DOI:10.3969/j.issn.1672-5565.20161008001

SCOP 数据库蛋白质折叠类型的自动分类分析

张业晓,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:蛋白质折叠规律研究是生命科学领域重要的前沿课题之一,蛋白质折叠类型分类是折叠规律研究的基础。本研究以SCOP数据库的蛋白质折叠类型分类为基础、以Astral SCOPe 2.05数据库中相似性小于40%的α、β、α+β及α/β类所属的折叠类型为研究对象,完成了989种蛋白质折叠类型的模板构建并形成模板数据库;基于折叠类型设计模板建立了蛋白质折叠类型分类方法,实现了SCOP数据库蛋白质折叠类型的自动化分类。家族模板自洽性检验与独立性检验所得的敏感性、特异性以及MCC的平均值分别为:95.00%、99.99%、0.94与90.00%、99.97%、0.92,折叠类型模板自洽性检验与独立性检验所得的敏感性、特异性以及MCC的平均值分别为:93.71%、99.97%、0.91与86.00%、99.93%、0.87。结果表明:模板设计合理,可有效用于对已知结构的蛋白质进行分类。

关键词:折叠设计合理;模板数据库;分类方法
中图分类号:Q51 **文献标志码**:A **文章编号**:1672-5565(2017)02-078-06

Study of automatic classification of protein folding type in SCOP database

ZHANG Yexiao, LI Xiaoqin
(*College of Life Science and Bioengineering*, *Beijing University of Technology*, *Beijing* 100124, *China*)

Abstract: The study of protein folding pattern is one of important topics in life science field. The classification of protein folding type is the base of study on folding pattern. In this paper, protein folding type classification of SCOP database was as the foundation, similarity less than 40% of the folding type was as the object of study in the class of α, β, $\alpha+\beta$ and α/β of Astral SCOPe 2.05 database, the template of 989 protein folding types was constructed and the template data base was formed. The method of protein folding type classification was constructed based on the design of folding type template, and the automatic classification of protein folding type of SCOP database was realized. In the self-consistency test and independence test of family template, the mean value of sensitivity, specificity and MCC: 95.00%, 99.99%, 0.94 and 90.00%, 99.97%, 0.92, respectively. In the self-consistency test and independence test of folding type template, the mean value of sensitivity, specificity and MCC: 93.71%, 99.97%, 0.91 and 86.00%, 99.93%, 0.87, respectively. The result shows that the design of template is reasonable and it can be effectively used for the classification of proteins with known structure.

Keywords: Classification of folding type; Template database; Method of classification

蛋白质折叠问题,是生命科学领域的前沿课题之一。蛋白质折叠类型反映了蛋白质的核心二级结构单元的连接方式[1]。包括二级结构单元(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等蛋白质分子空间结构组成的3个方面。对自然界存在的数千种折叠类型进行系统分类和识别,探索蛋白质折叠形成的经验规律,将有助于揭示蛋白质的折叠规律,为精确的蛋白质三级结构预测提供基础。

蛋白质三级结构复杂而不规则,但其所对应的蛋白质折叠类型却只有数百到数千种[2],蛋白质折叠类型分类是蛋白质折叠折叠首先需要解决的基本问题。SCOP数据库[3-5]是应用最广泛的结构分类数据库,为层状结构,包括蛋白质结构类、折叠类型、

收稿日期:2016-10-08;**修回日期**:2016-10-21.
基金项目:国家自然科学基金(21173014);北京市自然科学基金(4112010).
作者简介:张业晓,男,硕士研究生,研究方向:生物信息学;E-mail:867130994@qq.com;zhangyexiaobj@126.com.
*通信作者:李晓琴,女,教授,硕士生导师,研究方向:生物信息学;E-mail:lxq0811@bjut.edu.cn.

