

基于功能域组分的蛋白质折叠类型识别*

闫金丽 陈治伟 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 100124)

摘要 蛋白质空间结构研究是分子生物学、细胞生物学、生物化学以及药物设计等领域的重要课题。折叠类型反映了蛋白质核心结构的拓扑模式, 对折叠类型的识别是蛋白质序列与结构关系研究的重要内容。选取 LIFCA 数据库中样本量较大的 53 种折叠类型, 应用功能域组分方法进行折叠识别。将 Astral 1.65 中序列一致性小于 95% 的样本作为检验集, 全库检验结果中平均敏感性为 96.42%, 特异性为 99.91%, 马修相关系数(MCC)为 0.91, 各项统计结果表明: 功能域组分方法可以很好地应用在蛋白质折叠识别中, LIFCA 相对简单的分类规则可以很好地集中蛋白质的大部分功能特性, 反映了结构与功能的对应关系。

关键词 折叠类型, 折叠识别, 功能域, LIFCA 数据库

学科分类号 Q615, Q518, O629.7

DOI: 10.3724/SP.J.1206.2010.00345

蛋白质空间结构研究是分子生物学、细胞生物学、生物化学以及药物设计等领域的重要课题^[1]。但由于 X 射线晶体衍射以及核磁共振等实验方法的费时费力, 使得蛋白质结构的测定速度远远跟不上序列的增长速度。因此, 基于知识预测已知序列的蛋白质结构变得十分紧迫。目前, 国际上由蛋白质的氨基酸序列预测其空间结构的方法有以下三种: 同源模建法、折叠识别法和从头计算法^[2]。同源模建法受到序列相似度的限制, 从头计算运算量太大。介于两者之间的折叠识别被认为是最有潜力的方法。

蛋白质折叠类型分类是蛋白质折叠研究领域的基本问题, 同时也是蛋白质折叠识别的基础。SCOP^[3-4]、CATH^[5-6]等是蛋白质折叠分类数据库中应用较为广泛的数据库, 折叠分类也基本上是靠专家来完成, 但两者的分类依据及分类结果并不相同^[7-8], 分类的目的也并不是为了进一步的蛋白质折叠识别^[9]。为更好地开展蛋白质折叠机理及折叠类型稳定性研究, 以及进一步的折叠识别、蛋白质折叠结构进化研究, 建立基于统一规范的蛋白质折叠类型数据库特别重要。

蛋白质折叠识别研究在国内外已有报道并取得了一些进展, 涉及到的研究方法可以分为三类: 基于序列^[10-14], 基于结构^[15-16], 以及两者混合的特征

提取方法^[17-19]。这些方法都是基于 SCOP 折叠类型分类, 大多数研究只针对 SCOP 的少数折叠类型, 其中应用较广泛的是 Ding 的 27 种折叠类型: Ding 等^[12]利用氨基酸组分、极性、疏水性、二级结构作为特征向量, 用支持向量机(SVM)对这 27 种折叠类型进行识别, 平均预测精度为 56%; 施建宇等^[13, 17]用支持向量机融合网络以及图像特征进行识别, 总分类精度分别为 61.04%, 71.95%; Shen 和 Chou^[18]应用功能域组分和序列进化信息, 基于 Ding 的 27 种折叠类型进行识别研究, 平均识别准确率为 70.5%。这些研究中较注重的是方法的检验, 所用到的数据集样本量较小, 识别的准确率不超过 75%。

最近, 在规模更大或全库水平进行折叠分类的工作陆续被报道^[14, 19], 结果优于上述方法。Gewehr 等^[19]运用特异模式与类别的映射关系, 对 SCOP 库进行自动分类, 从多种角度构建检验集(包含全库检验)对结果进行检验, 在折叠水平的总体敏感性为 93.36%, 特异性为 98.13%, 分类效果较

* 国家自然科学基金(30570427)和北京市自然科学基金(4092008)资助项目。

** 通讯联系人。

Tel: 13552340552, E-mail: lxq0811@bjut.edu.cn

收稿日期: 2010-09-13, 接受日期: 2010-11-05

好. Jain 等^[14]运用多种机器学习算法对较大的数据集进行了自动结构分类, 最终一种集合决策树 (boosted random forest) 在对不同结构层次的划分中都取得了较好的结果, 其中交叉验证的平均准确率为 97.0%, 基于 SCOP 分类系统在结构类、折叠类型、超家族、家族水平的敏感性分别为 0.98、0.75、0.90、0.97.

