 8]，为蛋白质折叠识别奠定了基础．

折叠类型识别研究在国内外已有一些报道并取得一定进展[9, 10]，基本思路是：提取氨基酸组成成分、极性、疏水性、预测二级结构等信息作为特征参数，利用神经网络或支持向量机方法识别折叠类型，由于忽略了位点特异性等原因，平均精度约在 60%左右．

目前在蛋白质精确三维结构预测领域，profile

* 国家自然科学基金资助项目(30570427)和北京市自然科学基金资助项目(4063035).
** 通讯联系人.
Tel: 010-67391610, E-mail: lxq0811@bjut.edu.cn
收稿日期：2007-09-15，接受日期：2007-12-04

HMM 被认为是最有效的方法之一[11]，为了探索同一种折叠是否有相同的序列模式，是否使用较少的代表性样本即可实现高准确率的 HMM 识别，本文进行一个方法上的尝试：选取生物学研究比较充分的 Globin-like 折叠类型为研究对象，提取其中的少数低相似度序列，引入结构比对，得到基于结构的序列排布，训练 HMM，建立 Globin-like 折叠蛋白质的序列 HMM 识别方法.

1　样本选择

选取的代表性样本来自结构域序列数据库 Astral1.65，该数据库提供一个序列同一性小于 25%的子集，通过看图软件结合结构分析发现其中 17 个样本的折叠类型相同，且在 SCOP 数据库也同属于 Globin-like 折叠子，样本的家族分布如表 1.

Table 1　Target protein used in Globin-like fold of Astral1.65

Family	Amount of family members	Amount of target selected	Astral ID of target
Globins	196	12	d1a6m__ d1ash__ d1b0b__ d1cqxa1 d1ew6a_ d1h97a_ d1hlb__ d1irdb_ d1it2a_ d1itha_ d2gdm__ d3sdha_
Truncated hemoglobin	7	2	d1dlwa_ d1ngka_
Phycocyanin-like phycobilisome	28	2	d1jboa_ d1jbob_
Neural globin	1	1	d1kr7a_

Globin-like 折叠类型的蛋白质在 SCOP1.65 中包含 1 011 条记录，数目庞大，可以作为统计分析的数据集．其成员包括 globins、truncated hemoglobin、phycocyanin-like 及 neural globin 四个序列家族：globins 为铁血红素结合蛋白，truncated hemoglobin 为剪切血色素，neural globin 是神经组织血色素，这三个家族的核心结构由 6 个螺旋两两一组以三角架的方式紧密搭建，构成类球体，功能均和氧气的输送、结合有关，这三个家族的成员占整个折叠类型的绝大多数，而 phycocyanin- like 家族是后色胆素发光团受体，和前三个家族相比，功能有较大差异，其三维结构在 N 端有两个额外的螺旋片段，但核心部分 6 个螺旋的结构与其他成员非常相似，折叠结构参见示意图 1.

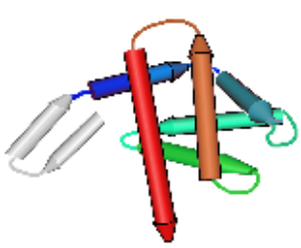

Fig. 1　A cartoon representation of Globin-like proteins

The core structure of Globin-like protein composed of six helices, which were drawn in color; two additional helices of phycocyanin-like family were shown in gray.

2　方　　法

2.1　结构比对

序列相似小于 30%时，序列比对准确度不高，这是因为在这些序列中具有相似结构功能的不同残基在序列比对中往往被错误配对[12]．因此，对于序列相似度小于 25%的 Globin-like 蛋白样本，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，得到基于结构的序列排布，以此来构建模型.

MUSTANG (a multiple structural alignment algorithm)[13]是 Lesk 等在 DALI 双结构比对获得成功的基础上于 2006 年发展的一种多结构比对方法，对于空间折叠、残基接触模式有较强的识别能力，其双结构比对结果与 DALI 相当，多结构比对结果与其他一些现有的工具如 POSA、CE-MC 等相比相似或者更好．我们将 MUSTANG 多结构比对算法与手工调整相集合：在家族内使用 MUSTANG 多结构比对算法，在家族之间使用手工进行调整，使核心部分的 6 个螺旋达到最大叠合，比对结果的叠合如图 2，通过结构比对得到的序列对齐如图 3 所示.

