doi:10.3969/j.issn.1672-5565.2014.02.08

α/β 类蛋白质折叠类型的分类方法研究

马 帅,王 勤,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:蛋白质折叠规律的研究是生命科学重大前沿课题之一,折叠分类是蛋白质折叠研究的基础。本文基于 LIFCA 数据库,选取样本量大于 2 的 55 种 α/β 类蛋白质折叠类型为研究对象。结合蛋白质折叠类型的定义及其保守拓扑结构特征,确定了 55 种蛋白质折叠类型的模板及其对应的特征参数。建立了基于模板的打分函数 Mul-Fscore,并结合二级结构参数信息,给出了 55 种 α/β 类蛋白质折叠类型的多模板分类方法。用此方法对 LIFAC 数据库中的 931 个样本进行检验,分类结果的平均特异性、平均敏感性、MCC 值分别为 99.58%、79.47%、79.39%。与 TM-score 分类结果对比发现,Mul-Fscore 分类的敏感性与 MCC 值好于 TM-score 的相应结果,平均特异性相近。

关键词:α/β 类蛋白质;折叠类型分类;打分函数;LIFCA 数据库
中图分类号:Q518.2 **文献标志码**:A **文章编号**:1672-5565(2014)-02-123-10

Research on the classification method of α/β protein fold type

MA Shuai, WANG Qin, LI Xiaoqin*
(*School of Life Science and Bioengineering, Beijing University of Technology, Beijing* 100124, *China*)

Abstract: The research of protein folding pattern is one of the major frontier subjects in life science, and folding is the basis of protein classification. Based on the LIFCA database, we selected research objects as 55 folding types of α/β, whose sample sizes are larger than 2. Combining with the definition of protein folding and its conservative topology characteristics, we determined the templates of the 55 folding types as well as their corresponding characteristic parameters. Based on the templates, we built a scoring function: Mul-Fscore. In LIFAC database, based on testing of selected 931 proteins, the average specificity, sensitivity, and MCC values are 99.58%, 79.47% and 79.39%, respectively. Compared with TM-score, we found that the sensitivity and MCC values of Mul-Fscore are slightly better, while the average specificity is quite similar.

Keywords: Protein of α/β; Folding type classification; Scoring function; LIFCA database

蛋白质折叠类型分类是蛋白质分类研究的重要问题,折叠类型分类反映了蛋白质核心结构的拓扑模式[1-4],包括蛋白质空间结构的二级结构单元、二级结构单元的相对排布位置以及蛋白质中肽链的走向三个主要方面。对蛋白质折叠类型进行系统化研究将为蛋白质相互作用界面、蛋白质的柔性和动态特性、远同源检测、酶和蛋白质的功能分类和预测等提供依据[5-6],对于相关蛋白质的设计和折叠动力学实验具有指导作用,并为精确的蛋白质三级结构预测奠定理论基础。

目前,SCOP、CATH 是两个重要的蛋白质结构分类数据库[7-8]。SCOP 是 Structural Classification of Proteins Database (蛋白质结构分类库) 的简称,是体现进化关系和结构相关性的分层次的结构分类数据库,其中,蛋白样本的折叠类型(fold)由专家凭经验指定。CATH 数据库是一个根据同源建模进行蛋白质结构分类的半自动分类数据库,其 4 个字母分别代表类型(Class)、构架(Architecture)、拓扑结构

收稿日期:2014-02-03;**修回日期**:2014-03-27.
基金项目:北京自然科技基金(4112010)资助。
作者简介:马帅,女,硕士,研究方向:生物信息学;E-mail:mashuai@emails.bjut.edu.cn.
*通信作者:李晓琴,女,教授,研究方向:生物信息学;E-mail:lxq0811@bjut.edu.cn.

(Topology)和同源性(Homologous),其中构架层次由人工指定。随着 PDB 数据库中晶体衍射和核磁共振解析的蛋白质结构数据的不断增加,SCOP 数据库已更新到 1.75C 版本[9],包含 1 194 种折叠类型,其中,全 α、全 β 以及 α/β 类蛋白质所对应的折叠类型数量分别为 284、174 和 147。CATH 数据库也更新到 v3.5 版本[10],包含 1 313 种折叠类型,全 α、全 β 以及 α/β 类蛋白质所对应的折叠类型数量分别为 386、299 和 594。可见,两者的分类并不相同。

在前期蛋白质折叠分类的研究中,建立了人工分类数据库——LIFCA 数据库[11]。LIFCA 数据库选自 ASTRAL-1.65 数据库中序列相似性小于 25% 的非冗余子集,与 SCOP 和 CATH 数据库不同,它是以参与折叠核心部分形成的规则结构片段组成、空间排布和拓扑连接为依据,建立起的一种基于统一原理的低冗余蛋白质折叠类型分类数据库。目前,LIFCA 数据库包含 α 类蛋白质的 44 种折叠类型,β 类蛋白质的 70 种折叠类型,α/β 类蛋白质的 145 种折叠类型。在实验室的前期研究工作中,针对 LIFCA 数据库中典型的蛋白质折叠类型,建立了基于 HMM 模型和功能域组分的识别方法[12-14],识别方法的敏感性、特异性均在 90%以上。蛋白质折叠类型识别方法的成功进一步证明:尽管 LIFCA 数据库的分类规则非常简单,却抓住了折叠类型分类的关键因素,也反映了蛋白质结构与功能的对应关系。对典型的蛋白质折叠类型建立自动的分类方法,是目前迫切需要解决的问题。

基于现有数据库中已知的折叠类型,预测蛋白质结构的问题实质上就转化为分类的问题,即折叠识别问题(Fold Recognition)[15],就是将待测蛋白与已知蛋白进行结构的相似性比对,从而判断待测蛋白的折叠类型。从实验蛋白质结构检测到基于计算机的蛋白质折叠与结构预测、从蛋白质拓扑分类到基于结构的蛋白质功能标注、从蛋白质匹配比对到新化合物筛选与药物设计,蛋白质结构比对几乎在现代结构生物学的每一个方面都很重要[16]。