在前期工作中, 我们给出了蛋白质框架结构构建及结构类识别方法^[20], 以蛋白质结构核心的拓扑连接和空间排布为依据构建^[21-22]了 LIFCA 数据库, 以该分类数据库为基础, 建立了基于 HMM 的蛋白质折叠类型识别方法, 并应用于 Globin-like^[23]及 36 种样本量较大的折叠类型识别中^[24], 36 个折叠类型的平均识别敏感性为 90.36%, 特异性为 99.99%, 马修相关系数 (MCC) 值为 0.95. 其中我们发现: 样本量小的折叠类型可以通过结构比对建立统一的 HMM 模型, 样本量较大的折叠类型无法建立统一的 HMM 模型, 需要分类建模^[24], 过程比较繁琐.

本文利用在亚细胞定位^[25]、蛋白酶类别划分^[26]、蛋白质结构类预测领域^[27]均有广泛应用并得到了很好的预测效果的功能域组分方法, 对 LIFCA 中样本量较大的折叠类型进行了折叠识别研究, 在多个检验集上对所方法进行检验.

1 材料和方法

1.1 实验数据

1.1.1 实验集. LIFCA 是以蛋白质结构拓扑不变性为原则、以蛋白质折叠核心规则结构片段的组成、连接和空间排布为依据建立的低相似度蛋白质折叠分类数据库^[21-22], 包含同一性低于 25% 的 2 406 个蛋白质样本, 共 259 种折叠类型. 与 SCOP 数据库^[3-4]相比, LIFCA 数据库的折叠类型分类涉及部分 SCOP 数据库中不同折叠类型家族的合并, 也涉及小部分同一折叠类型家族的拆分.

本文选取了 LIFCA 数据库中样本量大于 10 的 53 种折叠类型, 排除在 Pfam 数据库中无搜索结果的样本, 共得到 1 826 个样本作为实验集. 这些样本总共涉及 688 个 SCOP 家族, 其中, 在 LIFCA 数据库折叠类型分类中涉及家族拆分的为 31 个.

1.1.2 检验集.

为了评估本文所用方法的优劣, 以及其对涉及 SCOP 家族拆分样本的识别敏感性, 我们分别从不同层次进行了识别结果检验.

检验集 a: 选取 Astral1.65 中序列同一性低于

95% 的样本, 排除在 Pfam 中无搜索结果的样本. 共得到 9 211 个样本作为检验集. 其中属于实验集 53 种折叠类型的样本量为 3 857 个, 对检验集的覆盖度为 41.87%.

检验集 b: 选取检验集 a 中属于实验集 53 种折叠类型的样本, 剔除与实验集 1 826 个蛋白质相同的样本, 得到一个独立的检验集, 共 2 035 个蛋白质, 其中在 LIFCA 数据库折叠类型分类中涉及家族拆分的样本为 176 个.

1.2 方法

1.2.1 折叠类型及样本的功能域组分表示. 蛋白质通常由一个或多个功能区域组成, Pfam 数据库对大量的已知功能域进行了多序列比对和 profile-HMM 建模^[28]. 在 Pfam 中可以通过序列搜索确定查询序列中包含的功能域信息, 从而确定蛋白质所属折叠类型涵盖的功能特征. 将实验集中 1 826 个样本在 Pfam 中进行搜索, 为了获得高敏感性搜索结果, 设定搜索 E 值 < 0.01 , 共得到 1 032 个功能域组分.

