

文章编号:1673-1689(2007)05-0038-06

一种识别蛋白质的保守氨基酸残基的新方法

郭俊恩, 王士同, 徐红林

(江南大学 信息工程学院, 江苏 无锡 214122)

摘要: 保守氨基酸残基在生物进化过程中具有较强的稳定性, 一般不会发生太大的变化; 生物信息学认为: 氨基酸序列决定蛋白质结构, 蛋白质结构决定蛋白质功能。因此, 作者基于氨基酸的生化特性、几何特性和动态特性描述了一个新颖的算法去发现蛋白酶家族中的保守氨基酸残基, 而保守氨基酸残基将对蛋白质的结构和功能的研究具有重要的意义。

关键词: 保守氨基酸残基; 生物信息学; 序列; 结构; 功能; 蛋白酶家族

中图分类号: TP 18

文献标识码: A

A New Method of Identifying Protein Conserved Amino Residues

GUO Jun-en, WANG Shi-tong, XU Hong-lin

(School of Information Technology, Jiangnan University, Wuxi 214122, China)

Abstract: Conserved amino residues have good stability and do not change during the process of the biology evolution. it is accepted that: amino sequence decides protein structure and protein structure decides protein functions in bioinformatics. Based on these facts, a novel algorithm is designed to identify each protein conserved amino residues of the protease families. The proposal algorithm is based on the dynamic, biochemical and geometric features of amino acid. By the way, the conserved amino residues play key role on the protein structural and functional.

Key words: conserved amino residues; bioinformatics; sequence; structure; function; protease family

在残基-残基比对中, 某些位置的氨基酸残基相对于其它的残基具有较高的保守性, 这个信息揭示了某些残基对于一些蛋白质的结构和功能是极为重要的^[1]。处于活性位点的残基都是极为保守的, 比如形成二硫键的半胱氨酸, 参与电子传递的氨基酸残基以及决定底物特异性的氨基酸残基。这些保守残基对于保持蛋白质的结构和功能是非常重要的。

理解分子生物学功能基础时, 对蛋白质三维结构的理解是主要的一步。然而理解蛋白质功能机制的知识是不充足的, 因为生物功能经常依赖它的动态构象。通常, 蛋白质的功能是和特殊的序列或结构相关的, 并且, 功能模式的识别和它们在蛋白质的整个动态特性中的作用要求额外的数据和分析。在蛋白质数据库(PDB)^[2]中除了大量的结构是以实验的方法得到的外, 还有很多是用计算机方法

收稿日期: 2006-08-22.

基金项目: 江苏省自然科学基金项目(BK2003017).

作者简介: 郭俊恩(1978-), 男, 河南唐河人, 计算机应用技术专业硕士研究生.

通讯作者: 王士同(1964-), 男, 江苏镇江人, 教授, 博导, 主要从事人工智能研究. Email: wxwangst@yahoo.com.cn

得到的,并且用计算机得到的数量以指数增长,这些方法通常是基于序列比对的,现在已经用来取代实验的方法了,然而还没有系统的引入结构的动态特性。利用模式识别的方法和以前的生物学知识可能找到具有统计意义的模式,并且能用实验测试和核实这种模式。

通常报导少量的残基是直接参与蛋白质功能的。这些暗示了在功能和微环境之间有很强的相互关系。假如残基是在封闭的空间的话微环境指的是残基的局部区域,而不是必须沿着序列邻近的。蛋白质功能只是一个结构属性的集合,并且可以想象的到,整个蛋白质的动态性是和蛋白质的活性部位的局部结构相结合的。

从原始文献资料中手工抽取信息建立了一个催化剂残基数据集^[3]。根据残基的二维结构、溶解力、适应性、保守性、四维结构和功能,对178种的酶活性部位直接引入的具有催化作用的残基做了一个十分彻底的分析。这项工作提供了比较好的对分子特征的理解,这些特征影响催化剂的功能,尤其是适应性。然而它没有提供自动检索信息的功能或者发现新的频度模式的方法。