基于已知空间结构的蛋白质折叠类型分类,需要通过对蛋白质结构相似性的评判来完成。目前,蛋白质结构相似性的比较通常用 RMSD[17]、GDT-score、MaxSub、TM-score 来判定。1999 年 Zemla 提出了 GDT-score[18],旨在待测蛋白与已知蛋白的结构处于最佳叠合状态时精确地、非连续地识别待测蛋白质的最大子结构。在一些不同的截断距离下计算 GDT-score,打分值通常随着截断距离的增大而增大。2000 年 Siew 提了 MaxSub 方法,它能够进一步构建和扩展一些评估蛋白质结构相似性的方法[19]。MaxSub 与 GDT-score 有着相似的原理,旨在当已知蛋白与待测蛋白处于最佳叠合状态时,识别 C_α 原子的最大子结构,并且产生一个单一的标准化打分。由于 GDT-score 和 MaxSub 中截断距离是主观设定的,而且对于不同模板蛋白质的分类需要做手动的调整,再者,类似于 RMSD,对于待测蛋白,GDT-score 和 MaxSub 的打分值大小与蛋白质的长度有幂率的依赖[20]。为了解决这些问题,张扬和 SKolnic[21-22] 在 2004 年提出了 TM-score,旨在计算待测蛋白与已知蛋白的所有对应残基之间的距离并产生一个标准化打分,克服了打分值与蛋白质大小的幂率依赖,是对蛋白质结构相似性的全自动评估的有效补充[23]。

本文对 LIFCA 数据库中 α/β 类蛋白的 55 种典型折叠类型,提取模板信息及模板特征参数,建立 α/β 类蛋白质折叠类型多模板打分函数 Mul-Fscore,并结合二级结构参数信息,给出了 55 种 α/β 类蛋白质折叠类型的多模板分类方法。

2 材料的选择

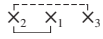
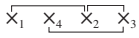
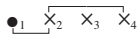
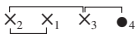
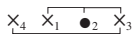
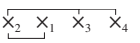
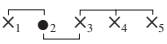
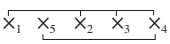
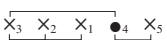
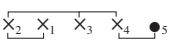
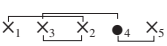
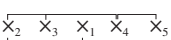
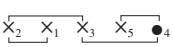
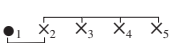
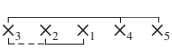
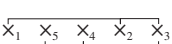
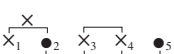
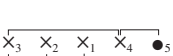
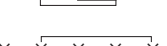
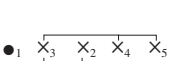
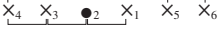
2.1 实验集及模板蛋白的选择

为了保证分类方法的适普性与精确性,用来做检验的实验集样本数目越多越好。与此同时,为了能与已有的研究结果比较,要尽可能选择多数研究者共同使用的数据集。根据上述要求选取了 LIFCA 数据库中样本量大于 2 的 55 种 α/β 类蛋白质的折叠类型,共得到 931 个样本作为实验集。

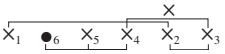
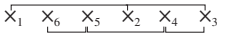
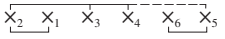
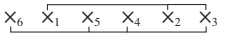
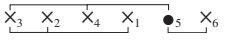
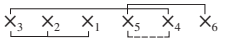
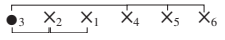
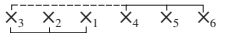
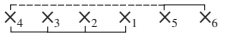
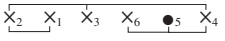
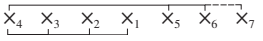
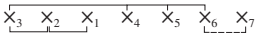
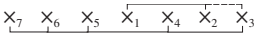
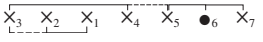
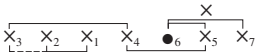
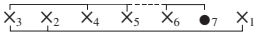
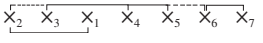
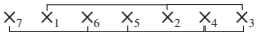
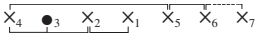
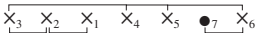
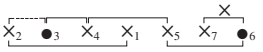
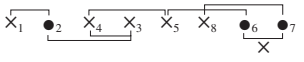
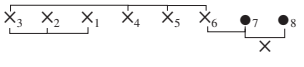
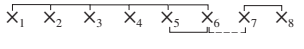
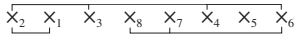
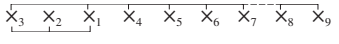
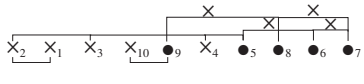
模板选取严格遵守以下原则:一种蛋白质折叠类型,选取一个模版,模板具有折叠类型特有的保守拓扑结构,且结构冗余最少,无氨基酸残基缺失。55 种 α/β 类蛋白质折叠类型的 931 个样本中,选取符合要求的 55 个模板蛋白(见表 1)。其中 Fold type 为折叠类型,Name 为每一个折叠类型的模板蛋白名称,Number of samples 为每一个折叠类型中样本的个数,Trend and sequence of β strand 为模板蛋白核心结构中 β 片段的走向和顺序,Topology of the template 为模板蛋白的空间向量展开图。拓扑图中,并行排列的"×"为方向指向里的 β 片段,"•"为方向指向外的 β 片段,相应下面的阿拉伯数字为 β 片段的连接顺序;粗实线表示两个 β 片段由 α 螺旋连接,虚实线表示两个 β 片段由 3_{10}-螺旋连接,细实线表示两个 β 片段由无规则卷曲 c 连接;连接线上方的"×"表示方向指向里的 α 螺旋或者 3_{10}-螺旋,连接线上方无"×"表示方向指向外的 α 螺旋或者 3_{10}-螺旋。