超家族、家族等不同层次,与蛋白质折叠类型对应的是 fold 层次,它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,根据专家经验人工完成折叠类型的指认。2013 年,在 SCOP 已有分类的基础上,SCOPe[6]数据库建立。尽管 SCOPe 中部分蛋白质样本通过序列比对可自动获得分类结果,但所用自动分类结果与手动分类结果并不相同。新发布的 ASTRAL 现在依然使用 SCOP 中的手动分类结果。最近 7 年,SCOP 数据中折叠层所包含的折叠类型总数基本保持在1 393种左右,4 种主要结构类包含的折叠类型总数保持在1 000种左右,折叠类型总数基本稳定。对已有 SCOP 的人工分类结果进行数据挖掘、建立蛋白质折叠类型分类方法,实现蛋白质折叠类型的自动分类,是迫切需要解决的问题。

模板的选取是建立蛋白质折叠类型分类方法的基础,也直接左右了分类结果的好坏[7]。通常会选取一个结构冗余少、折叠核心清晰的天然蛋白质作为折叠类型模板[8-10]。结构冗余少、折叠核心清晰的天然蛋白质主要靠人工凭经验挑选,不同的模板挑选结果会影响蛋白质折叠类型分类结果[9];同时,对部分家族、超家族数量较多的蛋白质折叠类型,以一个以结构简单的天然样本作为模板的分类结果并不理想[8-10],其原因是由于家族及超家族的分布比较宽泛,使得单一模板无法表现不同家族及超家族的共同特征,即普适性不够,需要多模板才能解决问题。如何克服人工挑选模板的局限性及对部分折叠类型单模板的普适性问题,迫切需要设计反应蛋白质折叠类型共同特征的单模板或多模板来解决上述问题。

本文将在前期工作基础上[10-12],提出系统的蛋白质折叠类型模板设计方法,对 SCOP 数据库 4 种主要结构类的近千种蛋白质折叠类型进行模板设计建模,形成完成蛋白质折叠类型模板数据库,利用成熟的结构比对方法——TM-align 和打分函数——TM-score,建立基于设计模板的蛋白质折叠分类的分类方法,解决 SCOP 数据库的自动分类问题。

1　材料

本课题主要选取 Astral SCOPe 2.05 数据库中相似性小于 40%,且分辨率高于25 nm的 All alpha proteins(α), All beta proteins (β), Alpha and beta proteins (α/β), Alpha and beta proteins ($\alpha+\beta$)4 类蛋白所属的折叠类型为研究对象,其中共有 989 种折叠类型、12 165个样本,相应数据记为 Set-I。表 1 列举了 4 类蛋白包含的折叠类型数目、家族数目以及样本数目。

表 1　4 类蛋白包含的折叠类型、家族及样本数目

Table 1　Number of folding type, family and sample of four class

种类	折叠类型	家族	样本
α	286	962	2 366
β	175	865	2 714
α/β	148	919	3 833
$\alpha+\beta$	380	1 195	3 252
ALL	989	3 941	12 165

实验集中,有 359 种蛋白质折叠类型仅包含一个家族,且家族中仅包含一个样本,对于这部分折叠类型,需要利用 Astral SCOPe 2.05 数据库中相似性小于 95%的数据信息,相应数据记为 Set-I-1;其余 630 种蛋白质折叠类型含有两个及两个以上家族,对应的家族数及样本数分别为 3 582、11 806,相应数据记为 Set-I-2。

独立检验集:SCOPe astral 2.06 数据库[6]中剔除 SCOPe astral2.05 所含样本,余下 2 142 样本,涉及 368 种蛋白质折叠类型,记为 Set-II。

2　蛋白质折叠类型模板设计及模板数据库的构建

蛋白质折叠类型的分类以蛋白质折叠核心的规则结构片段组成、连接和空间排布为依据,其中的规则结构片段即 α-螺旋或 β-折叠,其骨架结构主要由 α-碳原子连接而形成。因此折叠类型模板的设计就是确定折叠核心的片段并对其骨架结构的 α-碳原子坐标进行建模。

2.1　家族模板设计方法及家族模板数据库

以 BRD-like 折叠类型模板设计方法[12]为基础并修改完善,建立系统的家族模板设计方法。具体步骤为:对家族样本利用 MUSTANG[13]进行多结构比对,获得多结构比对信息;提取多结构比对信息中完全匹配的片段(即家族样本共同参与的折叠核心片段),形成该家族模板的折叠核心结构;对折叠核心片段进行骨架结构建模(即提取骨架坐标信息),形成家族模板。