几个基于网站的研究相似结构的数据库在公共域中是可以免费利用的。PROCAT利用了一种几何散列算法研究和建立了从保守几何学中得来的三维酶活性部位模板^[4,14]。WEBFEATURE^[5-6]为场微环境的简洁特性而采用了一种贝叶斯监督学习算法,然而这个结果对假定的后台分类和训练场所的选择是过敏的。PINTS服务^[7]允许以3种不同的方式对大量的有理由的预定义的结构和功能动机进行研究:蛋白质对模式数据库;模式对蛋白质数据库;蛋白质成对比较。在没有任何相关的知识的情况下,当它探测到蛋白质结构中的侧链的空间排列的相似性时,这项工作是非常有意义的。在寻找到最佳的匹配结构模式的叠印之后,它也设计出了一种统计方法去计算等价氨基酸空间位置间方均根的意义^[7]。然而这种算法在非同源蛋白质中能找到局部模式,特别是如果两种相似的蛋白质进行比较时将是难以忍受的。这种算法排斥带有仅仅包含H和C原子侧链的氨基酸,不能在正确与错误的匹配间明确而有效地区分这些原子。

已经证明两种其它的无监督学习方法成功的产生了新颖序列结构模式。I-site^[8]是一种短序列模式库,它与蛋白质的三维结构强有力的联系起来了,也为局部结构预测提供了一种新的方法论。TRILOGY^[9]把序列和结构都视作是模式的成分,

它们在搜索过程中同时被识别和延伸。在这项工作中发现了成千上百的重要的序列-结构模式。但是,保守氨基酸残基在蛋白质序列中还没有进行必要的认识,并且它们能以任何顺序产生,比如胰岛素,催化剂三元组。这些算法不能探测到缺少序列-模式成分的模式。

Shann-Ching Chen^[10]首先用一种无监督学习算法发现蛋白质家族的频度模式,然后再利用频度模式识别蛋白酶家族中每种蛋白质的保守氨基酸残基。除了序列和结构相似外,还考虑了结构的动态性。因此这些模式的特征是根据它们在微环境中的动态特性、生物化学特性和几何学特性所刻画的。没有经过任何序列比对,能把结构上的保守残基识别出来,这些残基的序列和顺序是没有必要进行维持的。实验表明在探测的模式中能正确的区分3种蛋白酶子家族微环境中的不同模式。但是这种方法在微机上的运行速度较慢,它的时间主要花费在合并特征链上了。

作者基于氨基酸的生化特性、几何特性和动态特性,描述了一个能快速发现蛋白质酶家族中保守氨基酸残基的新颖算法。这个算法用较短时间就可以发现丝氨酸蛋白酶家族中的两个子家族胰凝乳蛋白酶子家族和枯草杆菌蛋白酶子家族,以及巯基丙氨酸蛋白酶家族中的木瓜蛋白酶子家族中的保守氨基酸残基。目前,对于数据量呈指数级增长的蛋白质数据来说,这种方法是发现保守氨基酸残基的一种较有前途的方法。

1 数据来源和方法

整体思路由图1表示:首先选出给定的蛋白酶家族中的蛋白质集合作为数据集。特征从这个家族中的全部氨基酸中抽取。每种氨基酸对应一个氨基酸类型+索引号(以下简称氨基酸索引),索引号代表了该氨基酸在氨基酸序列中的位置。然后对每个氨基酸索引进行特征抽取,这些特征包括残基的动态特性、生化特性和几何特性。最后根据氨基酸的类型把氨基酸索引分成20组。这种迭代算法被应用到每一组氨基酸中去发现不同蛋白质相应的氨基酸残基所对应的保守氨基酸残基。所有能展现氨基酸保守特征的残基都能被一个迭代搜索算法识别出来。

1.1 数据集

酶分类数据库^[11]包含了780个对应于丝氨酸蛋白酶类 E. C. 3. 4. 21 的 PDB 入口和 122 个对应于巯基丙氨酸蛋白酶类 E. C. 3. 4. 22 的 PDB 入口。