表 1　模板蛋白折叠类型信息

Table 1　Folding type information of template proteins

Fold type	Names	Number of samples	Trend and sequence of β strand	Topology of thetemplate
c_3_2	d1b6ra2	4	PPP/213	
c_4_2	d2gsq_2	29	PPNP/2134	
c_4_7	d1mkp__	12	PPPP/1423	
c_4_8	d1m0da_	4	NPPP/1234	
c_4_10	d1k2yx2	4	PPPN/2134	
c_4_12	d1sfe_2	3	PPNP/4123	
c_4_13	d1iiba_	28	PPPP/2134	
c_5_1	d1d02a_	4	PNPPP/12345	
c_5_3	d1lvl_2	16	PPPPP/15234	
c_5_7	d1knga_	7	PPPNP/32145	
c_5_9	d1g5ca_	5	PPPPN/21345	
c_5_11	dlerv__	11	PPPNP/13245	
c_5_12	d1g8fa3	32	PPPPP/23145	
c_5_13	d1cfza_	4	PPPPN/21354	
c_5_14	d1h4xa_	3	NPPPP/12345	
c_5_16	d2uagal	103	PPPPP/32145	
c_5_19	d1e0cal	7	PPPPP/15423	
c_5_20	d1a79al	7	PNPPN/12345	
c_5_21	d1e0ta3	5	PPPPN/32145	
c_5_22	d1leha2	5	PNPPP/12435	
c_5_23	d1iqpa2	17	PPPPP/51432	
c_5_24	d1fjgk_	37	PNPPP/32145	
c_5_25	d1h4vbl	14	PPPNP/21345	
c_5_28	d1tmy__	102	PPPPP/21345	
c_5_31	d1ik6a2	4	NPPPP/13245	
c_6_2	d1chmal	4	PPNPPP/432156	
c_6_5	d2uaga2	3	NPPPPP/126345	

续(表 1)

Fold type	Names	Number of samples	Trend and sequence of β strand	Topology of thetemplate
c_6_6	d3cla__	4	PNPPPN/165423	
c_6_13	d1c4oa2	11	PPPPPP/165243	
c_6_14	d1poxa2	16	PPPPPP/213465	
c_6_16	d1g5ta_	8	PPPPPP/615423	
c_6_20	d1bif_2	7	PPPPNP/324156	
c_6_27	d1ipaal	4	PPPPPP/321546	
c_6_29	d1atza_	6	NPPPPP/321456	
c_6_34	d1lldal	79	PPPPPP/321456	
c_6_35	d1gvha3	5	PPPPPP/432156	
c_6_37	d1g8la3	3	PPPPNP/213654	
c_7_2	d1k6da_	3	PPPPPPP/4321567	
c_7_4	d1XCdal	6	PPPPPPP/3214567	
c_7_6	d1g3qa_	8	PPPPPPP/7651423	
c_7_7	d1zpdal	7	PPPPPNP/3214567	
c_7_8	d1e5ka_	13	PPPPNPP/3214657	
c_7_13	d1lc5a_	34	PPPPPNP/3245671	
c_7_14	d2uaga3	7	PPPPPPP/2314567	
c_7_16	d1qdea_	6	PPPPPPP/7165243	
c_7_18	d1edzal	6	PNPPPPP/4321567	
c_7_22	d1ej0a_	26	PPPPPNP/3214576	
c_7_23	d2pth__	9	PNPPPPN/2341576	
c_8_6	d1qq9a_	7	PNPPPPNN/12435867	
c_8_9	d1i36a2	6	PPPPPPNN/32145678	
c_8_13	d1pii_2	157	PPPPPPPP/12345678	
c_8_21	d1d3va_	3	PPPPPPPP/21387456	
c_9_1	d1rkd_	5	PPPPPPPNP/321456789	
c_10_10	d1cnza_	3	PPPPNPNNNN/21310945867	

2.2　模板特征参数的提取

α/β 类蛋白的标准二级结构序列为 $\beta_1\alpha_1\beta_2\alpha_2\cdots\beta_i\alpha_i$,其中 $\beta_i\alpha_i(i=1\cdots n)$ 及 $\alpha_{\beta i+1}(i=1\cdots n)$ 均通过无规卷曲片段 c 连接,其标准的二级结构序列可以看成由基本单元 $\beta_i\alpha_i\beta_{i+1}(i=1\cdots n)$ 顺序连接而成(β_{i+1}为两个单元共同所用)。基于前期提出的蛋白质折叠类型分类方法,结合 α/β 类蛋白折叠类型的拓扑特征,并考虑到 α/β 类蛋白质核心结构单元 $\beta_i\alpha_i\beta_{i+1}$ 满足右手连接关系原则,确定以形成折叠类型核心的规则 β-Sheet 的数量及其对应的长度为模板特征参数,特征参数取自 DSSP[22] 数据库的相应蛋白质条目。

DSSP 数据库中,标注了 8 种二级结构类型:H(α-螺旋)、G(3_{10}-螺旋)、I(π-螺旋)、B(β 桥)、E(β 股)、T(转角)、S(卷曲)和空白(环或无规则卷曲),其中连续 2 个及 2 个以上 E 定为 1 个 β-sheet,连续 3 个及 3 个以上 H 定为 1 个 α-螺旋,连续 3 个及 3 个以上 G 定为 1 个 3_{10}-螺旋,β-Sheet、α-螺旋、3_{10}-螺旋等规则二级结构之间的连续部分记为 C。对模板蛋白 j,提取 β 片段在序列中的位置信息,形成模板特征参数见表 2,其中 Fold type 为折叠类型,每一折叠类型对应的上行数据为组成核心结构的 β-Sheet 片段的起止序号,下行数据为该 β-Sheet 片段的长度;对任意待测蛋白,提取二级结构片段 β-sheet、α-螺旋、3_{10}-螺旋,形成由 β(β-sheet)、α(α-螺旋、3_{10}-螺旋)组成的二级结构序列,记作:$A_1A_2\cdots A_i\cdots A_n$($A=\beta$ or α),其中三连体 βαβ 的出现的频次记为 $N_{\beta\alpha\beta}$。

