Bromodomain-like 折叠类型模板的设计

李晓琴，张春城
（北京工业大学生命科学与生物工程学院，北京 100124）

摘 要： 针对折叠类型分类中所选天然模板普适性不足的问题，提出了 Bromodomain-like 折叠类型模板的设计方法. 选 SCOPe Astral 2.03 序列相似度小于 40% 并且分辨率高于 0.25 nm 的 52 个可用 Bromodomain-like 折叠样本，基于多结构比对结果及数据分析，建立了折叠类型家族模板的设计方法. 利用系统聚类方法构建了家族模板的系统聚类图，提出了蛋白质折叠类型模板的设计方法，并用于该折叠类型的模板设计. 结果表明：设计模板具有普适性，可用于蛋白质折叠类型分类.

关键词： 折叠类型分类；模板设计；结构比对；系统聚类
中图分类号： O 641 **文献标志码：** A **文章编号：** 0254－0037(2016) 10－1572－09
doi： 10.11936/bjutxb2015100078

Design of a Bromodomain-like Folding Type Template

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

Abstract: For the problem that the universal shortage of natural template for folding type classification, a design method of the Bromodomain-like folding type template was presented. 52 Bromodomain-like folding type samples whose sequence similarity is less than 40% were chosen and the resolution was higher than 0.25 nm in the database of the SCOPe of astral 2.03. Based on the results of multiple structure alignment and data analysis, the design method of the folding type family template was established. The clustering graph of family template was constructed using the system clustering method, and the design of the template of the folding type was completed. Results show that the design templates have universality, and the templates can be used for protein folding type classification.

Key words: folding type classification; template design; structure comparison; system clustering

蛋白质的结构能够提供蛋白质的很多信息，有助于了解蛋白质的功能和分子机制[1]. 目前，研究蛋白质结构的方法有 2 种，分别为实验测定方法和理论预测方法. 实验测定蛋白质三维结构的方法主要采用 X-ray 晶体衍射法[2] 和核磁共振波谱法. 传统的基于 Anfinsen “热力学假说”[3] 原理的蛋白质三维结构预测方法通常可分为 3 类：同源模建方法、折叠识别方法和从头预测法[4-5]. 同源模建受到序列相似度的限制，从头计算运算量太大，介于 2 种方法之间的折叠识别被认为是最有前途的方法，其基本思想是：预测的蛋白质折叠类型与某一已知结构的蛋白质折叠类型相同，这样蛋白质的折叠问题就转化为在已知空间结构的蛋白质中，选择一种最有可能的折叠类型，从而大大减少了预测蛋白质三维结构的难度. 尽管蛋白质空间结构预测的理论方法比较成熟，但空间结构即原子坐标的预测依然困难.

收稿日期： 2015-10-29
基金项目： 国家自然科学基金资助项目(21173014)；北京市自然科学基金资助项目(4112010)
作者简介： 李晓琴(1966—)，女，教授，主要从事生物信息学理论方面的研究，E-mail：lxq0811@bjut.edu.cn

蛋白质的空间结构十分复杂，但它的框架结构(折叠类型或拓扑结构)却较为简单，粗粒意义下的蛋白质结构研究越来越得到研究者的关注[6-7]. 蛋白质折叠类型是一种粗粒化的结构，包括蛋白质分子空间结构的 3 个主要方面：二级结构单元、二级结构单元的相对排布位置、蛋白质多肽链的整个路由关系(肽链走向)[8].

研究表明，蛋白质的折叠类型也只有数百到数千种[9-10]，远小于蛋白质分子折叠的自由度数，并且，蛋白质的折叠速率和折叠机制在很大程度由天然状态的拓扑所决定[11]. 对自然界存在的数百到数千种折叠类型进行系统研究，探索构建蛋白质折叠类型模板的方法，为进一步的蛋白质折叠类型分类及识别研究奠定基础，并有助于揭示蛋白质的折叠规律.

模板的选取是蛋白质折叠类型分类的关键问题. 在以往选择模板时，通常在结构数据库中选择天然蛋白质为模板，其依据以环区结构冗余小、折叠核心清晰且结构数据所占存储空间小的天然蛋白质为模板. 环区和折叠核心的清晰程度都影响预测的准确性. 研究表明，模板的好坏直接影响了预测模型的好坏，即预测的模型倾向于模板的模型[12].

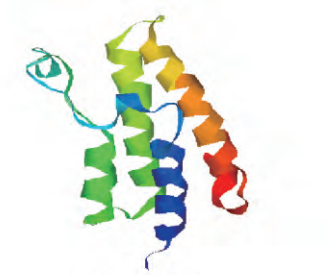
蛋白质的折叠类型主要由形成折叠核心的规则二级结构片段组成、排布、走向决定，蛋白质折叠类型的模板应该围绕折叠核心的的规则二级结构片段来构建. 通常选取结构简单的天然样本作为模板，这样折叠核心以外的其他结构就成为折叠类型分类的干扰因素，如何去除干扰、提取反应折叠类型拓扑特征的模板成为解决折叠类型分类的关键问题之一；另外，在一个蛋白质折叠类型内部，通常会包含多了家族和多个超家族，以结构简单的天然样本为模板，该模板具有所在家族的个性化结构特征，但不足以代表折叠类型所属全部超家族样本的共性特征，模板的普适性会比较差，如何克服天然模板的局限性、提高折叠类型模板的普适性成为解决折叠类型分类的又一关键问题.