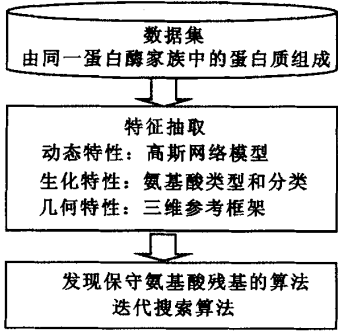


图 1 识别保守氨基酸残基方法的概述

Fig. 1 The overview of the method of identifying conserved amino residues

在这个工作中,两个最大的子家族,丝氨酸蛋白酶的胰凝乳蛋白酶子家族(S1)和枯草杆菌蛋白酶子家族(S8),以及巯基丙氨酸蛋白酶的木瓜蛋白酶子家族(C1)是作者的研究对象。为降低结构冗余,90%的序列特性是用 PDB-REPRDB 从 PDB 中选择的具有代表性的侧链,S1、S8 和 C1 子家族分别产生了 5、7 和 6 个代表性蛋白质,这些蛋白质包含在作者对各个子家族的分析中了:S1 包括了 1gct, 1sgt, 1ton, 3est 和 4ptp 5 个蛋白质;S8 包括了 2sec, 1lw6, 1bh6, 1svn, 1mee, 1ic6 和 1thm 7 个蛋白质;C1 包括了 1huc, 1the, 3pbh, 1icf, 1ppn 和 1ppo 6 个蛋白质。所有代表 S1、S8 和 C1 子家族的蛋白质都列在了表 1 中。

表 1 3 个蛋白酶家族中的代表性蛋白质

Tab. 1 The representative proteins in the three protease families

蛋白酶子家族	代表性蛋白质
S1	Gamma-chymotrypsin A, 1gct, 3.4.21.1
	Trypsin, 1sgt, 3.4.21.4
	Tonin, 1ton, 3.4.21.35
	Native elastase, 3est, 3.4.21.36
	Beta trypsin, 4ptp, 3.4.21.4
S8	Subt. carlsberg complex, 2sec, 3.4.21.62
	Complex of subt. bpn, 1lw6, 3.4.21.62
	Subtilisin dy in complex, 1bh6, 3.4.21.62
	Savinase, 1svn, 3.4.21.62
	Mesentericopeptidase, 1mee, 3.4.21.62
	Proteinase k, 1ic6, 3.4.21.64
	Thermitase, 1thm, 3.4.21.66
C1	Cathepsin b, 1huc, 3.4.22.1
	Cathepsin b, 1the, 3.4.22.1
	Procathepsin b, 3pbh, 3.4.22.1
	Cathepsin l, 1icf, 3.4.22.15
	Papain cys-25, 1ppn, 3.4.22.2
	Protease omega, 1ppo, 3.4.22.30

1.2 氨基酸特征抽取

每个残基附近的微环境被定义为以它的 C^α 为中心的以 7.0×10⁻⁹ cm 为半径的球形区域。对每种氨基酸的特征进行动态特性、生化特性和几何特性的抽取。

1.2.1 动态特性的抽取:高斯网络模型 高斯网络模型(GNM)^[12]是一个弹性的网络模型,它能描述蛋白质的动态平衡性,在 7.0×10⁻⁹ cm 的范围内通过统一的弹性系数,可以用来刻画蛋白质的动态特性。在 GNM 中,C^α形成网络结点。结点相互作用。网络的连接能力用基尔霍夫矩阵描述如下:如果元素 i 和 j 连接,则元素 Γ_{ij} = 1,否则 Γ_{ij} = 0,并且 Γ_{ii} = -Z_i,其中 Z_i 是连接到结点 i 的结点总数,以下简称 CN。

另外 CN 在大量的研究中已经被证明对动态平衡性有很大的影响。PDB 结构实验显示连接数有一个很大的变动范围(2≤CN≤16)。因此可见的连接数被映射到 4 个连续的水平 CN-1、CN-2、CN-3、CN-4,它们代表的范围分别是 2≤CN≤4、5≤CN≤8、9≤CN≤12 和 13≤CN≤16。

在特征抽取中,首先需要从 PDB 数据库中下载对应的蛋白质文件,如 S1 中的 1gct 等,然后再从蛋白质文件中找到氨基酸序列中每个氨基酸的 C^α 的空间坐标,最后这项工作需要找出每个残基的连接数水平以及与该氨基酸相连接的每一个残基作为抽取的氨基酸动态特性。如与 S1 中蛋白质 1gct 的组氨酸 His57 相邻的氨基酸残基为 5 个,分别是 ALA、ALA、CYS、ASP 和 GLY,因此蛋白质 1gct 的组氨酸 His57 残基的 CN 达到了 CN-2。