表 2　55 种折叠类型的模板特征参数

Table 2　Characteristic parameters of 55 folding type

Fold type	1	2	3	4	5	6	7	8	9	10
c_3_2	2~7 6	25~30 6	45~48 4							
c_4_2	3~7 5	29~32 4	52~55 4	58~60 3						
c_4_7	11~14 4	30~35 6	50~54 5	85~89 5						
c_4_8	40~41 2	47~51 5	74~78 5	104~106 3						
c_4_10	20~24 5	44~48 5	83~87 5	94~98 5						
c_4_12	4~10 7	13~19 7	24~30 7	48~49 2						
c _4_13	2~8 7	32~38 7	51~54 4	74~76 3						
c_5_1	35~39 5	79~83 5	89~96 8	143~148 6	172~175 4					
c_5_3	8~9 2	23~27 5	46~50 5	78~80 3	110~112 3					
c_5_7	34~39 6	63~69 7	89~93 5	109~113 5	119~124 6					
c_5_9	26~31 6	52~56 5	80~86 7	149~155 7	162~166 5					
c_5_11	3~5 3	23~28 6	53~58 6	76~81 6	84~90 7					
c_5_12	7~11 5	49~41 3	63~66 4	83~87 5	94~95 2					
c_5_13	3~8 6	36~42 7	57~63 7	74~78 5	113~119 7					
c_5_14	2~8 7	11~20 10	43~54 12	75~79 5	98~99 2					
c_5_16	8~11 4	32~35 4	52~54 3	67~70 4	90~92 3					
c_5_19	10~11 2	25~29 5	45~46 2	83~87 5	111~113 3					
c_5_20	19~22 4	30~33 4	46~52 7	75~81 7	87~90 4					
c_5_21	21~24 4	43~47 5	62~66 5	95~100 6	111~116 6					
c_5_22	14~20 7	25~33 9	43~46 4	77~84 8	110~111 2					
c_5_23	47~51 5	78~82 5	111~116 6	140~146 7	160~164 5					
c_5_24	6~12 7	19~23 5	29~34 6	69~75 7	94~100 7					

续(表 2)

Fold type	1	2	3	4	5	6	7	8	9	10
c_5_25	6~10 5	32~34 3	55~59 5	68~73 6	79~83 5					
c_5_28	3~7 5	27~32 6	49~53 5	77~81 5	99~102 4					
c_5_31	10~13 4	18~22 5	41~45 5	67~74 8	100~104 5					
c_6_2	41~44 4	66~69 4	74~79 6	94~98 5	119~122 4	143~146 4				
c_6_5	7~12 6	15~19 5	42~47 6	67~72 6	87~89 3	109~112 4				
c_6_6	27~35 9	84~91 8	96~101 6	138~144 7	166~170 5	178~188 11				
c_6_13	7~11 5	33~37 5	58~61 4	84~87 4	100~105 6	134~138 5				
c_6_14	20~23 4	44~47 4	69~73 5	96~102 7	128~131 4	155~161 7				
c_6_16	5~9 5	33~37 5	57~60 4	97~101 5	129~133 5	147~150 4				
c_6_20	2~6 5	48~51 4	69~70 2	136~140 5	169~175 7	180~186 7				
c_6_27	12~17 6	39~43 5	66~69 4	85~88 4	105~109 5	126~128 3				
c_6_29	3~10 8	41~48 8	52~56 5	106~113 8	133~140 8	162~164 3				
c_6_34	3~7 5	28~32 5	57~61 5	72~75 4	113~116 4	139~141 3				
c_6_35	9~14 6	37~43 7	65~71 7	86~87 2	105~109 5	132~135 4				
c_6_37	2~8 7	42~49 8	69~72 4	93~98 6	106~111 6	116~119 4				
c_7_2	20~23 4	47~50 4	70~75 6	94~97 4	148~159 12	181~192 12	211~213 3			
c_7_4	46~50 5	69~73 5	85~86 2	99~102 4	126~130 5	151~156 6	180~182 3			
c_7_6	3~8 6	34~38 5	81~84 4	114~118 5	136~141 6	164~173 10	193~198 6			
c_7_7	25~29 5	53~56 4	70~74 5	92~96 5	117~120 4	124~127 4	130~133 4			
c_7_8	4~9 6	47~50 4	65~66 2	91~96 6	119~123 5	128~136 9	164~167 4			
c_7_13	77~80 4	98~103 6	119~124 6	147~151 5	179~183 5	206~211 6	224~227 4			
c_7_14	13~17 5	40~44 5	60~64 5	80~83 4	114~118 5	130~133 4	140~143 4			
c_7_16	50~53 4	81~84 4	109~112 4	121~124 4	145~149 5	175~180 6	199~201 3			
c_7_18	31~35 5	55~59 5	63~68 6	79~85 7	100~103 4	121~125 5	140~144 5			
c_7_22	25~29 5	49~54 6	64~68 5	89~94 6	126~137 12	153~159 7	170~178 9			
c_7_23	4~7 4	39~41 3	46~53 8	56~63 8	88~94 7	102~106 5	130~135 6			
c_8_6	51~59 9	62~70 9	75~85 11	121~128 8	153~159 7	170~171 2	194~195 2	215~219 5		
c_8_9	2~6 5	25~28 4	47~48 2	59~62 4	83~86 4	106~112 7	124~128 5	144~147 4		
c_8_11	2~10 9	41~44 4	60~63 4	77~78 2	96~100 5	114~123 10	138~145 8	150~156 7		
c_8_13	4~5 2	24~28 5	52~57 6	76~79 4	100~106 7	121~125 5	149~151 3	169~172 4		
c_8_21	3~7 5	42~47 6	88~92 5	114~118 5	169~174 6	191~194 4	222~227 6	265~271 7		
c_9_1	3~7 5	54~62 9	84~87 4	134~137 4	159~162 4	180~181 2	216~220 5	226~230 5	233~237 5	
c_10_10	5~12 8	38~42 5	70~75 6	106~112 7	133~139 7	189~194 6	220~225 6	242~245 4	268~272 5	278~282 5