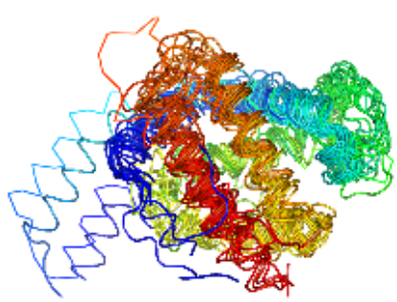

Fig. 2　Multiple structure superposition of the 17 Globin-like proteins used for building of HMM

Only backbone is shown, This figure was generated by PyMol (Delano Scientific, San Carlos,CA).

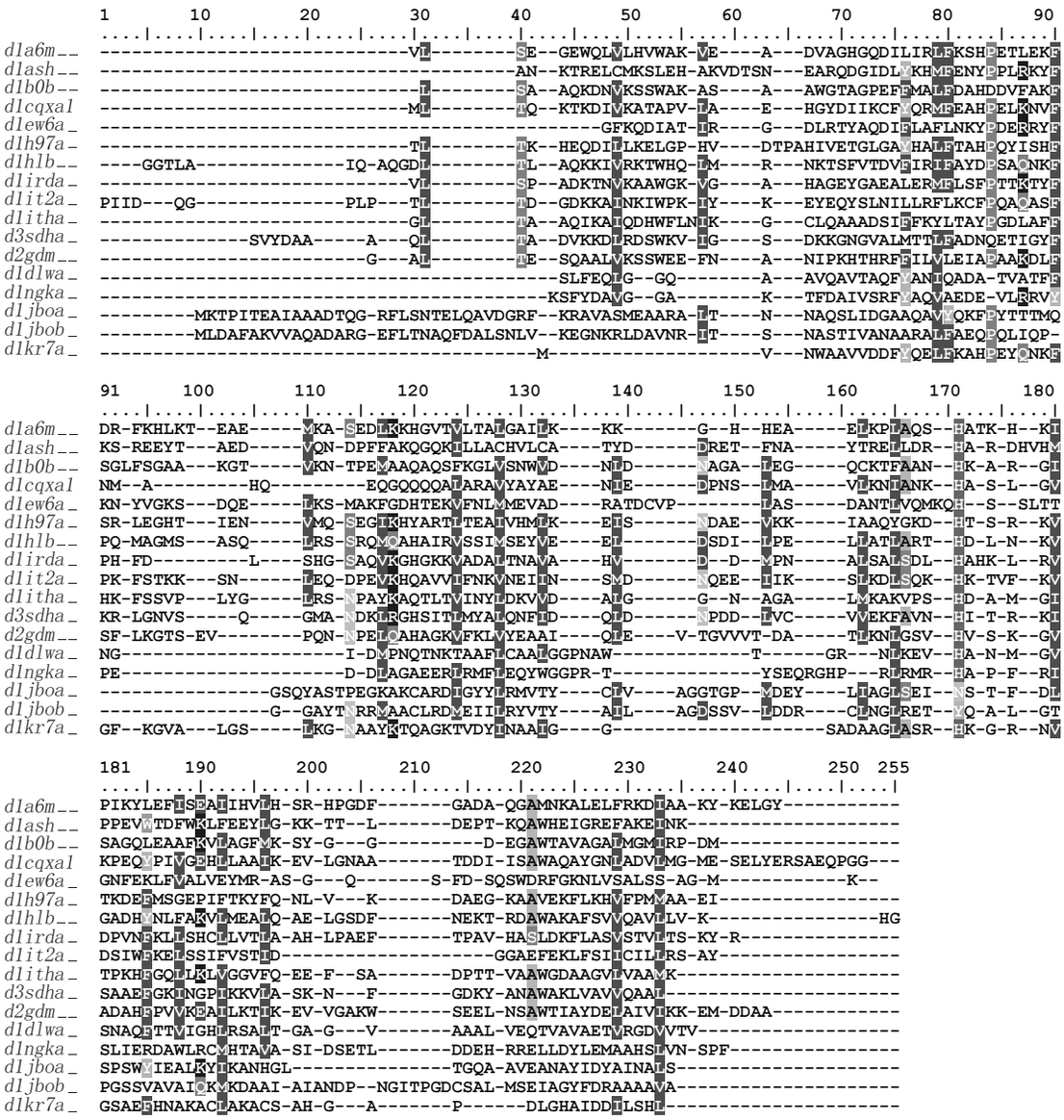

Fig. 3　Structural-based sequence alignment of 17 proteins used for building of HMM

Colors indicate by BioEdit [14], threshold for shading is > 50%.