基于此，通过对数据库中 Bromodomain-like 折叠类型的家族分类及样本进行分析，抓住形成蛋白质折叠类型的折叠核心结构，提出了 Bromodomain-like 折叠类型模板的设计方法，并用于该折叠类型的模板设计.

1　材料

Bromodomain(BRD)蛋白是一种进化高度保守的约含有 110 个氨基酸的溴蛋白功能结构域，这个家族在人体内能够唯一特异性的识别蛋白质中的乙酰化赖氨酸(KAc)[13]，使得 BRD 蛋白具有辨别不同蛋白结合物的能力[14-16]，因此它成为蛋白质交互模块中不断探索药物发现领域的代表. 大部分的 BRD 蛋白都在调控如组蛋白乙酰酶、依赖 ATP 的染色质重塑、甲基化转移酶和转录激活因子等基因转录过程中发挥重要的作用，并与肿瘤、神经紊乱、炎症、肥胖和心血管疾病发生相关[17]，是近年来的研究热点.

选取 Bromodomain-like 折叠类型为研究对象，在 SCOPe Astral 2.03 数据库中其对应编号为 a.29. 该折叠类型为左手四螺旋束结构，包含 15 个超家族、20 个家族. 为避免冗余序列对模板设计的影响，选取序列相似度小于 40%、分辨率高于 0.25 nm 的 52 样本，样本蛋白的 Astral SCOPe ID 如表 1 所示. 图 1 为 BRD 蛋白质模型及拓扑结构模型. SCOPe Astral 2.03 中相似度小于 40%、分辨率高于 0.25 nm 非 Bromodomain-like 折叠类型样本数为 12 065.



(a) 蛋白质模型

(b) 拓扑结构模型

图 1　BRD 蛋白质模型

Fig. 1　BRD protein model

对于核磁共振样本，利用其对应的多套骨架模型信息，参照 2.1 家族模板设计方法，建立单骨架样本模型；对于原子信息缺失较多的样本，不用于构建，仅用于折叠类型的模板的验证.

表 1 Bromodomain-like 折叠类型 52 个样本

Table 1 BRD folding type 52 samples

家族	样本名称			样本数
a. 29. 2. 0	d4ldfa_	d3qzta_	d3ljwa_	13
	d4ir5a_	d2grca_	d3q2ea_	
	d3daia_	d3daia_	d3rcwa_	
	d3jvla_	d3mb4a_	d3rcwa_	
	d3u5la_			
a. 29. 2. 1	d1e6ia_	d1eqfa1	d1eqfa2	4
	d3p1fa_			
a. 29. 3. 0	d2pg0a2	d4hr3a2	d3r7ka2	5
	d3swoa2	d3mkha2		
a. 29. 3. 1	d1u8va1	d2d29a1	d1rx0a1	6
	d1ivha1	d2reha1		
a. 29. 3. 2	d1w07a1	d1w07a2		2
a. 29. 5. 1	d1gkza1	d2pnra1		2
a. 29. 6. 0	d1xg2b_			1
a. 29. 6. 1	d1x91a_	d2cj4a_		2
a. 29. 7. 1	d1pi1a_			1
a. 29. 8. 1	d1tdpa_			1
a. 29. 8. 2	d2bl8a1			1
a. 29. 9. 1	d2etda1*			1
a. 29. 10. 1	d1v9va1			1
a. 29. 11. 1	d2fefa1			1
a. 29. 12. 1	d2fug11			1
a. 29. 13. 1	d3dbya1	d3dbya2	d3d19a1	4
	d3d19a2			
a. 29. 14. 1	d2qzga1	d2qsba1		2
a. 29. 15. 1	d2hi7b1			1
a. 29. 16. 1	d2rlda1	d2gsca1		2
a. 29. 17. 1	d2hgka1**	1		

备注: 标* 样本存在信息缺失; 标** 样本结构为核磁共振结果.

2 Bromodomain-like 折叠类型模板设计

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 家族模板的设计与生成

蛋白质折叠类型所属家族模板的设计,就是确定家族样本中共同参与折叠核心形成的结构片段,并对其骨架结构的 α-碳原子坐标建模,家族模板是构建蛋白质折叠类型模板的基础.

对 Bromodomain-like 折叠类型所属的任意家族,根据以下步骤建模: 1) 对家族样本进行多结构比对,获得多结构比对信息; 2) 对获得的多结构比对信息进行分析,确定并提取折叠核心片段; 3) 对折叠核心片段进行骨架结构建模. 根据分类结果,家族包含有 2 个及 2 个以上样本的,依据上述步骤建模. 家族内只含一个样本的,将其作为本家族的模板. 家族模板设计的流程如图 2 所示.