3　方　法

3.1　打分函数的建立

利用 TM-align 结构比对软件,将已知空间结构信息的待测蛋白质与模板 j 进行比对,得到待测蛋白与模板蛋白 j 的结构比对信息。定义待测蛋白与模板 j 的打分函数 Mul-Fscore 为:

$$\text{Mul-Fscore}_j = \frac{1}{N_j}\sum_{j=1}^{N_j}\left(\frac{L_{ij}}{L_{ijo}}\right)^2 。\quad (1)$$

其中,Mul-Fscore_j 表示待测蛋白与第 j 个模板的比对打分值,N_j 表示模板 j 中形成折叠核心部分的 β-Sheet 片段的个数,L_{ijo} 表示模板 j 的第 i 个规则 β-Sheet 片段的残基数,L_{ij} 表示与模板 j 的第 i 规则 β-Sheet 片段匹配上的待测蛋白的残基数。其中,$i=1,2\cdots N$,$j=1,2\cdots 55$。

3.2　Mul-Fscore 分类方法

对模板 $j(j=1\cdots 55)$,形成折叠核心的规则 β-Sheet 的数量为 N_j,其中平行 β-Sheet 的数量为 N'_j(即表 1 中模板空间向量展开图中"×"的数量)。对任意给定的已知空间结构的 α/β 类蛋白,利用 1.2 节给出的方法得到 $N_{\beta\alpha\beta}$,该蛋白折叠类型的归属通过以下两步完成:

第一步:待测蛋白候选模板的确定。当待测蛋白的 $N_{\beta\alpha\beta}$ 小于等于模板 j 对应的 N'_j 时,则模板 j 为待测蛋白折叠类型的候选模板。

第二步:待测蛋白折叠类型的确定。计算待测蛋白与候选模板的 Mul-Fscore 取值,最大取值对应模板的所属折叠类型即为该样本所属折叠类型,若出现多个模板 Mul-Fscore 取值最大且相同时,则该样本属于 N_j 最大的模板所对应的折叠类型。

3.3　评估参数

本文为检验 Mul-Fscore 多模板打分分类方法的效果,利用敏感性(Sensitivity)、特异性(Specificity)和 Matthew 相关系数三个指标进行评估,参数定义如下:

敏感性:

$$S_n = \frac{t_p}{t_p+f_n}\times 100\%。\quad (2)$$

特异性:

$$S_p = \frac{t_n}{t_n+f_p}\times 100\%。\quad (3)$$

Matthew 相关系数:

$$MCC = \frac{(t_p\times t_n)-(f_p\times f_n)}{\sqrt{(t_p+f_n)\times(t_n+f_p)\times(t_p+f_p)\times(t_n+f_n)}}。\quad (4)$$

其中,t_p 为真阳性个数,t_n 为真阴性个数,f_p 为假阳性个数,f_n 为假阴性个数。

4　结　果

模板信息数据库在 LIFCA 数据库中建立,样本量大于 2 的 α/β 类蛋白质折叠类型有 55 种,共计 931 个蛋白质样本。为了检验 Mul-Fscore 打分方法对多模板识别的准确性,对 931 个样本利用2.2 节给出的方法确定所属折叠类型,计算 55 种折叠类型样本分类的敏感性、特异性及 Matthew 相关系数,同时与 TM-score 的结果进行比较(见表 3)。表 3 中,Number 为 55 种折叠类型的序号;Fold type 为折叠类型;Name 为 55 种折叠类型的名称;S(S′)为该折叠类型的样本数量(931 个样本中非该种折叠类型的样本数量);标有"′"的为 TM-score 的结果,未标有"′"的为 Mul-Fscore 的结果。

表 3 中第 7 和 10 列的前组数据为两种方法的敏感性检验结果。统计显示,Mul-Fscore 分类方法的敏感性检验结果大于等于 TM-score 敏感性结果的有 41 个,占 55 种折叠类型的 75%,Mul-Fscore 的平均敏感性结果高于 TM-score 结果 1.11%,说明对于同一种折叠类型的样本,Mul-Fscore 较 TM-score 有更高的辨识力。

表 3 中第 7 和 10 列括号中的数据为特异性检验结果。数据统计,Mul-Fscore 分类方法的平均特异性结果为 99.58%,TM-score 的平均特异性结果为 99.55%,两个结果非常相近。说明本文的方法与 TM-score 对于区分不同折叠类型有相似的能力。

表 3 中最后一列为 Matthew 相关系数的结果统计,55 组数据中,Mul-Fscore 方法计算出的 MCC 值不小于 TM-score 的占 67%,其中 MCC 值高出 TM-score 方法 20%以上的占 27%。