骨架坐标提取方法:对由 n 个样本组成的家族,利用 MUSTANG 进行多结构比对,获得多结构比对结果,提取完全匹配片段,对匹配片段中任一残基 i 的 α-碳原子匹配坐标信息—— (x_i, y_i, z_i) ,计算匹配坐标的平均值—— $(\bar{x}, \bar{y}, \bar{z})$,将其作为该残基的骨架 α-碳坐标信息,形成匹配片段的骨架坐标信

息。求坐标平均值公式如下：

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i.$$

利用 MUSTANG[13] 进行程序蛋白质多样本的多结构比对，是因为与 POSA[14]、CE-MC[15]、MALECON[16] 和 MultiProt[17] 等多结构比对软件相比，该软件它在空间折叠、残基的接触模式中有较强的识别能力。

利用上述方法，对 989 种蛋白质折叠类型涵盖的 3 941 家族分别构建家族模板，形成蛋白质家族模板数据库。数据库中的家族模板编号为 SCOPe astral 中相应家族代码，模板在 4 种结构类中的分布如图 1 所示。

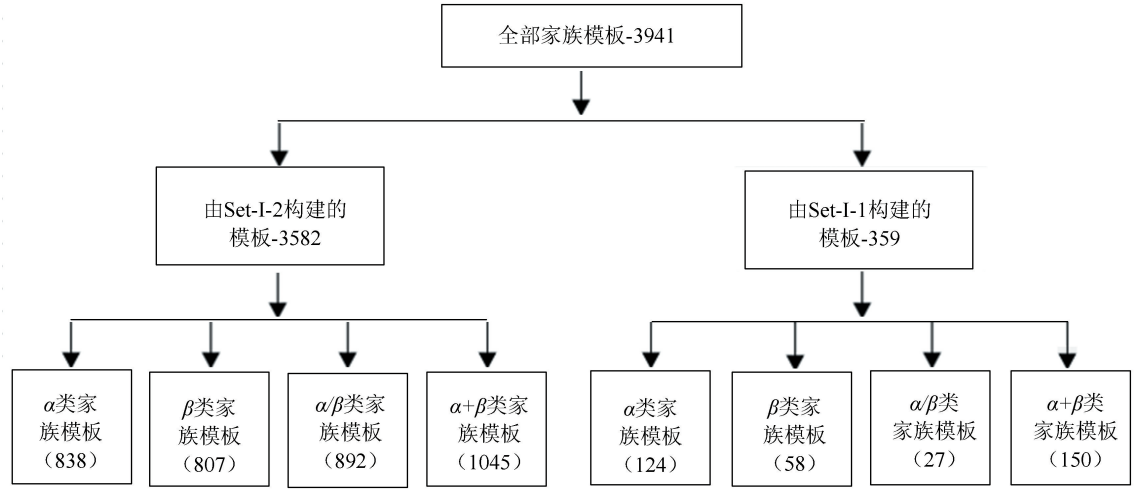

图 **1** 家族模板数据库分布

Fig.1 Database distribution diagram of family template

2.2 折叠类型模板设计方法及折叠类型模板数据库

蛋白质折叠类型模板是以家族模板为单位通过系统聚类并经过筛选和验证最终得到。系统聚类的基本思想：对任意蛋白质折叠类型所属的 n 个家族模板，先将 n 个家族模板看成不同的 n 类，然后将性质最接近（距离最近）的两类合并为一类，再从 $n-1$ 类中找到最接近的两类加以合并，依此类推，直到所有的家族模板被合为一类，得到 n 个家族模板的系统聚类图。家族模板通过 TM-align[18] 进行两两比对，以 TM-score[19] 作为距离参数，将 TM-score 取值最大（即距离最小）的两家族合并，合并方法与模板数据库的蛋白质折叠类型分类方法相同。

在 Bromodomain-like 折叠类型模板的设计基础上[12]，并通过对系统聚类图中节点对应初始模板的计算分析及检验，提出任意蛋白质折叠类型 i 模板筛选的经验标准：具有折叠类型 i 特有全部折叠核心片段；分布于系统聚类图中的独立分支；由家族模板首次合并形成；对蛋白质折叠类型 i 所属样本的识别率不低于 80%。