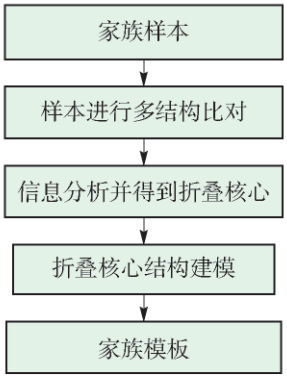


图 2 家族模板设计的流程

Fig. 2 Family template design flow chart

家族模板的折叠核心结构通过多结构比对信息获得. 多结构比对信息中,完全匹配的片段即家族样本共同参与折叠核心的片段,提取全部的匹配片段,形成该家族模板的折叠核心结构. 目前结构比对算法如 CE[18]、DaliLite[19]、SSM[20]、TM-align[21]、MUSTANG[22]、GOSSIP[23]. 本文利用 MUSTANG 多结构比对算法,MUSTANG 是在 DALI 双结构比对获得成功的基础上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.

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.

求坐标平均值公式为

$$\begin{cases} \bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i \\ \bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i \\ \bar{z} = \dfrac{1}{n}\sum_{i=1}^{n} z_i \end{cases}$$

2.2　家族模板系统聚类图的建立及稳定性分析

通过家族模板的设计流程得到各个家族的模板，由于家族 a. 29. 9. 1 已经被舍弃，于是共生成 19 个家族模板，以 19 个家族模板为基础构建本折叠类型模板.

折叠类型模板设计的流程如图 3 所示.

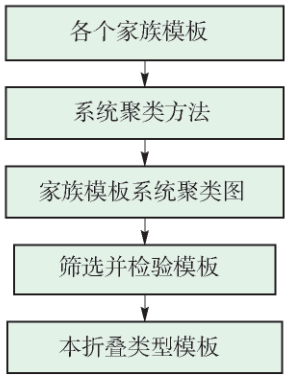


图 3　折叠类型模板设计的流程

Fig. 3　Flow chart for the design of the folding type template

系统聚类是将多个样本分成若干类的方法，其基本思想是：先将所有 n 个样本看成不同的 n 类，然后将性质最接近（距离最近）的 2 类合并为 1 类，再从这 $n-1$ 类中找到最接近的 2 类加以合并，依此类推，直到所有的样本被合为 1 类. 两样本的合并与生成方法同 2. 1. 利用 TM-align 结构比对程序给出的 TM-score（或 RMSD）参数作为距离指标，构建家族模板的系统聚类图. TM-score 取值 [0,1]，值越高代表 2 样本结构越相似；RMSD 越小，说明两样本结构越相似. 依据 TM-score 的家族模板系统聚类图如图 4 所示，各分支点的对应的 RMSD 及 TM-score 参数如表 2 所示. 英文字母代表形成的模板，例如 a 代表 3. 0 家族和 3. 1 家族形成的模板，字母顺序代表构建模板顺序.

由图 4、表 2 可知：1）随着聚类的进行，RMSD 总体呈现一个递增的趋势，TM-score 总体呈现递减的趋势，这是由于模板之间的差异性逐渐变大所导致. 2）模板间的 RMSD 都在 4 以内，TM-score 都在 0. 5 以上（r 模板除外），说明模板间的稳定性良好，相似性良好. 3）蛋白质的最先聚类是在超家族内部，而且具有很高的 TM-score 打分值以及较低的 RMSD，如家族 3. 0 和 3. 1，6. 0 和 6. 1，2. 0 和 2. 1 的聚类，TM-score 都在 0. 9 以上并且 RMSD 在 1. 3 以下，说明超家族内部的样本差异性小；其次聚类的是折叠相似的超家族之间，例如 13. 1 和 16. 1、5. 1 和 8. 2，RMSD 在 2. 4 左右，打分值分别为 0. 79 和 0. 67，说明超家族之间的特异性逐渐变大.