蛋白质折叠类型 i 的功能域组分向量为 F_i :

$$F_i = \begin{bmatrix} a_{i1} \\ a_{i2} \\ \cdots \\ a_{ij} \\ \cdots \\ a_{i1\ 032} \end{bmatrix} \quad (i=1\cdots 53, j=1\cdots 1\ 032)$$

i 为蛋白质折叠类型编号, j 为蛋白质功能域编号, a_{ij} 为第 i 种折叠类型的第 j 个向量元素. 当折叠类型 i 所属样本在 Pfam 中搜索到功能域 j 时 a_{ij} 为 1, 否则为 0.

相应的, 任意蛋白质样本 p 的功能域组分向量为 P :

$$P = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_j \\ \cdots \\ b_{1\ 032} \end{bmatrix} \quad (j=1\cdots 1\ 032)$$

若蛋白质样本 p 在 Pfam 中搜索到功能域 j 时 b_j 为 1, 否则为 0.

1.2.2 蛋白质样本的折叠类型判定. 蛋白质 p 与折叠类型 i 的相关程度可以表示为:

$$\Lambda(P, F_i) = \frac{P \cdot F_i}{\|P\| \cdot \|F_i\|} \quad (i=1 \cdots 53)$$

式中, $P \cdot F_i$ 表示两个向量之间的点乘, $\|P\|$ 和 $\|F_i\|$ 是向量的模, $\Lambda(P, F_i)$ 的取值分布于 0、1 间. 若蛋白质 p 与折叠类型 k 的相关程度最大, 即: $\Lambda(P, F_k) = \text{Max}\{\Lambda(P, F_1), \Lambda(P, F_2), \cdots, \Lambda(P, F_N)\}$, 则确定蛋白质 p 所属折叠类型为 k . 如果蛋白质 p 与任何折叠类型的相关性均为 0, 则蛋白质 p 不属于任何上述 53 种折叠类型.

1.2.3 评估参数.

为了评估本文方法的有效性, 我们采用了多种分类系统评估参数: 总分类精度 Q , 敏感性 (Sensitivity), 特异性 (Specificity), 马修相关系数 (Matthew's correlation coefficient, MCC 综合评价指数).

各参数定义如下:

$$Q = \sum_{i=1}^c p_i / N \quad c \text{ 为样本类别总数; } p_i \text{ 为第 } i \text{ 类样本识别正确的数目; } N \text{ 为样本总数.}$$

$$S_n = \frac{t_p}{t_p + f_n} \times 100\% \quad S_p = \frac{t_n}{t_n + f_p} \times 100\%$$

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

t_p : 真阳性数目, t_n : 真阴性数目, f_p : 假阳性数目, f_n : 假阴性数目.

2 结果与讨论

2.1 功能域组分方法的全库检验

表 1 是功能域组分方法的全库检验结果, 给出了所有折叠类型的详细识别结果, 包含每个折叠类型在实验集中涉及建立功能向量集的样本数、全库检验集中的样本数、预测的敏感性、特异性和相应的 MCC 值.

Table 1 Results of fold recognition on the test set a.