2.2　比对结果分析

为确定比对结果的可信性，对比对结果做了两两空间坐标均方根偏差(RMS)分析：以图 3 为依据，将所有两两匹配残基的 C_α 原子纳入 RMS 计算，结果见图 4，整个结构比对的 RMS 为 3.54Å．通过统计，在 FSSP(families of structurally similar proteins)[15]中，结构相似蛋白质 RMS 的均值约为 4.1Å．由于在计算中所有可能的残基均被纳入，而 Globin-like 折叠类型中，螺旋走向的角度差异通常会使 RMS 较大，可以认为比对结果显著．

尽管样本间序列同一性很低，但通过结构比对得到的序列对齐也表明：仍然存在一些保守的位点，如图 3 中功能相关的 171HIS 等以及螺旋内的一些强疏水性位点，为 Globin-like 折叠类型建立统一的序列 HMM 提供了依据．

ID/sequence length	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. d1a6m__/1-151		138	137	137	129	141	143	140	132	139	135	144	108	116	124	128	107
2. d1ash__/1-147	2.16		137	129	127	140	140	132	131	137	136	137	108	112	125	128	107
3. d1b0b__/1-141	1.86	2.47		128	127	140	138	131	134	135	134	136	105	110	123	125	106
4. d1cqxa1/1-150	2.80	3.37	2.93		118	132	133	134	124	128	128	138	107	115	122	126	96
5. d1ew6a_/1-137	2.13	2.50	2.48	3.02		129	127	123	121	126	123	126	101	106	112	114	105
6. d1h97a_/1-147	2.44	2.92	2.31	3.15	2.96		142	135	135	139	138	140	108	113	127	129	107
7. d1hlb__/1-157	2.48	2.83	2.46	3.64	3.06	3.17		137	137	139	137	141	108	115	131	135	107
8. d1irda_/1-141	1.48	2.38	2.10	2.94	2.64	2.51	2.63		127	133	131	138	108	116	124	128	101
9. d1it2a_/1-146	1.78	2.59	1.91	3.14	2.32	2.13	2.69	1.76		130	130	133	100	105	125	124	101
10. d1itha_/1-141	1.83	2.61	1.95	3.11	2.07	2.38	2.85	2.16	1.99		136	136	107	111	125	127	107
11. d3sdha_/1-145	2.01	2.63	2.14	2.96	2.43	2.93	2.62	1.98	2.06	2.36		137	106	109	132	134	105
12. d2gdm__/1-153	2.64	3.04	2.82	3.45	3.00	2.97	3.10	3.10	2.87	3.10	3.29		109	115	127	131	104
13. d1dlwa_/1-116	3.71	4.14	3.80	3.91	3.49	3.54	3.77	3.57	3.68	3.81	3.48	3.73		113	101	103	94
14. d1ngka_/1-126	4.83	5.12	5.01	5.06	4.92	4.52	4.89	4.87	4.85	4.94	4.91	4.76	2.43		105	109	94
15. d1jboa_/1-162	5.23	4.97	5.07	5.71	4.48	4.87	6.25	5.42	5.94	5.47	6.87	4.76	4.90	5.22		158	94
16. d1jbob_/1-171	5.79	5.61	5.55	6.33	5.09	5.23	6.59	5.92	6.22	6.03	7.37	5.67	5.08	5.27	2.70		96
17. d1kr7a_/1-110	2.47	3.27	2.52	2.90	2.51	2.83	2.86	2.71	2.36	2.45	2.13	5.95	3.02	3.20	4.14	4.42	

Fig. 4　RMS of structure alignment

: Number of residues pairs involved in RMS. ■: RMS (Å).