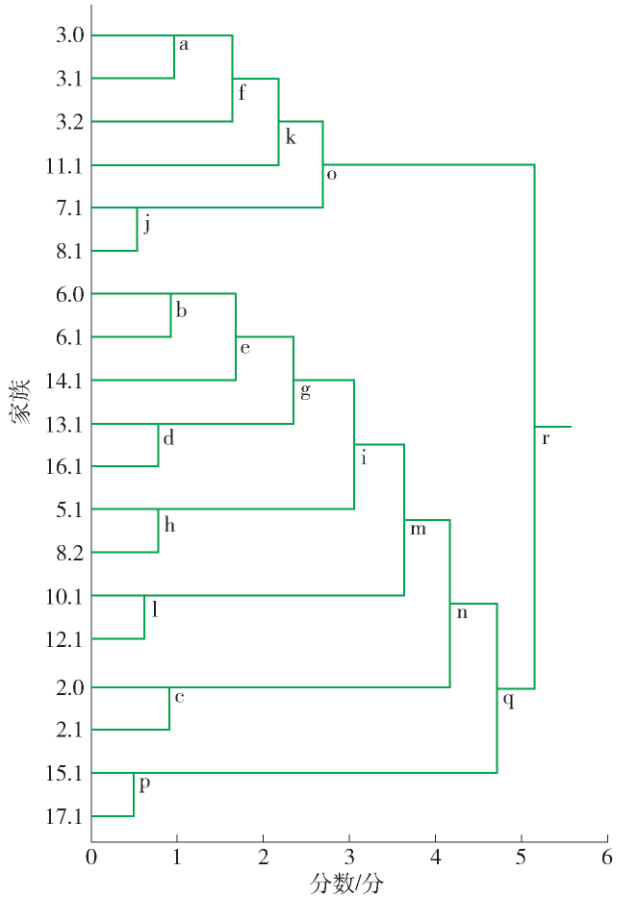


图 4　根据 TM-score 的模板系统聚类图

Fig. 4　According to TM-score template system cluster chart

表 2　图 4 分支点对应的 RMSD 及 TM-score 的参数

Table 2　Corresponding parameters of the RMSD and TM-score in Fig. 4

模板	RMSD	分数	模板	RMSD	分数
a	0. 08	0. 96	j	0. 39	0. 55
b	0. 13	0. 93	k	0. 38	0. 55
c	0. 07	0. 92	l	0. 31	0. 54
d	0. 24	0. 79	m	0. 24	0. 59
e	0. 19	0. 75	n	0. 28	0. 53
f	0. 22	0. 68	o	0. 29	0. 51
g	0. 22	0. 67	p	0. 34	0. 5
h	0. 24	0. 63	q	0. 24	0. 55
i	0. 18	0. 72	r	0. 25	0. 43

为进一步检验家族模板聚类图稳定性，以 RMSD 为距离参数获得的家族模板系统聚类图如图 5 所示. 各分支点的对应的 RMSD 及 TM-score 参数如表 3 所示.

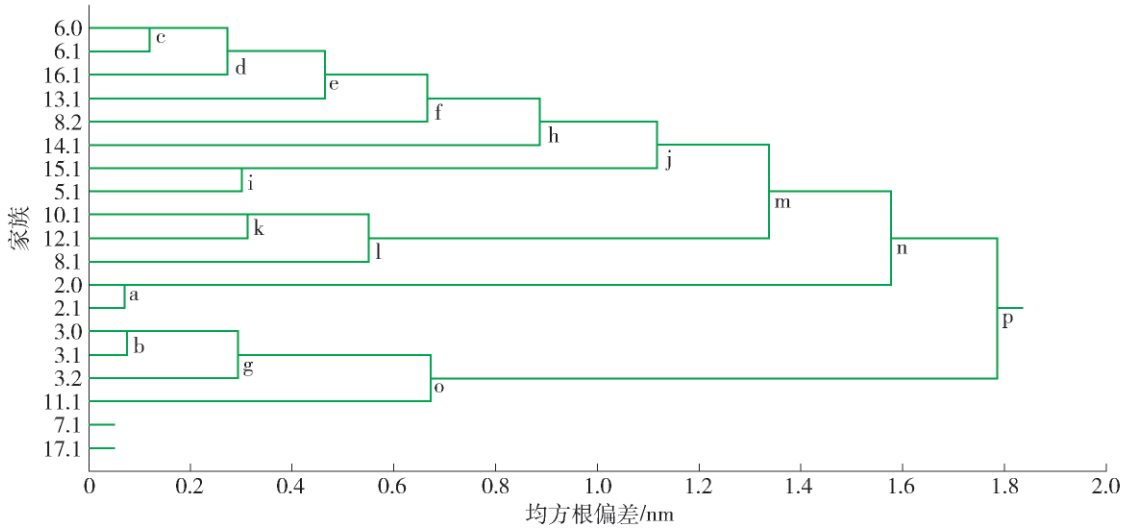


图 5 根据 RMSD 的模板系统聚类图

Fig. 5 According to RMSD template system cluster chart

表 3 图 5 中各分支点对应的 RMSD 及 TM-score 参数

Table 3 Corresponding parameters of the RMSD and TM-score in Fig. 5

模板	RMSD	分数	模板	RMSD	分数
a	0.07	0.92	i	0.30	0.52
b	0.08	0.96	j	0.23	0.68
c	0.12	0.93	k	0.31	0.54
d	0.15	0.74	l	0.24	0.51
e	0.19	0.64	m	0.22	0.60
f	0.20	0.60	n	0.24	0.53
g	0.21	0.68	o	0.38	0.54
h	0.22	0.58	p	0.21	0.37