利用上述方法，对 989 种蛋白质折叠类型分别构建模板，组成折叠类型模板数据库，模板分布如图 2 所示。其中，由数据集 Set-I-1 构建的模板 359 种，由于这些蛋白质折叠类型仅含一个家族，家族模板即为折叠类型模板；由数据集 Set-I-2 构建的模板数共1 258，其中 508 种蛋白质折叠类型成功筛选到了模板，另外的 122 种折叠类型未能筛选到满足条件的模板，以家族模板替代折叠类型模板。

蛋白质折叠类型模板的具体数据信息，见表 2。Fold 代表折叠类型，Number of template 代表每种折叠类型中模板的数量，Mode-ID 为编号，TM-score 为合并家族模板的打分值。以 b.1.5.1_29.8 为例，其中 b 代表结构类，即全 β 类，1 代表 SCOP 数据库中全 β 类的折叠类型，5.1_29.8 代表形成该模板的 5.1 和 29.8 家族，相应的 TM-score 列对应的单元格为空。

由表 2 可知，折叠类型模板识别率及 TM-score 的平均值分别为 96.17%、0.83，模板的平均识别率明显高于筛选标准，由此说明，模板本身抓住了折叠类型的基本特征，模板设计具有合理性及适用性。

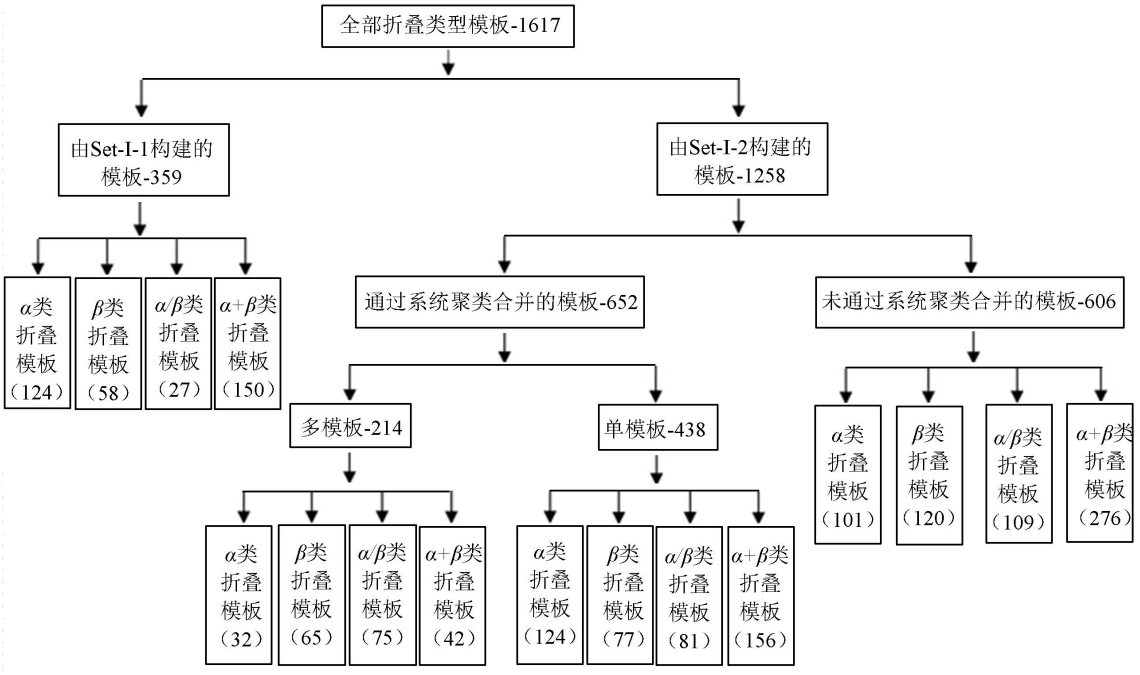

图 **2**　折叠类型模板数据库分布

Fig.2　Database distribution diagram of folding type template

表 2　折叠类型模板的具体信息

Table 2　Specific information about folding type templates

Fold	Number of template	Mode-ID	TM-score	识别率/%
a.1	2	a.1.1.0_1.2	0.78	89.09
		a.1.1.1_1.3	0.63	89.09
…	…	…	…	…
b.1	2	b.1.5.1_29.8	0.84	91.84
		b.1.13.0_13.1	0.81	80.00
…	…	…	…	…
MEAN			0.83	96.17