All targets are sorted by families; sequence length indicated the number of amino acids for each target.

2.3　HMM 构建

HMM 是目前蛋白质序列的分类识别最成功的方法之一，见图 5，HMM 将氨基酸在比对中的状态分为匹配、插入、删除三个状态，氨基酸序列看作是由这三种状态之间以不同概率的跳转得到的，状态匹配或者插入的时候，20 种氨基酸的概率取值不同，如图 5 所示，M 表示匹配，I 表示插入，D 表示删除，箭头表示状态跳转．当处于匹配态 M 时，还引入一个 20 种氨基酸的概形，表征该位点出现氨基酸的不同概率，在相邻匹配的位点 M 之间，有插入状态 I，每个位置的插入突变概率不同，同一位置插入不同氨基酸的概率也不同，不相邻的氨基酸之间存在删除状态，表示某真实序列与此模型相比在该位点可能存在删除突变，删除多个连续位点可以通过连续地跳转到 D 来实现．

通过图 5 可以看到一个隐马尔科夫模型的框架包含以下几个要素：a．状态的数目 M，也就是 HMM 的长度，一般和训练集的平均氨基酸序列长度相当．b．不同状态之间的跳转概率，对应于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数值．c．每个 M 态及 I 态，20 种氨基酸出现的概率，实质上是一个 profile．

从理论上说，HMM 的框架模型对所有位点的所有状态都可以描述，包含了所有可能的氨基酸序列，非常灵活，由于决定模型三个要素的参数都是通过训练得到的，对于一个 HMM，起决定性作用的事实上是训练样本[16, 17]．

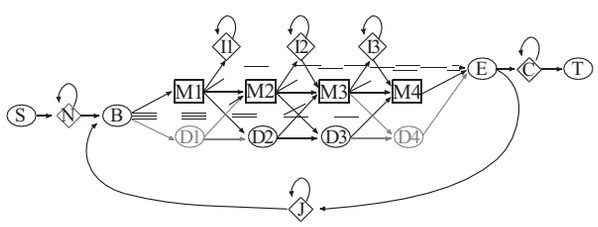

Fig. 5　An architecture of profile HMM

Squares indicate match states (modeling consensus positions in the alignment). Diamonds indicate insert states (modeling insertions relative to consensus) and special random sequence emitting states.Circles indicate delete states (modeling deletions relative to consensus) and special begin /end states. Arrows indicate state transitions.

对于 Globin-like 蛋白质，以图 3 基于结构的序列比对结果作为训练集，用 HMMer 算法构建模型[19]，训练后确定了 HMM 模型的各个参数，其 Martin Madera 示意图如图 6．包括 167 个位点，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，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．训练集中的保守位点均在 HMM 中得到体现，如图 3 中的 83Pro 对应于 HMM 的 33Pro，90Phe 对应于 HMM 的 39Phe，功能位点

171His 对应于 HMM 的 108His．相对于螺旋内部，在螺旋之间插入空位要容易得多，所以除去 40 位点以前 Phycocyanin-like family 家族在 N 端的两个螺旋，插入几率曲线(底部浅绿线)有几个峰值，藉此可以识别出 Globin-like 折叠蛋白质核心部分的 6 个较长的螺旋．