由图 5 可知：1）最先聚类是在超家族内部，而且具有很高的 TM-score 打分值以及较低的 RMSD，与图 4 结果一致；2）与图 4 相比，图 4 中模板 o 所在聚类区间与图 5 中模板 o 所在聚类区间，都是由家族 3.0、3.1、3.2、11.1 聚类而成，其差别在于图 5 的家族 8.1 在模板 n 所在区间，家族 7.1 在图 5 中没有与任何模板聚类；3）图 4 中模板 q 所在的区间和图 5 中模板 n 所在的区间，都是由家族 6.0、6.1、16.1、13.1、8.2、14.1、15.1、5.1、10.1、12.1、8.1、2.0、2.1 聚类而成，只是聚类的顺序不同，差别在于家族 8.1 分别聚类在图 4 中模板 o 区间和图 5 中模板 n 区间，而家族 17.1 在图 5 中没有参与聚类. 图 4、5 的总体差别在于家族 8.1 和家族 8.2，2 个家族都只有 1 个样本，其中 a.29.8.1 家族模板是由核磁共振结构建立的模板，而 a.29.8.2 家族是 1 个含有很长冗余的结构. 通过以上对不同参数获得的家族模板聚类结果的分析可知，以 TM-score 为参数的聚类图稳定性很好，可以将 TM-score 的聚类结果作为本文的聚类依据.

2.3 基于系统聚类图的 Bromodomain-like 折叠类型模板的选取标准

根据图 4 的 TM-score 系统聚类图，共生成 a ~ r 共 18 个模板，将各个模板对本折叠类型的 52 个样本及非本折叠类型的 12 065 个样本进行 TM-algin 比对，得到 TM-score，并以 TM-score 的取值 0.5 作为阈值，当 TM-score 大于等于 0.5 时，待测蛋白与模板属于同一折叠类型，否则为不同折叠类型. 计算各个模板用于折叠类型分类的识别率、MCC 值及尤登指数[24]，结果如表 4 所示. 识别率、MCC 值及尤登指数反映了设计模板用于折叠类型分类的有效性.

虽然各家族模板最终聚类为一个 r 模板，但是 r 模板在阈值为 0.5 时的识别率为所有模板中最低的，由于 r 模板的折叠核心片段较短，因此不能将 r 模板作为最后的折叠类型模板；处于各独立分枝中的最先聚类的模板识别率等指标相对较好. 基于上述结果，并结合蛋白质折叠类型的确定标准，提出以下折叠类型模板筛选标准：1）模板的折叠核心片段清晰；2）模板分布于各独立分枝；3）模板的识别率在 80% 以上；4）模板由家族模板首次合并形成. 满足以上 4 个标准的只有 4 个

模板，分别为 c、d、h、j 模板，将这 4 个模板作为折叠类型模板. 如图 6 所示，为各个待选模板和 r 模板的骨架模型.

表 4　各个模板的识别率及 MCC 值、尤登指数对比

Table 4　Recognition rate and MCC value, Youden index of each template

模板	识别数	识别率/%	MCC 值	尤登指数
a	27	51.9	0.34	0.51
b	26	50.0	0.28	0.49
c	42	80.8	0.59	0.80
d	52	100.0	0.34	0.97
e	36	69.2	0.29	0.67
f	27	51.9	0.28	0.50
g	35	67.3	0.28	0.65
h	49	94.2	0.42	0.92
i	17	32.7	0.20	0.32
j	47	90.4	0.55	0.90
k	40	76.9	0.33	0.75
l	33	63.5	0.27	0.61
m	14	26.9	0.18	0.26
n	25	48.1	0.30	0.47
o	24	46.2	0.29	0.45
p	30	57.7	0.37	0.57
q	26	50.0	0.29	0.49
r	18	34.6	0.22	0.34

3　设计方法分析及讨论

3.1　模板坐标提取方法的讨论

利用结构匹配的 α-碳原子三维坐标提取模板相应的 α-碳原子坐标，模板的 α-碳原子坐标应该体现匹配的 α-碳原子三维坐标的聚集性. 本文中均值这一反映聚集性的参数建立了模板坐标提取方法. 在统计学中反映一组数据聚集性的参数还有调和均值、几何均值和中位数等，分别利用上述 3 参数同样可以建立模板坐标提取方法. 不同模板坐标提取方法得到的模板是否具有同一性?

为检验模板坐标提取方法对提取模板的影响，以一个家族模板的生成为例做了检验，检验结果表明，不同方法得到的模板坐标的差别不具有统计学意义.