1.2.2 生化特性的抽取:氨基酸的类型和分类 20 种常见氨基酸按照氨基酸的特性和侧链化学性质的不同进行如下分类^[13]:脂肪类 (AL1) 的包括 Gly, Ala, Val, Len 和 Ile;含硫类(SULFUR)的包括 Cys 和 Met;芬芳类(ARO)的包括 Phe, Tyr, Asn 和 Gln;极性类 (NEUTR) 的包括 Ser, Thr, Asn 和 Gln;酸性类(ACID)的包括 Asp 和 Glu;碱性类(BASIC)的包括 Lys, Arg 和 His;亚氨基酸类(IMI)的包括 Pro。由于同一种氨基酸可能不仅属于一类,如赖氨酸残基(Lys)有两种特征含硫类和碱性类。因此在特征提取过程中作者要判断与某一个氨基酸残基相邻的所有氨基酸残基属于哪一类,并把它作为该氨基酸残基的一个生化特性。如与 S1 中蛋白质 1gct 的组氨酸 His57 残基相邻的 5 个氨基酸残基分别具有的生化特性是 ALI、ALI、SULFUR、ACID 和 ALI。

1.2.3 几何特性的抽取:三维参考框架 三维参考框架利用氨基酸的3个骨干原子N、C^{*}和C'来描述每一个残基,这3个原子形成了一个框架三棱锥^[4]。在图2(A)中,3个原子唯一定义了残基在空间中的位置和方向,三维参考框架的坐标原点为C^{*},坐标是一样的。在图2(B)中定义了3个方向的向量 e_1 、 e_2 和 e_3 。在图2(C)中每个残基的微环境被 $\pm e_1$ 、 $\pm e_2$ 和 $\pm e_3$ 分成8个区域,即Q1~Q8。然后利用一个几何哈希算法^[3,13]判断与该氨基酸残基相邻的氨基酸残基以及其对应的生化特性分别属于哪个区域。如与S1中蛋白质1gct的组氨酸His57残基相邻的5个氨基酸残基及其生化特性在组氨酸His57的8个区域中的分布情况如下:ALA0-Q3, ALI0-Q3, ALA1-Q3, ALI1-Q3, YS0-Q7, ULFUR0-Q7, GLY0-Q7, ALI0-Q7, ASP0-Q4, ACID0-Q4。

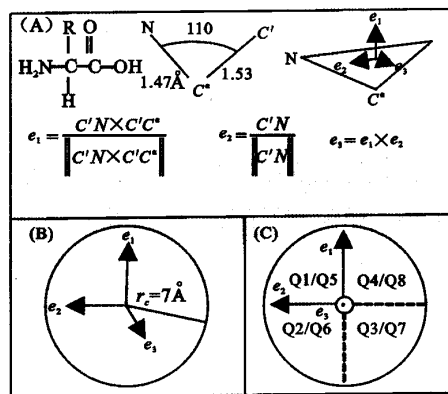


图2 氨基酸几何特性抽取的方法

Fig. 2 The method of extracting amino acid geometric features

1.2.4 一个完整的特征抽取的实例:1gct-Ser-63 作者用PDB文件1gct为例对丝氨酸Ser-63进行特征抽取,抽取过程如下:利用高斯网络模型可以发现与Ser-63相连的残基数为5,因此连接数达到了CN2。这5个残基分别是Lys、Thr、Thr、Asp和Val,它们分别具有以下化学性质碱性(BASIC)、中性(NEUTR)、中性(NEUTR)、酸性(ACID)和脂肪类(ALI)。最后运用三维参考框架可以发现在该残基的微环境中的8个区域内分别分布了Ser-63的以下特征(在同一个区域中存在相同的多个残基或者化学性质时作者用0,1,2...对其进行编号):在Q1内有THR,THR具有NEUTR的性质;在Q2内有THR,THR具有NEUTR的性质;在Q4内有ASP和VAL,分别具有ACID和ALIRR的性质;在Q8内有LYS具有BASIC性质。因此作者最终

抽取的PDB文件1gct中的残基Ser-63的特征就是:CN2、THR0-Q1、NEUTR0-Q1、THR0-Q2、NEUTR0-Q2、ASP0-Q4、ACID0-Q4、VAL0-Q4、ALI0-Q4、LYS0-Q8和BASIC0-Q8。

1.3 发现保守氨基酸残基的算法:迭代搜索算法

发现保守氨基酸残基的迭代搜索算法描述如下:

1) 设定要选取的特征链中的节点元素个数的最大值MAX、最小值MIN和向前查看长度LEN,同时令循环控制变量N=MAX。

2) 由抽取的氨基酸特征构造同一蛋白酶家族中某一类氨基酸残基(如Ser)的原始表,原始表每一项记录包括两个字段:氨基酸索引和该氨基酸残基对应的特征。

3) 由原始表构造工作表。工作表每条记录包括3个字段:特征、绝对支持率和相对支持率。绝对支持率指某一特征在原始表中出现的次数;相对支持率指的是绝对支持率除以原始表的表长所得到的一个比值。当然在创建工作表时绝对支持率和相对支持率均为0。