Fold type	Number in LIFCA	Number in Astral1.65(95%)	t_p	t_n	f_n	f_p	Sensitivity	Specificity	MCC
a_12_1	11	12	12	9 193	0	6	100.00%	99.93%	0.82
a_2_1	54	109	90	9 095	19	7	82.57%	99.92%	0.87
a_3_1	40	52	49	9 151	3	8	94.23%	99.91%	0.90
a_3_2	11	64	61	9 147	3	0	95.31%	100.00%	0.98
a_3_3	69	97	93	9 094	4	20	95.88%	99.78%	0.89
a_3_4	13	19	17	9 192	2	0	89.47%	100.00%	0.95
a_3_5	39	83	82	9 126	1	2	98.80%	99.98%	0.98
a_3_9	11	20	20	9 184	0	7	100.00%	99.92%	0.86
a_4_1	34	57	57	9 134	0	20	100.00%	99.78%	0.86
a_4_11	14	32	31	9 179	1	0	96.88%	100.00%	0.98
a_4_12	9	24	23	9 129	1	58	95.83%	99.37%	0.52
a_4_13	12	53	53	9 154	0	4	100.00%	99.96%	0.96
a_4_14	22	46	46	9 165	0	0	100.00%	100.00%	1.00
a_4_3	14	30	26	9 178	4	3	86.67%	99.97%	0.88
a_4_4	16	34	34	9 177	0	0	100.00%	100.00%	1.00
a_4_5	17	28	28	9 179	0	4	100.00%	99.96%	0.94
a_4_6	14	28	23	9 181	5	2	82.14%	99.98%	0.87
a_4_9	21	29	26	9 175	3	7	89.66%	99.92%	0.84
a_6_1	10	13	13	9 198	0	0	100.00%	100.00%	1.00
a_6_3	24	87	87	9 123	0	1	100.00%	99.99%	0.99
a_7_2	14	29	29	9 182	0	0	100.00%	100.00%	1.00
b_12_3	24	60	60	9 145	0	6	100.00%	99.93%	0.95
b_5_3	56	113	110	9 095	3	3	97.35%	99.97%	0.97
b_5_6	13	23	23	9 183	0	5	100.00%	99.95%	0.91
b_5_9	68	117	111	9 090	6	4	94.87%	99.96%	0.96
b_6_11	10	15	15	9 194	0	2	100.00%	99.98%	0.94
b_6_14	20	40	40	9 171	0	0	100.00%	100.00%	1.00
b_6_5	28	47	47	9 162	0	2	100.00%	99.98%	0.98
b_7_1	154	330	307	8 881	23	0	93.03%	100.00%	0.96

Continued									
Fold type	Number in LIFCA	Number in Astral1.65(95%)	t_p	t_n	f_n	f_p	Sensitivity	Specificity	MCC
b_8_1	40	83	83	9 126	0	2	100.00%	99.98%	0.99
b_8_14	21	36	36	9 149	0	26	100.00%	99.72%	0.76
b_8_7	99	193	190	9 011	3	7	98.45%	99.92%	0.97
b_8_8	32	63	63	9 139	0	9	100.00%	99.90%	0.93
b_9_1	52	514	514	8 676	0	21	100.00%	99.76%	0.98
b_9_2	24	45	44	9 162	1	4	97.78%	99.96%	0.95
c_4_13	27	45	45	9 162	0	4	100.00%	99.96%	0.96
c_4_2	27	57	53	9 150	4	4	92.98%	99.96%	0.93
c_4_7	12	19	19	9 192	0	0	100.00%	100.00%	1.00
c_5_11	11	19	18	9 184	1	8	94.74%	99.91%	0.81
c_5_12	30	47	46	9 145	1	19	97.87%	99.79%	0.83
c_5_16	101	151	118	9 019	33	41	78.15%	99.55%	0.76
c_5_23	16	34	32	9 168	2	9	94.12%	99.90%	0.86
c_5_24	37	73	73	9 130	0	8	100.00%	99.91%	0.95
c_5_25	14	16	16	9 195	0	0	100.00%	100.00%	1.00
c_5_28	97	160	156	9 038	4	13	97.50%	99.86%	0.95
c_5_3	12	20	20	9 163	0	28	100.00%	99.70%	0.64
c_6_13	10	14	14	9 192	0	5	100.00%	99.95%	0.86
c_6_14	15	24	21	9 184	3	3	87.50%	99.97%	0.87
c_6_34	79	138	133	9 038	5	35	96.38%	99.61%	0.87
c_7_13	34	53	53	9 158	0	0	100.00%	100.00%	1.00
c_7_22	26	38	34	9 169	4	4	89.47%	99.96%	0.89
c_7_8	12	17	17	9 194	0	0	100.00%	100.00%	1.00
c_8_13	156	307	284	8 874	23	30	92.51%	99.66%	0.91
Total(53)	1 826	3 857	Average				96.42%	99.91%	0.91