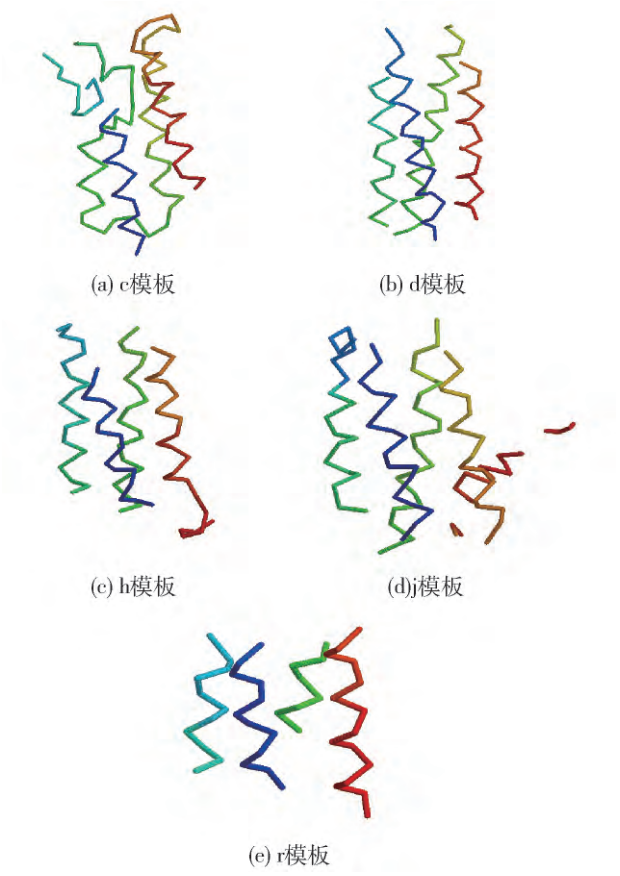


图 6　各个待选模板和 r 模板的骨架模型

Fig. 6　Skeleton model of each selected template and r template

具体检验过程如下:

选取 a.29.2.1 家族的 4 个样本，分别为 d1e6ia_、d1eqfa1、d1eqfa2、d3p1fa_，运行 MUSTANG 程序后得到的匹配位点为 113 个，分别依靠调和均值、几何均值、均值和中位数得到 4 个对应模板. 将 4 个样本的 X、Y、Z 坐标分别同 4 个模板的三维坐标 X、Y、Z 进行极距分布分析，得到极距值. 表 5 为各个样本与模板间的平均极距值.

由表 5 可知，在 X 坐标下，4 个模板的极距值分布相差不大，但调和均数模板和中位数模板相对较好，其极距值值偏低，在 Y 坐标和 Z 坐标下，调和均数模板、均值模板、几何均数模板的全局值相等，而中位数模板平均极距值除去在样本 3 处偏高之外，在其他样本处都偏低，说明中位数模板较其他 3 类模板稳定. 综合 X、Y、Z 三个坐标下的平均极距值分布，得到中位数模板较为稳定.

本文对其三维坐标的平均极距值进行单因素方差分析，检验不同模板是否对平均极距值有差异. 方差分析的前提是在各个水平下的总体服从方差相

等的正态分布,正态分布的要求并不是很严格,但方差相等的要求是比较严格的. 本文方差相等的检验方法是 homogeneity of variance test 方法,该方法是统计推断的方法,其零假设是各水平下总体方差没有显著差异,本实验显著水平选择 0.05. 如表 6 所示,为各个坐标的单因素方差结果.

表 5 各个样本与模板间的平均极距值

Table 5 Average interpolar distance between each sample and the template

模板类型	X 坐标平均全距值	Y 坐标平均全距值	Z 坐标平均全距值	样本
中位数	0.290 8	0.230 4	0.203 3	1
	0.246 9	0.297 6	0.247 2	2
	0.319 5	0.316 0	0.322 8	3
	0.192 5	0.358 7	0.286 0	4
几何平均	0.297 4	0.268 5	0.203 0	1
	0.237 9	0.328 3	0.288 1	2
	0.305 5	0.301 0	0.311 3	3
	0.211 2	0.361 3	0.295 8	4
调和均数	0.316 6	0.268 9	0.203 0	1
	0.255 5	0.329 0	0.287 8	2
	0.322 5	0.300 8	0.311 2	3
	0.222 7	0.361 2	0.295 4	4
平均值	0.303 7	0.268 3	0.203 3	1
	0.256 2	0.327 8	0.288 1	2
	0.332 0	0.301 1	0.311 2	3
	0.226 2	0.361 5	0.296 0	4

表 6 各个坐标的平均极距值的单因素方差分析

Table 6 Single factor analysis of variance of average value of each coordinate distance

平均极距值	前提检验相伴概率	方差检验 F 值	相伴概率
X 坐标	0.982	0.153	0.926
Y 坐标	0.968	0.106	0.955
Z 坐标	0.995	0.039	0.989

在 X 坐标下相伴概率为 0.982,大于显著性差异 0.05,可以认为各个组总体方差是相等的,满足方差检验的前提条件. 方差检验的 F 值为 0.153,相伴概率为 0.926,相伴概率大于显著水平 0.05,表示 4 种模板在 X 坐标下的平均极距值无明显区别,即 4 种模板无显著差别. 同样的,分别在满足方差检验的前提条件下,本实验对 Y 和 Z 坐标分别计算其相伴概率,分别为 0.955 和 0.989,都大于显著差异 0.05,说明 4 中模板在 Y 和 Z 坐标下的平均极距值无明显差别,4 种模板无显著差别.