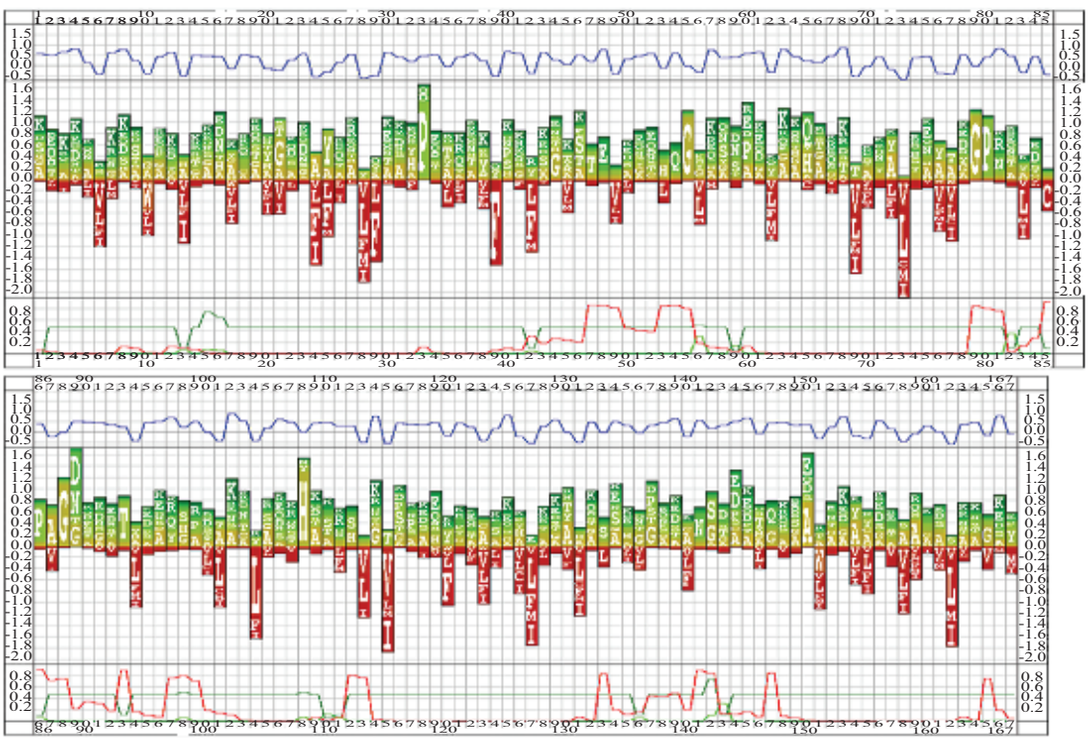

Fig. 6　A graphical representation of the HMM model for Globin-like fold

Thanks Martin Madera′s script for drawing this Graphic of profile HMM[18].

3　结　　果

为检验 HMM 识别效果，我们在数据库 Astral1.65 的基础上，以 SCOP 数据库人工分类为准，建立了由 1 101 个蛋白质氨基酸序列组成的 Globin-like 折叠类型数据库，以及由 66 926 个蛋白质氨基酸序列组成的非 Globin-like 折叠类型数据库．在使用 HMMer 算法默认参数的条件下，用得到的 HMM 分别对上述两个数据库所属蛋白质进行 Globin-like 折叠类型识别，定义统计评估指标 t_p 为 Globin-like 类正确预测的总数，f_p 为其他类错误预测为 Globin-like 类的总数，t_n 为其他类正确预测为非 Globin-like 的总数，f_n 为 Globin-like 类错误预测为其他类的总数，结果见表 2．

Table 2　Identify result of structure-based HMM

Identify result	Protein number
t_p	1 097
f_p	0
t_n	66 926
f_n	4

识别特异性：$S_p=\frac{t_n}{t_n+f_p}\times 100\%$，识别敏感性：$S_n=\frac{t_p}{t_p+f_n}\times 100\%$．

利用 HMM 识别 Globin-like 特异性为 100%，敏感性 99.637%．

Pfam[19]和 SUPERFAMILY[20]数据库是目前蛋白质机器分类识别中比较成功的两个例子，均使用 HMM 作为分类算法，Globin-like 折叠类型在这两个库中包含的 HMM 多达上百条，与之相比较，借助结构信息建立的序列 HMM 极大地减少了模型数量，如表 3．

Table 3　Compare between Structure-based HMM and Pfam，SUPERFAMILY

Method	Number of HMM model
Pfam	107
SUPERFAMILY	101
Structure-based HMM	1

4 结　论

通常将 HMM 用于家族和超家族的分类识别是成功的，但是在折叠类型层面，结构相似的蛋白质之间序列相似性可能很小，多序列比对方法得不到一个有意义的排布，这给使用 HMM 进行机器识别造成了困难，本文以 Globin-like 折叠类型为例，只抽取少数代表性序列，借助结构比对得到序列排布，利用该结果进行 HMM 训练，识别准确性达 99.994%，与 Pfam，SUPERFAMILY 相比，显著减少了 HMM 模型的数目和训练所需的样本，识别效果基本一致．这说明：具有相同折叠的蛋白质，虽然序列两两比对的同一性很低，但通过结构比对仍然可能找到该结构序列上的概形分布和突变模式．但是由于 HMM 算法的限制，这种模式还有待进一步的深入研究．