表 3 敏感性、特异性及 MCC 值的比较

Table 3 The comparison of sensitivity, specificity and MCC value

Number	Fold type	Name	$S(S')$	$t_p(t_n)$	$f_n(f_p)$	$S_n(S_p)\%$	$t'_p(t'_n)$	$f'_n(f'_p)$	$S'_n(S'_P)\%$	MCC(MCC′)
1	c.3.2	d1b6ra2	4 (927)	2 (907)	2 (20)	50 (97.84)	2 (906)	2 (21)	50 (97.73)	0.21 (0.20)
2	c.4.2	d2gsq_2	29 (902)	29 (895)	0 (7)	100 (99.22)	26 (901)	3 (0)	89.66 (100)	0.89 (0.95)
3	c.4.7	d1mkp__	12 (909)	12 (907)	0 (2)	100 (99.78)	12 (909)	0 (0)	100 (100)	0.92 (1)
4	c.4.8	d1m0da_	4 (927)	4 (920)	0 (7)	100 (99.24)	1 (927)	3 (0)	25 (100)	0.60 (0.50)
5	c.4.10	d1k2yx2	4 (927)	4 (918)	0 (9)	100 (99.03)	3 (927)	1 (0)	75 (100)	0.55 (0.87)
6	c.4.12	d1sfe_2	3 (928)	3 (913)	0 (15)	100 (98.38)	3 (923)	0 (5)	100 (99.46)	0.40 (0.61)
7	c.4.13	d1iiba_	28 (903)	17 (856)	11 (47)	60.71 (94.80)	11 (876)	17 (27)	39 (97.01)	0.37 (0.31)
8	c.5.1	d1d02a_	4 (927)	3 (926)	1 (1)	75 (99.89)	2 (927)	2 (0)	50 (100)	0.75 (0.71)
9	c.5.3	d1lvl_2	16 (915)	16 (915)	0 (0)	100 (100)	16 (912)	0 (3)	100 (99.67)	1 (0.92)
10	c.5.7	d1knga_	7 (924)	6 (924)	1 (0)	85.71 (100)	6 (924)	1 (0)	85.71 (100)	0.93 (0.93)
11	c.5.9	d1g5ca_	5 (926)	4 (926)	1 (0)	80 (100)	4 (926)	1 (0)	80 (100)	0.89 (0.89)
12	c.5.11	dlerv__	11 (920)	7 (920)	4 (0)	63.64 (100)	11 (916)	0 (4)	100 (99.57)	0.80 (0.85)
13	c.5.12	d1g8fa3	32 (899)	22 (898)	10 (1)	68.75 (99.89)	20 (896)	12 (3)	62.50 (99.67)	0.81 (0.73)
14	c.5.13	d1cfza_	4 (927)	2 (926)	2 (1)	50 (99.89)	2 (926)	2 (1)	50 (99.89)	0.58 (0.58)
15	c.5.14	d1h4xa_	3 (928)	3 (928)	0 (0)	100 (100)	3 (918)	0 (10)	100 (98.92)	1 (0.48)
16	c.5.16	d2uagal	103 (828)	88 (793)	15 (35)	85.44 (95.77)	89 (772)	14 (56)	86.41 (93.24)	0.75 (0.69)
17	c.5.19	d1e0cal	7 (924)	7 (917)	0 (7)	100 (99.24)	7 (924)	0 (0)	100 (100)	0.70 (1)
18	c.5.20	d1a79al	7 (924)	6 (924)	1 (0)	85.71 (100)	6 (921)	1 (3)	85.71 (99.68)	0.93 (0.75)
19	c.5.21	d1e0ta3	5 (926)	4 (922)	1 (4)	80 (99.57)	4 (925)	1 (1)	80 (99.89)	0.63 (0.80)
20	c.5.22	d1leha2	5 (926)	5 (926)	0 (0)	100 (100)	5 (926)	0 (0)	100 (100)	1 (1)
21	c.5.23	d1iqpa2	17 (914)	13 (912)	4 (2)	76.47 (99.78)	6 (914)	11 (0)	35.29 (100)	0.81 (0.59)
22	c.5.24	d1fjgk_	37 (894)	14 (894)	23 (0)	37.84 (100)	30 (893)	7 (1)	81.08 (99.89)	0.61 (0.88)
23	c.5.25	d1h4vbl	14 (917)	9 (916)	5 (1)	64.28 (99.89)	13 (914)	1 (3)	92.86 (99.67)	0.76 (0.87)
24	c.5.28	d1tmy__	102 (829)	66 (807)	36 (22)	64.71 (97.35)	82 (789)	20 (40)	80.39 (95.17)	0.66 (0.70)
25	c.5.31	d1ik6a2	4 (927)	3 (925)	1 (2)	75 (99.78)	4 (914)	0 (13)	100 (98.60)	0.67 (0.48)
26	c.6.2	d1chmal	4 (927)	4 (925)	0 (2)	100 (99.78)	4 (927)	0 (0)	100 (100)	0.82 (1)
27	c.6.5	d2uaga2	3 (928)	3 (928)	0 (0)	100 (100)	3 (928)	0 (0)	100 (100)	1 (1)
28	c.6.6	d3cla__	4 (927)	4 (927)	0 (0)	100 (100)	4 (927)	0 (0)	100 (100)	1 (1)
29	c.6.13	d1c4oa2	11 (920)	10 (920)	1 (0)	90.91 (100)	7 (918)	4 (2)	63.64 (99.78)	0.95 (0.70)
30	c.6.14	d1poxa2	16 (915)	15 (914)	1 (1)	93.75 (99.89)	16 (915)	0 (0)	100 (100)	0.94 (1)
31	c.6.16	d1g5ta_	8 (923)	7 (923)	1 (0)	87.50 (100)	7 (915)	1 (8)	87.50 (99.13)	0.93 (0.64)
32	c.6.20	d1bif_2	7 (924)	4 (924)	3 (0)	57.14 (100)	7 (924)	0 (0)	100 (100)	0.75 (1)
33	c.6.27	d1ipaal	4 (927)	4 (927)	0 (0)	100 (100)	4 (927)	0 (0)	100 (100)	1 (1)
34	c.6.29	d1atza_	6 (925)	6 (925)	0 (0)	100 (100)	6 (923)	0 (2)	100 (99.78)	1 (0.87)
35	c.6.34	d1lldal	79 (852)	48 (839)	31 (13)	60.76 (98.47)	17 (851)	62 (1)	21.52 (99.88)	0.67 (0.43)
36	c.6.35	d1gvha3	5 (926)	3 (923)	2 (3)	60 (99.68)	3 (923)	2 (3)	60 (99.78)	0.55 (0.55)
37	c.6.37	d1g8la3	3 (928)	2 (926)	1 (2)	66.67 (99.78)	3 (927)	0 (0)	100 (100)	0.58 (1)
38	c.7.2	d1k6da_	3 (928)	3 (928)	0 (0)	100 (100)	3 (928)	0 (0)	100 (100)	1 (1)
39	c.7.4	d1XCdal	6 (925)	4 (925)	2 (0)	66.67 (100)	5 (923)	1 (2)	83.33 (99.78)	0.82 (0.77)
40	c.7.6	d1g3qa_	8 (923)	7 (922)	1 (1)	87.50 (99.89)	6 (923)	2 (0)	75 (100)	0.87 (0.87)
41	c.7.7	d1zpdal	7 (924)	3 (924)	4 (0)	42.86 (100)	2 (922)	5 (2)	28.57 (99.78)	0.65 (0.37)
42	c.7.8	d1e5ka_	13 (918)	9 (918)	4 (0)	69.23 (100)	13 (918)	0 (0)	100 (100)	0.83 (1)
43	c.7.13	d1lc5a_	34 (897)	34 (897)	0 (0)	100 (100)	34 (897)	0 (0)	100 (100)	1 (1)
44	c.7.14	d2uaga3	7 (924)	3 (924)	4 (0)	42.86 (100)	3 (924)	4 (0)	42.86 (100)	0.65 (0.65)
45	c.7.16	d1qdea_	6 (925)	5 (925)	1 (0)	83.33 (100)	3 (921)	3 (4)	50 (99.57)	0.91 (0.46)
46	c.7.18	d1edzal	6 (925)	3 (925)	3 (0)	50 (100)	4 (924)	2 (1)	66.67 (99.89)	0.71 (0.73)
47	c.7.22	d1ej0a_	26 (905)	7 (905)	19 (0)	26.92 (100)	13 (905)	13 (0)	50 (100)	0.51 (0.70)
48	c.7.23	d2pth__	9 (922)	9 (921)	0 (1)	100 (99.89)	1 (922)	8 (0)	12.50 (100)	0.95 (0.33)
49	c.8.6	d1qq9a_	7 (924)	5 (924)	2 (0)	71.43 (100)	3 (924)	4 (0)	42.86 (100)	0.84 (0.65)
50	c.8.9	d1i36a2	6 (925)	6 (925)	0 (0)	100 (100)	6 (922)	0 (3)	100 (99.68)	1 (0.82)
51	c.8.11	d1vdra_	8 (923)	8 (923)	0 (0)	100 (100)	8 (923)	0 (0)	100 (100)	1 (1)
52	c.8.13	d1pii_2	157 (774)	152 (774)	5 (0)	96.82 (100)	151 (774)	6 (0)	96.18 (100)	0.98 (0.98)
53	c.8.21	d1d3va_	3 (928)	2 (928)	1 (0)	66.67 (100)	3 (928)	0 (0)	100 (100)	0.81 (1)
54	c.9.1	d1rkd_	5 (926)	4 (926)	1 (0)	80 (100)	4 (926)	1 (0)	80 (100)	0.89 (0.89)
55	c.10.10	d1cnza_	3 (928)	2 (928)	1 (0)	66.67 (100)	3 (928)	0 (0)	100 (100)	0.82 (1)
Ave.						79.47 (99.58)			78.36 (99.55)	0.794 (0.776)