3　模板数据库的蛋白质折叠类型分类方法及结果

3.1　模板数据库的蛋白质折叠类型分类方法

将任意待测蛋白样本与模板数据中的所有模板进行 TM-align[18] 比对,计算 TM-score[19] 值。TM-score 取值最大的模板所在的折叠类型即为待测蛋白样本所属折叠类型。

分类结果利用敏感性、特异性、Matthew 相关系数 3 个指标对其进行评估,参数定义如下:

$$\text{敏感性 } S_n = \frac{t_p}{t_p + f_n} \times 100\%,$$

$$\text{特异性 } S_p = \frac{t_n}{t_n + f_p} \times 100\%,$$

相关系数

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

式中: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。

3.2　自洽性检验

为验证模板设计及分类方法的合理性,以数据集 Set-I 中的样本为研究对象,分别利用家族模板数据库与折叠类型模板数据库进行蛋白质折叠类型分类的自洽性检验,检验结果见表 3、4。其中 S 表示折叠类型所含样本数量,S' 为真阳性与假阳性数量之和。

表 3　家族模板的自洽性检验

Table 3　Self consistency test of family template

Fold	$S(S')$	$t_p(t_n)$	$f_n(f_p)$	$S_n(S_p)$/%	MCC
a.1	56(55)	55(12109)	1(0)	98.21(100.00)	0.99
…	…	…	…	..	…
b.1	515(504)	502(11 648)	13(2)	97.48(99.98)	0.98
…	…	…	…	…	…
c.1	477(499)	477(11 666)	0(22)	100.00(99.81)	0.97
…	…	…	…	…	…
d.16	24(21)	21(12 141)	3(0)	87.50(100.00)	0.94
…	…	…	…	…	…
MEAN				95.00(99.99)	0.94

表 4 折叠类型模板的自洽性检验

Table 4 Self consistency test of folding type template

Fold	$S(S')$	$t_p(t_n)$	$f_n(f_p)$	$S_n(S_p)/\%$	MCC
a.1	56(47)	46(12 108)	10(0)	82.14(99.99)	0.90
…	…	…	…	..	…
b.121	62(68)	57(12 092)	5(11)	91.94(99.91)	0.88
…	…	…	…	…	…
c.1	477(506)	477(11 659)	0(29)	100.00(99.75)	0.97
…	…	…	…	…	…
d.16	24(20)	20(12 141)	4(0)	83.33(100.00)	0.91
…	…	…	…	…	…
MEAN				93.71(99.97)	0.91

由检验结果可知,基于家族模板数据库自洽性检验结果的敏感性、特异性及 MCC 的均值分别高达 95.00%、99.99%、0.94,基于折叠类型模板数据库自洽性检验结果的敏感性、特异性以及 MCC 的均值分别为 93.71%、99.97%及 0.91。两种类型模板对相同数据集的分类检验结果相当,前者的分类结果略高于后者。说明家族模板及折叠类型模板设计合理,模板反映了折叠类型的基本特征;前者的模板总数为3 941,后者仅为1 617,后者模板数仅为前者的 2/5,分类速度后者远远优于前者,分类精度家族模板略优于折叠类型模板。

3.3 独立性检验

为进一步检验模板数据库及分类方法的普适性,以数据集 Set-II 中的样本为研究对象,分别对家族模板数据库与折叠类型模板数据库进行独立性检验,检验结果见表 5、6。$S+$为数据集 Set-II 中样本数量。