参 考 文 献

1 Eisenberg D. Into the black of night. Nature Structural Biology, 1997, **4**: 95～97

2 Shortle D. Structure Prediction: Folding proteins by pattern recognition. Current Biology, 1997, **7**(3): 151～154

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

4 David B. A surprising simplicity to protein folding. Nature, 2000, **405**(6782): 39～42

5 Novotny M, Madsen D, Kleywegt G J. Evaluation of protein fold comparison servers. Proteins, 2004, **54**(2): 260～270

6 Matsuda K, Nishioka T, Kinoshita K. Finding evolutionary relations beyond superfamilies: Fold-based superfamilies. Protein Science, 2003, **12**(10): 2239～2251

7 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
Liu X H, Li X Q. Acta Biophys Sin, 2006, **22**(Suppl): 370～371

8 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
Zhang W, Li X Q. Acta Biophys Sin, 2006, **22**(Suppl): 387～388

9 Ding C H Q, Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics, 2001, **17**(4): 349～358

10 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
Shi J Y, Pan Q. Prog Biochem Biophys, 2006, **33**(2): 155～162

11 Dunbrack R L. Sequence comparison and protein structure prediction. Curr Opin Struc Biol, 2006, **16**(3): 374～384

12 李 菁, 王 炜. 氨基酸残基归类及用简化后的字符识别蛋白质结构保守区域. 中国科学 C 辑, 2006, **36**(6): 552～562
Li J, Wang W. Science in China Series C-Life Sciences, 2006, **36**(6): 552～562

13 Konagurthu A S, Whisstock J C, Stuckey P J, *et al*. MUSTANG: A multiple structural alignment algorithm. PROTEINS-Structure Function and Bioinformatics, 2006, **64**(3): 559～574

14 Hall T A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series, 1999, **41**: 95～98

15 Holm L, Sander C. Mapping the protein universe. Science, 1996, **273**(5275): 595～602

16 Eddy S R. Profile hidden Markov models. Current Opinion in Structural Biology, 1996, **6**(3): 361～365

17 Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics, 1998, **14**(10): 846～856

18 Madera M, Vogel C, Chothia C. The SUPERFAMILY database in 2004: additions and improvements. Nucleic Acids Research, 2004, **32**(Database issue): 235～239

19 Durbin R, Eddy S, Krogh A. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge: Cambridge University Press, 1998. 100～132

20 Finn R D, Mistry J, Benjamin Schuster-Böckler. Pfam: clans, web tools and services. Nucleic Acids Research, 2006, **34** (Database issue): 247～251

Identification Proteins of Globin-like Fold*

REN Wen-Ke, XU Hai-Song, LI Xiao-Qin**

(*Bioengineering Center, Beijing University of Technology, Beijing* 100021, China)

Abstract Identifying protein fold is an important issue in protein structure research. Based on the classification of SCOP1.65, 17 Globin-like proteins from four homology families (< 25% sequence identity) are selected from Astral 1.65. The sequence alignment result, from structure alignment tool MUSTANG combined with manual inspection, has been used to generate a profile HMM of Globin-like fold. In a fold identify test on 68 057 sequences of Astral-1.65, the model identified 1 097 Globin-like proteins rightly, only 4 proteins of this fold are not correctly distinguished. The sensitivity and specificity of the profile HMM reach to 99.64% and 100%, respectively. Compared with Pfam and SUPERFAMILY which construct HMM based on merely sequence alignment, the model number is reduced from about 100 to 1, while keeping the sensitivity at the same level. The result shows that, for those proteins with same fold type but low sequence identity, a unified HMM could be constructed by introducing structure alignment to fold identify with high accuracy.

Key words protein, fold identify, Globin-like, profile HMM, structure alignment

---

*This work was supported by grants from The National Natural Science Foundation of China (30570427) and Natural Science Foundation of Beijing (4063035).

**Corresponding author . Tel: 86-10-67391610, E-mail: lxq0811@bjut.edu.cn

Received: September 15, 2007 Accepted: December 4, 2007