5　讨　论

Mul-Fscore 分类方法是在给出蛋白质折叠类型定义、并建立人工分类数据库及蛋白质折叠类型识别方法的基础上提出的,目的是实现基于模板的 α/β 类蛋白折叠类型自动分类。Mul-Fscore 函数通过提取参与折叠核心部分组成的匹配残基的数量,建立打分函数,并利用二级结构参数信息建立了 α/β 类蛋白折叠类型分类方法,获得了优于 TM-Fscore 的敏感性、相关性检验结果及相似的特异性检验结果。由于 Mul-Fscore 分类方法只考虑参与折叠核心的规则二级片断的匹配长度,相较于利用匹配残基空间结构信息的 TM-score 分类方法可以减少计算量。

在蛋白质折叠类型中,存在一种折叠类型结构与另一种折叠类型结构互为“子母”关系的现象,当出现这个问题时,仅仅依赖 Mul-Fscore 打分值将削减 α/β 类蛋白质折叠类型分类的准确性。因此,在进行 Mul-Fscore 分类方法中,增加了一个前处理:即待测蛋白候选模板确定。因为对于一个理想的 α/β 类蛋白样本,其 $N_{\beta\alpha\beta}$ 与 N'_j 的关系满足:$N'_j = N_{\beta\alpha\beta}+1$。增加前处理与没有前处理的分类方法相比,在 55 种折叠类型中,10 种折叠类型的敏感性结果有所提高,45 种折叠类型保持不变,整体的平均敏感性提高 3%;8 种折叠类型的特异性结果有所提高,45 种折叠类型保持不变,2 种折叠类型略有下降,整体平均特异性提高 0.04%。

虽然 Mul-Fscore 分类方法整体显现出较好的状态,但是从统计结果中依然发现个别折叠类型的统计指标不理想。表 3 中敏感性值较低的几个折叠类型编号分别为 12,22,23,32,41,47,对应的折叠类型分别为 c.5.11,c.5.24,c.5.25,c.6.20,c.7.7,c.7.22。分析原因是在这些折叠类型中,反平行 β-Sheet 片段与左右相邻的 β-Sheet 片段之间不是通过 α 螺旋连接,而是由无规则卷曲直接相连,再加上有些折叠类型模板中相邻平行 β-Sheet 片段之间也直接通过卷曲连接,这样统计出的该折叠类型所属样本的 βαβ 个数会比理论数目少很多,因此不能克服折叠类型之间的“子母”关系,造成错误分类的出现,而且这种错误分类的出现不能通过制定打分规则来人为地避免。

