蛋白质功能基团三维模体及其应用

叶玉珍 解 涛 丁达夫*

(中国科学院上海生物化学研究所, 上海 200031. * 联系人, Email: dingdafu@server.shcnc.ac.cn)

摘要 用一维序列模体和三维结构模体刻画与识别蛋白质功能区是蛋白质功能预测和分子设计中的重要课题. 目前三维模体的提取与搜索均以残基为单位进行, 特异性有限. 鉴于残基的功能基团才是其发挥功能的关键要素, 提出以功能基团为单位来表征三维模体(称为功能基团三维模体), 并发展了相应的搜索算法, 用于阐释蛋白质功能的结构基础以及预测未知蛋白质的功能. 以从胰蛋白酶(PDB 代码为 1mct)中提取的"三联体和氧穴"功能基团三维模体及其搜索为例, 在整体结构极不相似的胰蛋白酶、枯草杆菌蛋白酶和α/β水解酶中均搜索到了此功能基团三维模体, 为三者具有相似催化机制提供了解释. 对一些具有"变异三联体"的酶亦能用此功能基团三维模体给予识别. 对蛋白质结构库的搜索结果充分展示了其在功能预测中的应用. 与其他三维模体搜索及序列模体搜索的结果相比较表明, 功能基团三维模体是蛋白质功能区的更好的表示, 在蛋白质功能预测中有更高的灵敏度.

关键词 功能基团三维模体 距离矩阵均方根偏差 概率 期望值 功能预测

蛋白质的序列、结构和模体(Motif, 包括序列模体和结构模体)的比较是获取其功能信息的重要手段[1~3]. 为了在远缘同源和趋同进化情况下仍能获取蛋白质的功能信息, 许多学者发展了以残基为最小单位, 用残基的主链骨架(如 C_α原子), 或是侧链的代表原子组成的空间结构来刻画蛋白质的活性部位的三维模体方案及其搜索方法[4~9], 并用于基因的功能预测[4~9]和蛋白质分子设计[10~12]. 然而, 有些情况下这类三维模体方案不能有效地对功能区进行刻画和识别. 如胰蛋白酶、枯草杆菌蛋白酶、一些α/β水解酶, 整体结构很不相似, 但具有相似的催化机制, 其功能的结构基础是三联体, 为检测三维模体方案的可行性的很好例子[6,7]. 但最近发现的很多三联体"变异体", 如天冬酰胺酶中以 Lys 替代 His 执行碱的功能, 以 Thr 替代 Ser 执行亲核功能; 疥疮病链霉菌酯酶(*streptomyces scabies* esterase)中不是由 Asp 或 Glu, 而是由残基 Trp 的主链羰基执行酸的功能等[13,14], 这些蛋白质不能用基于残基的三维模体方案加以有效的识别. 另外很多实验表明, 除了三联体, 氧穴也是这些酶, 包括丝氨酸蛋白酶、脂酶以及羧肽酶发挥功能的重要组成部分[15]. 氧穴一般由两个或多个氢键供体组成, 用于稳定催化过程中酰化酶过渡态. 但是由于其不具备残基特异性, 该部位也难以用基于残基的三维模体方案加以描述, 因而需要发展一种更为细致的模体描述方法.

本文提出以功能基团这一更细致的单位来表征蛋白质三维模体, 称为功能基团三维模体(简称功能基团模体), 以期对蛋白质功能区作出更精确的描述. 本方案可以用于主链基团与侧链基团的匹配; 同一残基的两个基团与来自于两个残基的两个基团的匹配; 有相似基团而整体残基不相似的匹配等问题, 而这些问题是基于残基的表示方案难以解决的. 同时, 在搜索算法上, 采用深度优先策略寻找目标蛋白质中的匹配, 并建立统计显著性指标对结果进行检验. 对丝氨酸蛋白酶的催化三联体这一典型例子, 实现了 "三联体和氧穴" 功能基团三维模体的描述, 并进行了该功能基团三维模体的目标搜索, 显示其在功能预测及结构-功能关系研究中的应用.

1 数据与方法

(ⅰ) **数据**. (1) **猪胰蛋白酶**(PDB **代码为** 1mctA) **用于功能基团三维模体的构建**. (2) **非冗余蛋白质库** 1(**共** 968 **个蛋白质**) **用于功能基团三维模体随机分布的统计**. **从网站** [1] **提供的蛋白质聚类结果**, **取每一类的第** 1 **个蛋白质构成该非冗余蛋白质库**. (3) **蛋白质测试集**. Russell **搜索方法得到的可能具有三联体的蛋白质**[7]. α/β**水解酶**, **蛋白质列表来源于** ESTHER **网站** [2]. **疥疮病链霉菌酯酶** (*streptomyces scabies* esterase)(PDB **代码为** 1esc, 1esd **和** 1ese)[14]. **非冗余的蛋白质库** 2[8] (**共** 2 098 **个蛋白质**, **蛋白质之间的相似性小于** 95%). **所有蛋白质的三维坐标均从** RCSB **网站** [3]**下载**. PDB **代码为** 4 **字符**, **第** 5 **个字符表示所用到的蛋白链**.