3.2 模板提取数量及参数约束的讨论

折叠类型模板的筛选主要受折叠核心片段的组成、在系统聚类图中的分布、位置及模板的识别率限制. 由表 3 的各个模板的识别率可知,当降低识别率到 70%,能筛选出 e 模板和 k 模板,识别率分别为 69.2% 和 76.9%,通过计算 Matthew 相关系数分别为 0.29 和 0.33,尤登指数分别为 0.67 和 0.75. Matthew 相关系数反应真阳性和真阴性的相关程度,Matthew 相关系数越大说明模板对于区分真阴性和真阳性的能力越好. 尤登指数是敏感性和特异性之和减 1,指数介于 0 ~ 1,表示筛选方法发现本折叠类型样本和非本折叠类型样本的总能力,指数越大表示模板真实性越高. e 模板和 k 模板的 Matthew 相关系数在 0.3 左右,尤登指数在 0.8 以下,2 个值都较小,并且它们不是独立分支中的最先聚类形成的模板,与条件(4)违背. 当降低识别率到 60%,能筛选出 g 模板和 l 模板,其识别率分别为 67.3% 和 63.5%,MCC 值分别为 0.28 和 0.27,尤登指数分别为 0.65 和 0.61,2 个模板 MCC 值和尤登指数较小,且 g 模板是由 d 模板聚类而成,即 d 模板信息包含 g 模板信息,那么 g 模板相比 d 模板是多余的模板,可以舍弃. l 模板包含在 c 模板、d 模板和 h 模板形成的聚类区间,也可以舍弃. 当提高识别率到 90%,c 模板的识别率为 80.8% 而被舍弃,只能筛选出 d、h 和 j 模板,3 个模板对折叠类型所属家族及超家族的覆盖度降低,模板的完备性不够. 综合以上因素,将识别率定为 80%.

3.3 设计模板与天然模板的对比分析

为检验这 4 个设计模板的稳定性,分别统计了 4 个设计模板和 4 个天然模板间的等价 α-碳原子之间的距离 d_i,单位 nm.

$$d_i = \sqrt{(x_i - x_0)^2 + (y_i - y_0)^2 + (z_i - z_0)^2}$$

式中(x_i, y_i, z_i)和(x_0, y_0, z_0)分别代表 2 个匹配的 α-碳原子的坐标.

将 4 个设计模板 c、d、h、j 进行 MUSTANG 多结构比对之后,匹配的 α-碳原子有 59 个. 计算任意两两匹配的 α-碳原子之间的距离,得到分布图如图 7 所示. 在设计模板 c、d、h、j 所在家族内,挑选冗余结构少的天然蛋白样本作为天然模板,分别为

d1e6ia_、d2gsca1、d2bl8a1、d3qzta_，将这 4 个天然模板进行 MUSTANG 多结构比对，匹配的 α-碳原子有 70 个，计算任意两两匹配的 α-碳原子之间的距离，得到天然模板距离分布图，如图 8 所示.

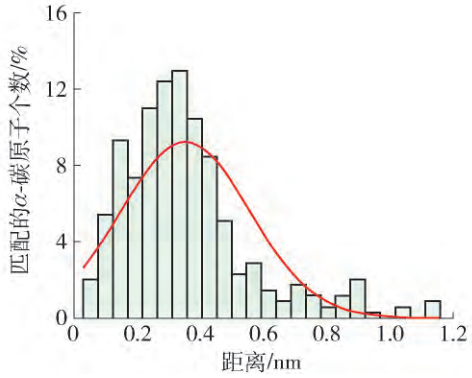


图 7　设计模板间距离的分布

Fig. 7　Distribution of distance between the design templates

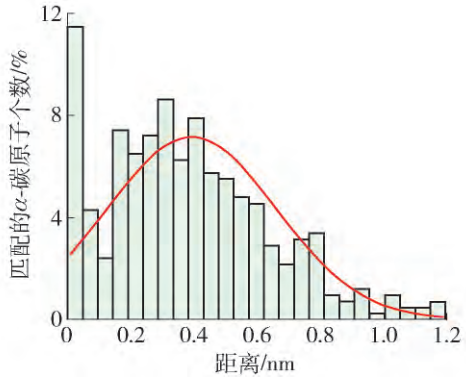


图 8　天然模板间距离的分布

Fig. 8　Distribution map of distance between natural templates

由图 7 可知，设计模板间的距离成正态分布，其平均值为 0.35 nm，平均值 95% 的置信区间为 [0.33，0.37]，标准差为 2.0. 其距离 25% 的分位点为 0.22，50% 的分位点为 0.32，75% 的分位点为 0.43. 由图 8 可知，天然模板间的距离成正态分布，其平均值为 0.30 nm，平均值 95% 的置信区间为 [0.37，0.42]，标准差为 2.7. 均值 25% 的分位点是 0.19，50% 的分位点是 0.35，75% 的分位点是 0.76. 通过以上数值分析可知：在平均值方面，设计模板间距离均值较天然模板小；在标准差方面，设计模板距离标准差较天然模板小，说明设计模板间距离更加稳定；在分位点数值分布方面，分位点表示密度函数在小于该点时与坐标横轴围成的面积，当分位点相同时，坐标横轴数值越小说明密度函数越大，即图 7、8 中的纵轴百分比越大，百分比越大说明模板距离间距数值越多，模板稳定性高. 设计模板在分位点为 25% 时，即面积为 0.25 时，所对应的横轴距离值为 0.22，而天然模板在面积为 0.25 时所对应的距离值为 0.18. 除去距离 25% 分位点时设计模板比天然模板距离值大以外，在 50% 和 75% 分位点设计模板都比天然模板的距离值小，说明设计模板稳定性好.