用于全库检验的检验集 a 共有 9 211 个样本, 属于实验集 53 种折叠类型的样本量为 3 857 个, 正确识别的为 3 695 个, 不属于实验集 53 种折叠类型的样本量为 5 354 个, 预测正确的为 5 021 个, 全库检验总体精度为 94.63%。所有折叠类型总体识别的敏感性为 96.42%, 特异性为 99.91%, MCC 值为 0.91。在全集检验中, 我们选取的检验集样本量较大, 但总体的识别结果较为理想, 这表明: 应用功能域组分方法建立的向量集合具有很好的覆盖

度和完整度, 相关性判定方法可以准确地对样本进行折叠类型识别。

2.2 独立检验集识别结果分析

为了评估本试验方法对涉及 SCOP 家族拆分样本的识别敏感性, 我们构建了独立检验集 b, 分别从不同层次进行了识别结果检验。表 2 给出了独立检验集 b 的识别结果统计, 分为总体、不涉及家族拆分样本、涉及家族拆分样本 3 个方面。

Table 2 Results of fold recognition on the independent test set b.

Test set	Sample number	Correctly classified	$Q/\%$
Independent test set b	2 035	1 922	94.45
Samples not involving family divided	1 859	1 781	95.80
Samples involving family divided	176	141	80.11

对于独立检验集 b, 总体的识别精度(94.45%)与全库的识别精度(94.63%)差别不大, 表明本文使用的方法对折叠类型之间的识别, 以及已知折叠类

型与未知折叠类型的识别都有很好的表现。但对于独立检验集 b 中涉及 SCOP 家族拆分的折叠类型, 总体的识别精度仅为 80.11%。说明虽然蛋白质间

拓扑结构不同, 功能上却是统一的, 但基于我们的分类, 功能域组分的方法并不能很好地对这部分样本进行识别.

2.3 基于不同数据库不同方法的折叠识别结果分析

SCOP 数据库是国际通用的蛋白质结构分类数据库, 它是由蛋白质的相似性和进化同源性分析得到的, 所以功能域组分方法也适用于 SCOP 折叠类型分类系统下的识别研究. 为比较 LIFCA 与 SCOP 对于功能域组分方法的敏感性, 我们针对检验集 b 中不涉及家族拆分的样本(1 859 个蛋白质), 基于

SCOP 分类, 用相同的方法进行了折叠识别研究, 涉及 SCOP 折叠类型共 107 种, 对应 LIFCA 折叠类型 51 种.

另外, 本文应用的功能域组分方法在国际上也有类似的工作, 如 Gewehr 等运用特异模式与类别的映射关系, 对 SCOP 数据库不同分类层次进行自动识别, 其中涉及的特异模式包含 Pfam 的功能域组分, 且占比重大. 表 3 给出了不同数据库分类下不同方法的识别结果.

Table 3 Results of fold recognition based on different method and different database

Method	Database	Average sensitivity	Average specificity	MCC
HMM ^[23-24]	LIFCA	90.36%	99.99%	0.95
Functional domain composition ¹⁾	LIFCA	96.42%	99.91%	0.91
Functional domain composition ²⁾	LIFCA	94.91%	99.97%	0.95
Functional domain composition ²⁾	SCOP	95.44%	99.99%	0.97
AutoSCOP ^[19]	SCOP	93.36%	98.13%	N/A

¹⁾ Functional domain composition: The result based on test set a.

²⁾ Functional domain composition: The result based on a part of samples in independent test set b, which do not involving family divided.

对于检验集 b 中不涉及家族拆分的样本, 涉及 LIFCA 折叠类型的数目是 SCOP 折叠类型数目的 47.66%, 从表 3 我们可以看出, 功能域组分方法在两种数据库分类下的总体敏感性、特异性、MCC 值区别并不大(0.53%、0.02%、0.02). 说明 LIFCA 分类规则虽然简单, 但却集中了蛋白质的大部分功能特性, 反映了结构与功能的对应关系.