表 5 家族模板的独立性检验

Table 5 Independence test of family template

Fold	$S+(S')$	$t_p(t_n)$	$f_n(f_p)$	$S_n(S_p)/\%$	MCC
a.1	4(4)	4(2 138)	0(0)	100.00(100.00)	1.00
…	…	…	…	..	…
b.121	3(4)	3(2 138)	0(1)	100.00(99.95)	0.87
…	…	…	…	…	…
c.1	77(104)	77(2 038)	0(27)	100.00(98.69)	0.85
…	…	…	…	…	…
d.113	8(8)	8(2 134)	0(0)	100.00(100.00)	1.00
…	…	…	…	…	…
MEAN				90.00(99.97)	0.92

表 6 折叠类型模板的独立性检验

Table 6 Independence test of folding type template

Fold	$S+(S')$	$t_p(t_n)$	$f_n(f_p)$	$S_n(S_p)/\%$	MCC
a.1	4(4)	4(2 138)	0(0)	100.00(100.00)	1.00
…	…	…	…	..	…
b.122	4(3)	3(2 138)	1(0)	75.00(100.00)	0.87
…	…	…	…	…	…
c.1	77(108)	77(2 034)	0(31)	100.00(98.50)	0.84
…	…	…	…	…	…
d.113	8(9)	8(2 133)	0(1)	100.00(99.95)	0.94
…	…	…	…	…	…
MEAN				86.00(99.93)	0.87

由表 5、6 可知,家族模板数据库及折叠类型模板数据库对扩充样本的分类效果稍差于自洽性检验中的结果,但是在独立性检验中家族模板与折叠类型模板的分类效果普遍高于 90%,说明模板数据库及其分类方法可用于对扩充蛋白样本进行折叠类型的分类,从而验证了模板设计及分类方法具有有效的普适性。

4 结 论

蛋白质折叠规律研究是生命科学重大前沿课题,折叠分类是蛋白质折叠研究的基础,折叠分类也将应用到蛋白质识别中去。本文基于 Astral SCOPe 2.05 数据库中相似性小于 40%的 α、β、$\alpha+\beta$ 及 α/β 所属的折叠类型为研究对象,通过对蛋白质折叠结构分析及信息挖掘,完善了蛋白质折叠类型模板设计方法,完成家族模板数据库及折叠类型模板数据库的构建,建立基于模板的蛋白质折叠类型分类方法,并用于蛋白质折叠类型的自动化分类。结果表明:1)模板设计方法合理,并可用于家族及折叠类型模板的构建;2)构建了完整的 α、β、α/β 以及 $\alpha+\beta$ 等 4 类蛋白所包含折叠类型模板数据库及家族模板数据库;3)蛋白质折叠类型分类方法能够有效对已知结构的蛋白进行折叠类型的归类。

致谢

本课题能够顺利完成,首先,感谢北京市自然科学基金资助项目的大力支持;其次,衷心的感谢导师李晓琴教授的悉心指导,从文章的选题、研究计划的制定,各个方面都离不开李老师热情耐心的帮助和教导;最后,感谢实验室的同学们对我提供的帮助。

参考文献（References）

[1]FINKELSTEIN A V, PTITSYN O B. Why do globular proteins fit the limited set of folding patterns[J]. Progress in Biophysics & Molecular Biology, 1987, 50(3): 171-190. DOI: 10.1016/0079-6107(87)90013-7.

[2]CHOTHIA C. Proteins. One thousand families for the molecular biologist[J]. Nature, 1992, 357(6379): 543-544. DOI: 10.1038/357543a0.

[3] 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. DOI: 10.1093/nar/gkh039.

[4]GANDHIMATHI A, GHOSH P, HARIHARAPUTRAN S, et al. PASS2 database for the structure-based sequence alignment of distantly related SCOP domain superfamilies: update to version 5 and added features[J]. Nucleic Acids Research, 2015, 44(D1): D410-D414. DOI: 10.1093/nar/gkv1205.

[5]ANDREEVA A, HOWORTHD, CHANDONIA J M, et al. Data growth and its impact on the SCOP database: new developments[J]. Nucleic Acids Research, 2008, 36(Suppl-1): D419-D425. DOI: 10.1093/nar/gkm993.

[6]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. DOI: 10.1093/nar/gkt1240.