4) 查原始表并统计工作表中的每个特征的绝对支持率和相对支持率。

5) 对工作表按绝对支持率或者相对支持率进行排序。

6) 工作表的表长如果大于等于MIN则转到7)执行;否则转到9)执行结束程序。

7) 取工作表中支持率最高的N项特征,同时考虑到最大可能性要向前查看LEN个特征共同组合成 C_{N+LEN}^N 个特征链表。

8) 检查7)中产生的每一个特征链表,看该记录是否在同一蛋白酶家族中的每一个蛋白质中至少产生一次,如果是则查原始表找出该特征集合所对应的原始表中的记录并将其输出,找出的记录所对应的氨基酸索引就是要找的保守氨基酸残基,同时将原始表中这条记录删除,并令 $N=N-1$,转到3)执行;否则只令 $N=N-1$,转到6)执行。

9) 结束程序的执行。

该算法中第1)步的参数MAX、MIN和LEN是待定的,根据实验结果一般情况下MAX的取值为20,MIN的取值为5,LEN的取值为1~5;第7)步的特征链表的长度可以由组合数计算得到,即 C_{N+LEN}^N ,其中N为要选取的特征集合的元素个数,LEN为向前查看的元素个数。

在同一蛋白酶子家族中把该迭代算法应用到20种氨基酸上,就可以找出每一蛋白酶子家族中每一种氨基酸所对应的保守氨基酸残基。

表 2 S1 的代表性蛋白质 1gct 的保守氨基酸残基
Tab. 2 The conserved amino residues of the representative protein 1gct in S1

实验结果	HIS	ASP	SER	ALA	CYS	GLY	TRP	ILE	OTHERS
Shann-Ching	57	194	195	55	42	196,44	141	212	N/A
Chen 的实验		102	214	56	191	142,140			
结果				183	58	43,193			
					182	197,211			
作者的	57	102	195	55	42	196,44	141	212	N/A
实验结果	40	194	214	56	191	142,140	27	6	
		35	26	183	58	43,193			
		64		5	182	197,211			
				22	122	184,187			
					201	502			

表 3 S8 的代表性蛋白质 1svn 的保守氨基酸残基
Tab. 3 The conserved amino residues of the representative protein 1svn in S8

实验结果	HIS	ASP	SER	ALA	ASN	LEU	MET	PRO	THR	VAL	PHE	OTHERS
Shann-Ching	64	32	125	223	155	96	222	201	66	177	66	N/A
Chen 的实验	67	181	190	228		126		225	220		189	
结果	226		207			233						
			221									
作者的	64	32	125	223	155	96	222	201	66	177	66	N/A
实验结果	67	181	190	228	18	126	119	225	220	93	189	
	226	41	207	69	62	233		131	57	95	50	
	17	60	221			257			71			
		197	106			262						
			166									
			265									

表 4 C1 的代表性蛋白质 1huc 的保守氨基酸残基
Tab. 4 The conserved amino residues of the representative protein 1huc in C1

实验结果	HIS	CYS	ASN	ALA	GLN	GLY	TRP	ILE	LEU	PHE	PRO	SER	TYR	OTHERS
Shann-Ching	199	29	219	31	23	229	30	201	216	32	76	28	146	N/A
Chen 的实验		26		34		27	221	203				152	214	
结果		71				73						220	94	
		62				70								
作者的	199	29	219	31	23	229	30	201	216	32	76	28	146	N/A
实验结果	239	26	72	34	13	27	221	203	181	5	251	152	214	
		71	149	3		73	80	44	182	231		220	94	
		62		7		70						4	11	
		14				24							83	
		43				68								

2 结果与讨论

在作者的实验中,分别用蛋白质 1gct 代表丝氨酸蛋白酶家族中的胰凝乳蛋白酶子家族 S1,用蛋白质 1svn 代表丝氨酸蛋白酶家族中枯草杆菌蛋白酶子家族 S8,用蛋白质 1huc 代表巯基丙氨酸蛋白酶家族中的木瓜蛋白酶子家族 C1,得出的结果列到表 2、表 3 和表 4 中,并与 Shann-Ching Chen^[10] 的实验结果进行了比较。