针对 SCOP 折叠类型, 我们的方法总体敏感性、特异性、MCC 值分别为: 95.44%, 99.99%, 0.97, 整体略高于 Gewehr 等所用的运用特异模式与类别的映射关系进行自动识别的方法.

我们同时也比较了针对 LIFCA 分类, 本方法与之前所用 HMM 方法的优劣. 在前期研究中, 应用 HMM 对 LIFCA 中 36 种折叠类型进行了建模识别, 所得结果较本文结果在敏感性方面有明显差距, 表明了功能域组分这种方法虽然简单, 但却也能很好地用于折叠识别.

在我们的识别中, 由于样本量的限制, 只涉及了部分折叠类型, 但总体较高的识别结果足以表明本文使用方法的有效性.

3 结 论

本文应用功能域组分方法对 LIFCA 中样本量

较大的 53 种折叠类型进行了折叠识别研究, 由 LIFCA 中数据进行功能向量集合构建, 通过不同检验集对结果进行了评估, 在全库检验与独立检验集检验中均获得了较好地识别结果. 但我们也发现, 对于涉及 SCOP 家族拆分的样本, 我们的方法尚不能很好地进行识别. 但总体而言, 文中使用的功能域组分向量集构建以及折叠类型判定方法可以很好地应用在折叠识别中.

参 考 文 献

- [1] Burley S K. An overview of structural genomics. *Nature Structural Biology*, 2000, **7**(Suppl.): 932-934
- [2] Baker D, Sali A. Protein structure prediction and structural genomics. *Science*, 2001, **294**(5): 93-96
- [3] Murzin A G, Brenner S E, Hubbard T, *et al.* SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Molecular Biology*, 1995, **247**(4): 536-540
- [4] Lo Conte L, Brenner S E, Hubbard T, *et al.* SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Research*, 2002, **30**(1): 264-267
- [5] Orengo C A, Michie A D, Jones S, *et al.* CATH: A hierarchic classification of protein domain structures. *Structure*, 1997, **5**(8): 1093-1108
- [6] 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**(1): D247–D251
- [7] Novotny M, Madsen D, Kleywegt G J. Evaluation of protein fold comparison servers. *Proteins*, 2004, **54**(2): 260–270
- [8] Matsuda K, Nishioka T, Kinoshita K. Finding evolutionary relations beyond superfamilies: Fold-based superfamilies. *Protein Science*, 2003, **12** (10): 2239–2251
- [9] Chen Y, Crippen G M. Fold recognition *via* a tree. *J Computational Biology*, 2006, **13**(9): 1565–1573
- [10] Dubchak I, Muchnik I, Holbrook S R, *et al.* Prediction of protein folding class using global description of amino acid sequence. *Proc Natl Acad Sci USA*, 1995, **92**(19): 8700–8704
- [11] Dubchak I, Muchnik I, Mayor C, *et al.* Recognition of a protein fold in the context of the SCOP classification. *Proteins: Structure, Function and Genetics*, 1999, **35**(4): 401–407
- [12] Ding C H, Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, 2001, **17**(4): 349–358
- [13] 施建宇, 潘 泉, 张绍武, 等. 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, Zhang S W, *et al.* *Prog Biochem Biophys*, 2006, **33**(2): 155–162
- [14] Jain P, Garibaldi J M, Hirst J D. Supervised machine learning algorithms for protein structure classification. *Computational Biology and Chemistry*, 2009, **33**(3): 216 – 223
- [15] Marsolo K, Parthasarathy S, Ding C. A multi-level approach to SCOP fold recognition. *Proceedings of the Fifth IEEE Symposium on Bioinformatics and Bioengineering*. Washington, DC: IEEE Computer Society, 2005: 57–64
- [16] Marsolo K, Parthasarathy S. Alternate representation of distance matrices for characterization of protein structure. *Proceedings of the Fifth IEEE International Conference on Data Mining*. Washington, DC: IEEE Computer Society, 2005: 298–305
- [17] 施建宇, 张艳宁. 使用图像特征构建快速有效的蛋白质折叠识别方法. *生物物理学报*, 2009, **25**(2): 106–116
- Shi J Y, Zhang Y N. *Acta Biophysica sinica*, 2009, **25**(2): 106–116
- [18] Shen H B, Chou K C. Predicting protein fold pattern with functional domain and sequential evolution information. *J Theoretical Biology*, 2009, **256**(3): 441–446
- [19] Gewehr J E, Hintermair V, Zimmer R. AutoSCOP: automated prediction of SCOP classifications using unique pattern-class mappings. *Bioinformatics*, 2007, **23**(10): 1203–1210
- [20] Luo L F, Li X Q. Recognition and architecture of the framework structure of protein. *Proteins*, 2000, **39**(1): 9
- [21] 刘晓辉, 李晓琴, 徐海松, 等. 构建基于折叠核心的全 α 类蛋白取代矩阵. *中国生物化学与分子生物学报*, 2008, **24**(8): 761
- Liu X H, Li X Q, Xu H S, *et al.* *Chin J Biochem Mol Biol*, 2008, **24**(8): 761
- [22] 张 玮, 李晓琴, 徐海松, 等. 蛋白质折叠类型识别方法研究. *生物物理学报*, 2008, **24**(1): 65
- Zhang W, Li X Q, Xu H S, *et al.* *Acta Biophysica Sinica*, 2008, **24**(1): 65
- [23] 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
- Ren W K, Xu H S, Li X Q. *Prog Biochem Biophys*, 2007, **35**(5): 548–554
- [24] 刘 岳, 李晓琴, 徐海松, 等. 蛋白质折叠类型的分类建模与识别. *物理化学学报*, 2009, **25**(12): 2558–2564
- Liu Y, Li X Q, Xu H S, *et al.* *Acta Physico-Chimica Sinica*, 2009, **25**(12): 2558–2564
- [25] Chou K C, Cai Y D. Predicting subcellular localization of proteins by hybridizing functional domain composition and pseudo-amino acid composition. *J Cell Biochem*, 2004, **91**(6): 1197–1203
- [26] Xu X C, Yu D, Fang W, *et al.* Prediction of peptidase category based on functional domain composition. *J Proteome Research*, 2008, **7**(10): 4521–4524
- [27] Chou K C, Cai Y D. Predicting protein structural class by functional domain composition. *Biochem Biophys Res Commun*, 2004, **321**(4): 1007–1009
- [28] Finn R D, Mistry J, Schuster-Bockler B, *et al.* Pfam: clans, web tools and services. *Nucleic Acids Res*, 2006, **34**(1): D247–D25