综合以上分析，设计模板在各个方面参数都较天然模板小，说明设计模板的空间构象比天然模板更为稳定.

3.4　模板的普适性与有效性分析

利用本文给出的方法，确定并提取 BRD 折叠类型的 4 个模板. 每个模板结构均包含该折叠类型的 4 个折叠核心片段，片段的空间坐标反映了片段的取向及片段间的排布，即设计模板成功提取了该折叠类型的叠核心片段及其取向和排布信息，具备结构上的普适性；从图 4 的系统聚类图上可以看到：4 个模板分布于各独立分支中，各自代表了其所属家族、超家族集团的结构特性，提取的 4 个设计模板代表了该折叠类型样本的共同属性；表 4 中，4 个设计模板对所属折叠类型样本的识别率均在 80% 以上，将该模板用于蛋白质折叠类型分类是有效的.

4　结论

1）针对折叠类型分类中所选天然模板的普适性不足的问题，提出了 Bromodomain-like 折叠类型模板的设计方法，并用于该折叠类型的模板设计.

2）利用该模板设计方法设计的模板，具有普适性，克服了天然模板的单一性，并且可用于蛋白质折叠类型的分类.

参考文献：

[1] ZHANG Y，SKOLNICK J. Segment assembly，structure alignment and iterative simulation in protein structure prediction [J]. Bmc Biology，2013，11(1)：1-4.

[2] 阎隆飞，孙之荣. 蛋白质分子结构 [M]. 北京：清华大学出版社，1999：211-213.

[3] ANFINSEN C B. Principles that govern the folding of protein chains. [J]. Science，1973，181(4096)：223-230.

[4] 赵国屏. 生物信息学 [M]. 北京：科学出版社，2002：160-164.

[5] BAKER D，SALI A. Protein structure prediction and structural genomics. [J]. Science，2001，294(5540)：

93-96.

[6] LUO L, LI X. Recognition and architecture of the framework structure of protein [J]. Proteins Structure Function & Bioinformatics, 2000, 39(1): 9-25.

[7] 张春霆. 蛋白质结构分类与结构类预测研究 [J]. 中国科学基金, 2000(5): 298-299.
ZHANG C T. Protein structure classification and prediction of structural classes [J]. Science Foundation in China, 2000(5): 298-299. (in Chinese)

[8] 阎隆飞, 孙之荣. 蛋白质分子结构 [M]. 北京: 清华大学出版社, 1999: 67.

[9] CHOTHIA C. One thousand families for the molecular biologist [J]. Nature, 1992, 357: 543-544.

[10] WANG Z X. How many fold types of protein are there in nature? [J]. Proteins Structure Function & Bioinformatics, 1996, 26(2): 186-191.

[11] BAKER D. A surprising simplicity to protein folding. [J]. Nature, 2000, 405(6782): 39-42.

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

[13] FILIPPAKOPOULOS P, KNAPP S. The bromodomain interaction module. [J]. Febs Letters, 2012, 586(17): 2692-2704.

[14] DHALLUIN C, CARLSON J E, ZENG L, et al. Structure and ligand of a histone acetyltransferase bromodomain [J]. Nature, 1999, 399(6735): 491-496.

[15] CONWAY S J. Bromodomains: are readers right for epigenetic therapy? [J]. Acs Medicinal Chemistry Letters, 2012, 3(9): 691-694.

[16] VOLLMUTH F, BLANKENFELDT W, GEYER M. Structures of the dual bromodomains of the P-TEFb-activating protein Brd4 at atomic resolution. [J]. Journal of Biological Chemistry, 2009, 284(52): 36547-36556.

[17] VIDLER L R, PANAGIS F, OLEG F, et al. Discovery of novel small-molecule inhibitors of BRD4 using structure-based virtual screening [J]. Journal of Medicinal Chemistry, 2013, 56(20): 8073-8088.

[18] SHINDYALOV I N, BOURNE P E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. [J]. Protein Engineering, 1998, 11(9): 739-747.

[19] HOLM L, PARK J. DaliLite workbench for protein structure comparison. [J]. Bioinformatics, 2000, 16(6): 566-567.

[20] KRISSINEL E H K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions [J]. Acta Crystallogr D Biol Crystallogr, 2004, 60(12-1): 2256-2268.

[21] ZHANG Y, SKOLNICK J. TM-align: a protein structure alignment algorithm based on the TM-score [J]. Nucleic Acids Research, 2005, 33(6): 2302-2309.

[22] 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.

[23] KIFER I, NUSSINOV R, WOLFSON H J. GOSSIP: a method for fast and accurate global alignment of protein structure [J]. Bioinformatics, 2011, 27(7): 925-932.

[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 [J]. 物理化学学报, 2009(12): 2558-2564.
LIU Y, LI X Q, XU H S, et al. Classification modeling and recognition of protein fold type [J]. Acta Physico-Chimica Sinica, 2009(12): 2558-2564. (in Chinese)

(责任编辑 杨开英)