上述表 2、表 3 和表 4 中的第一行为氨基酸残基类型,第二行和第三行为残基在氨基酸残基序列中对应的索引号,最后一列中 Others 代表其它氨基酸残基,N/A 代表没有保守氨基酸残基。

从上述结果可以看出 Shann-Ching Chen 的实验结果是作者的一个子集,这是因为 Shann-Ching Chen 采用的是逐步合并氨基酸残基的特征使其长度不断增大并最终求出一个含有特征支持率最高的一个特征链,然后利用这个特征链再找出保守氨基酸残基,但这种方法运行速度特别慢,所花费的时间主要在合并特征链上。而作者的实验采用的是对每个特征的支持率进行排序,选出支持率最高

的 N 个特征,然后再从这 N 个特征中选择 n 个特征作为节结元素组成一个特征链($n \leq N$),这样就组合出 C_N^n 个特征链,最后利用这 C_N^n 个特征链找出所有的保守氨基酸残基。可见 Shann-Ching Chen 找出的特征链只是找出的中的特征链中的一个,显然他的结果是作者的子集。作者简化了逐步合并氨基酸残基的特征使其长度不断增大并最终求出一个含有特征支持率最高的一个特征链这一步,从而节省了大量的时间,因此作者的算法要比 Shann-Ching Chen 的算法快。

3 结 语

作者通过对同一蛋白酶家族中每一个蛋白质的氨基酸残基进行动态特性、生化特性和几何特性的抽取,然后用一个迭代搜索算法对同一蛋白酶家族中的 20 种氨基酸进行处理,结果可以发现实验用到的 3 个蛋白酶家族中的保守氨基酸残基。在这个基础上,作者下一步的工作可以研究从保守氨基酸残基中发现同一蛋白酶家族的催化剂三元组,还可以进一步研究氨基酸序列上的保守氨基酸残基对蛋白质结构以及功能的影响。

参考文献(References):

- [1] Andreas D Baxevanis, B. F. Francis Ouellette. 生物信息学基因和蛋白质分析的实用指南[M]. 李衍达译. 北京:清华大学出版社, 2000.
- [2] Berman HM, Westbrook J, Feng Z et al. The Protein Data Bank[J]. *Nucleic Acids Res*, 2000, 28:235—242.
- [3] Bartlett G J, Porter C T, Borkakoti N, et al. Analysis of catalytic residues in enzyme active sites[J]. *J Mol Biol*, 2002, 324:105—121.
- [4] Wallace A C, Borkakoti N, Thornton J M. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases: application to enzyme active sites[J]. *Protein Sci*, 1997, 6: 2308—2323.
- [5] Bagley S C, Altman R B. Characterizing the microenvironment surrounding protein sites[J]. *Protein Sci*, 1995, 4:622—635.
- [6] Liang M P, Banatao D R, Klein T E, et al. Web Feature: an interactive web tool for identifying and visualizing functional sites on macromolecular structures[J]. *Nucleic Acids Res*, 2003, 13:3324—3327.
- [7] Stark A, Russell R B. Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures[J]. *Nucleic Acids Res*, 2003, 13:3341—3344.
- [8] Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence structure motifs[J]. *J Mol Biol*, 1998, 281:565—577.
- [9] Bradley P, Kim P S, Berger B. TRILOGY: discovery of sequence structure patterns across diverse proteins[J]. *Proc Natl Acad Sci USA*, 2002, 13:8500—8505.
- [10] Shann-Ching Chen, Ivet Bahar. Mining frequent patterns in protein structures: a study of protease families[J]. *Bioinformatics*, 2004, 20 (Suppl. 1): i77—i85.
- [11] Bairoch A. The Enzyme data bank[J]. *Nucleic Acids Res*, 1993, 21:3155—3156.
- [12] Sibsankar Kundu. Dynamics of proteins in crystals: comparison of experiment with simple modes[J]. *Biophysical Journal*, 2002, 83:723—732.
- [13] Koolman J Rohm. Colour Atlas of Biochemistry[M]. Georg Thieme Verlag; Thieme Medical Publishers, 2005.
- [14] Xavier Pennec, Nichloas, Ayache. A geometric algorithm to find small but highly similar 3D substructures in proteins[J]. *Bioinformatics*, 1998, 14(6):516—522.

(责任编辑:朱明)