[7]KELLEY L A, MACCALLUM R M, STEMBERG M J. Enhanced genome annotation using structural profiles in the program 3D-PSSM[J]. Journal of Molecular Biology, 2000, 299(2): 499-520. DOI: 10.1006/jmbi.2000.3741.

[8]马帅, 王勤, 李晓琴. α/β 类蛋白质折叠类型的分类方法研究[J]. 生物信息学, 2014, 12(2): 123-132. DOI: 10.3969/j.issn.1672-5565.2014.02.08.
MA Shuai, WANG Qin, LI Xiaoqin. The study of classification of protein folding types of α/β[J]. China Journal of Bioinformatics, 2014, 12(2): 123-132. DOI: 10.3969/j.issn.1672-5565.2014.02.08.

[9]孔令强, 李晓琴. 基于特征片段信息的 PH domain-like barrel 蛋白质折叠类型分类方法[J]. 生物信息学, 2012, 10(2): 125-129. DOI: 10.3969/j.issn. 1672-5565.2012.02.13.
KONG Lingqiang, LI Xiaoqin. A method of PH domain-like barrel protein fold classification based on characteristic fragments[J]. China Journal of Bioinformatics, 2012, 10(2): 125-129. DOI: 10.3969/j.issn. 1672-5565.2012.02.13.

[10]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. DOI: 10.3969/j.issn.1672-5565.2010.03.015.
LI Xiaoqin, REN Wenke, LIU Yue, et al. Protein fold type classify methods and classification database[J]. China journal of Bioinformatics, 2010, 8(3): 245-253. DOI: 10.3969/j.issn.1672-5565.2010.03.015.

[11]LUO Liaofu, LI Xiaoqin. Recognition and architecture of the framework structure of protein[J]. Proteins Structure Function & Bioinformatics, 2000, 39(1): 9-25. DOI: 10.1002/(SICI)1097-0134(20000401)39:1<9::AID-PROT2>3.3.CO;2-C.

[12]李晓琴, 张春城. Bromodomain-like 折叠类型模板的设计[J]. 北京工业大学学报, 2016, 42(10):1572-1580. DOI: 10.11936 /bjutxb2015100078.
LI Xiaoqin, ZHANG Chuncheng. Design of folding type template of Bromodomain-like[J]. Journal of Beijing University of Technology, 2016, 42(10):1572-1580. DOI: 10.11936 /bjutxb2015100078.

[13]KONAGURTHU A S, WHISSTOCK J C, STUCKEY P J, et al. MUSTANG: a multiple structural alignment algorithm [J]. Proteins Structure Function & Bioinformatics, 2006, 64(3): 559-574. DOI: 10.1002/prot.20921.

[14]YE Y, GODZIK A. Multiple flexible structure alignment using partial order graphs[J]. Bioinformatics, 2005, 21(10): 2362-2369. DOI: 10.1093/bioinformatics/bti353.

[15]GUDA C, LU S, SCHEEFF E D, et al. CE-MC: a multiple protein structure alignment server[J]. Nucleic acids research, 2004, 32(suppl 2): W100-W103. DOI: 10.1093/nar/gkh464.

[16]OCHAGAVIA M E, WODAK S. Progressive combinatorial algorithm for multiple structural alignments: application to distantly related proteins[J]. Proteins Structure Function & Bioinformatics, 2004, 55(2): 436-454. DOI: 10.1002/prot.10587.

[17]SHATSKY M, NUSSINOV R, WOLFSON H J. MultiProt-a multiple protein structural alignment algorithm[J]. Lecture Notes in Computer Science, 2002, 2452: 235-250. DOI: 10.1007/3-540-45784-4_18.

[18]ZHANG Yang, SKOLNICK J. TM-align: a protein structure alignment algorithm based on the TM-score[J]. Nucleic Acids Research, 2005, 33(7): 2302-2309. DOI: 10.1093/nar/gki524.

[19]XU Jinrui, ZHANG Yang. How significant is a protein structure similarity with TM-score = 0.5? [J]. Bioinformatics, 2010, 26(7): 889-895. DOL: 10.1093/bioinformatics/btq066.