Protein Fold Recognition by Functional Domain Composition*

YAN Jin-Li, CHEN Zhi-Wei, XU Hai-Song, LI Xiao-Qin**

(Bioengineering Center, Beijing University of Technology, Beijing 100124, China)

Abstract Research of protein 3D structures plays a key role in molecular biology, cell biology, biomedicine, and drug design. The protein fold type reflects the topological pattern of the structure's core. Fold recognition is an important method in protein sequence-structure research. On the 53 fold types which have more than 10 samples in LIFCA were selected. The functional domain composition is introduced to predict the fold types of a protein or a domain. After testing 9 211 proteins with less than 95% sequence identity from the Astral 1.65 database, the average sensitivity, specificity and Matthew's correlation coefficient (*MCC*) of the 53 fold types were found to be 96.42%, 99.91% and 0.91, respectively. The result indicates that using the functional domain composition to represent a protein is very promising for protein fold recognition. And though based on simple classification rules, LIFCA can concentrate the functional features of proteins, reflecting the corresponding relation between structure and function.

Key words fold type, fold recognition, functional domain, LIFCA

DOI: 10.3724/SP.J.1206.2010.00345

*This work was supported by grants from The National Natural Science Foundation of China (30570427) and Natural Science Foundation of Beijing (4092008).

**Corresponding author.

Tel: 86-13552340552, E-mail: lxq0811@bjut.edu.cn

Received: September 13, 2010 Accepted: November 5, 2010