对 LIFAC 数据库中全 α、全 β 结构类中的任一蛋白样本,分别与 55 种模板进行比对,并计算 Mul-Fscore 取值(见表 4)。从表 4 可以看出:全 α 及全 β 结构类的 1380 个蛋白样本。Mul-Fscore 取值分布在(0,0.6]、(0,0.7]及(0,0.8]的样本占总体的比例分别为 97.55%,99.42% 和 99.89%。而将 931 个实验样本与其所属折叠类型的模板进行比对,并计算 Mul-Fscore 取值,取值分布在(0,0.6]、(0,0.7]及(0,0.8]的样本占总体的比例分别为2.79%、5.59%和 10.63%,即利用 Mul-Fscore 可以将不属于模板范围的全 α 类及全 β 类蛋白样本排除。

表 4　Mul-Fscore 的排异性检验

Table 4　The rejection-inspection of Mul-Fscore

Type	Numbers	Score<0.6	Score<0.7	Score<0.8
全 α	550	99.79% *	99.98% *	100% *
全 β	830	96.03% *	99.04% *	99.81% *
Total	1380	97.55% *	99.42% *	99.89% *

注:* 为全 α、全 β 类蛋白质折叠类型的 Mul-Fscore 识别率。

Notes: * Au-α, au-β protein-folding type of Mul-Fscore recognition rate.

参考文献(References)

[1] DAGGETT V, FERSHT A. The present view of the mechanism of protein folding [J]. Nature Reviews Molecular Cell Biology, 2003, 4(6): 497-502.

[2] DAGGETT V, FERSHT A R. Is there a unifying mechanism for protein folding? [J]. Trends in Biochemical Sciences, 2003, 28(1): 18-25.

[3] ONUCHIC J N, WOLYNES P G. Theory of protein folding [J]. Current Opinion in Structural Biology, 2004, 14(1): 70-75.

[4] GIANNI S, GUYDOSH N R, KHAN F, et al. Unifying features in protein-folding mechanisms[J]. Proceedings of the National Academy of Sciences, 2003, 100(23): 13286-13291.

[5] CAO M, COWEN L J. Remote homology detection on alpha-structural proteins using simulated evolution[C]// Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine. ACM, 2012: 353-360.

[6] VOLKAMER A, Kuhn D, RIPPMANN F, et al. Predicting enzymatic function from global binding site descriptors [J]. Proteins: Structure, Function, and Bioinformatics, 2013, 81(3): 479-489.

[7] ANDREEVA A, HOWORTH D, BRENNER S E, et al. SCOP database in 2004: refinements integrate structure and sequence family data[J]. Nucleic Acids Research, 2004, 32(suppl 1): D226-D229.

[8] ORENGO C A, MICHIE A D, JONES S, et al. CATH-a hierarchic classification of protein domain structures[J]. Structure, 1997, 5(8): 1093-1109.

[9] FOX N K, BRENNER S E, CHANDONIA J M. SCOPe: Structural classification of proteins-extended, integrating SCOP and ASTRAL data and classification of new structures[J]. Nucleic Acids Research, 2014, 42(D1):

D304-D309.

[10] SILLITOE I, CUFFA L, DESSAILY B H, et al. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures[J]. Nucleic Acids Research, 2013, 41(D1): D490-D498.

[11] 李晓琴, 仁文科, 刘岳, 等. 蛋白质折叠类型分类方法及分类数据库[J]. 生物信息学, 2010, 8(3): 245-247.
LI Xiaoqin, REN Wenke, LIU Yue, et al. Protein fold type classify methods and classification database [J]. Bioinformatics, 2010, 8(3): 245-247.

[12] 闫金丽, 陈治伟, 徐海松, 等. 基于功能域组分的蛋白质折叠类型识别[J]. 生物化学与生物物理进展, 2011, 38(2): 166-172.
YAN Jinli, CHEN Zhiwei, XU Haisong, et al. Protein fold recognition by functional domain composition [J]. Progress in Biochemistry and Biophysics, 2011, 38(2): 166-172.

[13] 刘岳, 徐海松, 乔辉. 蛋白质折叠类型的分类建模与识别[J]. 物理化学学报, 2009, 25(12): 2558- 2564.
LIU Yue, XU Haisong, QIAO Hui. Classification modeling and recognition of protein fold type[J]. Acta Physico-Chimica Sinica, 2009, 25(12): 2558-2564.

[14] 任文科, 徐海松, 李晓琴. Globin-like 蛋白质折叠类型识别[J]. 生物化学与生物物理进展, 2008, 35(5): 548-554.
REN Wenke, XU Haisong, LI Xiaoqin. Identification proteins of globin-like fold[J]. Progress in Biochemistry and Biophysics, 2008, 35(5): 548-554.

[15] WANG H, HE Z, ZHANG C, et al. Transmembrane protein alignment and fold recognition based on predicted topology[J]. PloS one, 2013, 8(7): e69744.

[16] YANG Y, FARAGGI E, ZHAO H, et al. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates [J]. Bioinformatics, 2011, 27(15): 2076-2082.

[17] COUTSIAS E A, SEOK C, DILL K A. Using quaternions to calculate RMSD [J]. Journal of Computational Chemistry, 2004, 25(15): 1849-1857.

[18] ADAMZ. LGA: a method for finding 3D similarities in protein structures[J]. Nucleic Acids Research, 2003, 31(13): 3370-3374.

[19] SIEW N, ELOFSSON A, RYCHLEWSKI L, et al. MaxSub: An automated measure for the assessment of protein structure prediction quality [J]. Bioinformatics, 2000, 16(9): 776-785.

[20] XU J, ZHANG Y. How significant is a protein structure similarity with TM-score = 0.5? [J]. Bioinformatics, 2010, 26(7): 889-895.

[21] ZHANG Y, SKOLNICK J. Scoring function for automated assessment of protein Structure template quality [J]. Proteins, 2004, 57(4): 702-710.

[22] ZHANG Y, SKOLNICK J. TM-align: a protein structure alignment algorithm based on the TM-score [J]. Nucleic Acids Research, 2005, 33(7): 2302-2309.

[23] PANDIT S B, SK J. Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score[J]. BMC Bioinformatics, 2008, 9(1): 531.