(ⅱ) **方法**. **首先利用已知的结构和功能知识构建功能基团三维模体**, **然后统计该模体各基团的组合在非冗余蛋白质库中的随机分布**, **进而利用这种分布来评估三维模体搜索的统计显著性**.

(1) **残基的功能基团拆分**. **不同于仅用** C_α **原子的简化的残基表示方法**[8]**和侧链基团的残基表示方法**[7], **我们将残基拆分为多个基团**, **包括两个主链基团**, **氨基** (**记为** NH) **和羰基** (**记为** CO), **和一个或多个侧链基团**. **这里认为不同残基的主链基团没有残基差异**. **同一基团可以有几组组成原子**, **如** Ser/Thr **的羟基**(**记为** OH), **当该基团起亲核功能**, **需考虑氧原子以及与其共价连接的碳原子**(**见图** 1(d)**的** 1 **方框**)**的空间坐标**; **而当该基团作为氢键供体的时候**, **考虑氧原子及其上的氢原子的空间坐标**(**见图** 1(d)**的** 2 **方框**). **氢原子的坐标根据** CHARMM **力场**[16] **中的标准键长**、**键角和二面角搭建**. **图** 1 **示意几种残基的基团分割情况**.

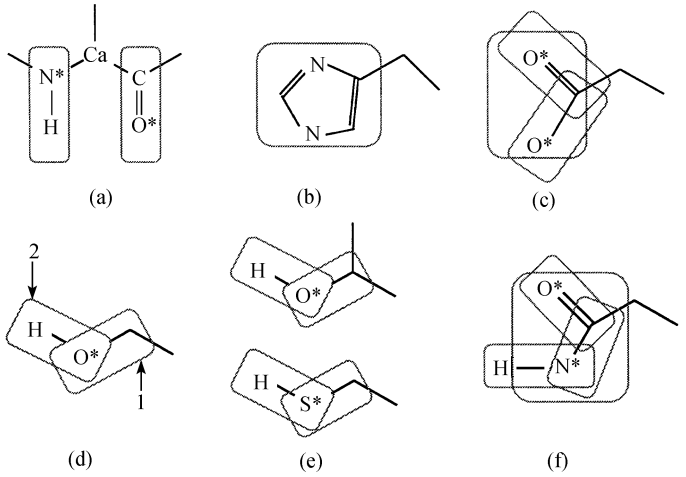

图 1 **显示部分残基的功能基团拆分**

方框中为各功能基团的组成原子, **其中未标出原子类型的均为碳原子**. * **表示该原子作为功能基团的中心原子**

每种类型的基团赋予一个理化性质谱面, **分别表示疏水性**、**氢键供受体**、**亲核性**、**芳香性**、**带电性等**, **根据这个谱面计算任意一对基团之间的相似性分数**(**取值范围为**[0~1]), **用于**

1) http://www-lmmb.ncifcrf.gov/~tsai/index.html
2) http://www.ensam.inra.fr/cholinesterase
3) http://www.rcsb.org

刻画基团间的相似性.

(2) 功能基团三维模体的定义. 假如已知一个蛋白质活性位点的关键残基, 将这些关键残基拆分为功能基团, 然后选取其中与活性直接相关的基团来定义 3D 功能基团三维模体. 一个功能基团三维模体, 包括所有组成基团的类型及其组成原子的空间坐标. 每个基团赋一个相似性阈值表示不同的相似性要求, 如相似性阈值为 1.0, 表示搜索过程中只有理化性质完全相同的基团才能与之匹配; 小于 1.0 表示允许相似基团的匹配.

以由两个基团组成的模体为例(如图 2(a) 所示的 A 和 B 组成的模体), 每个基团分别由两个原子组成, 基团组成原子之间形成一个距离矩阵(d_1, d_2, d_3, d_4), 假定在一个蛋白质中找到与功能基团 A 和 B 相似的基团 A′ 和 B′, 如图 2(b)所示. 定义这两个结构之间距离矩阵的均方根偏差 (记为 $\mathrm{RMSD_{dm}}$, distance matrix RMSD) 来刻画它们之间的立体结构相似程度, 低的 $\mathrm{RMSD_{dm}}$ 值表示高的立体结构相似性.

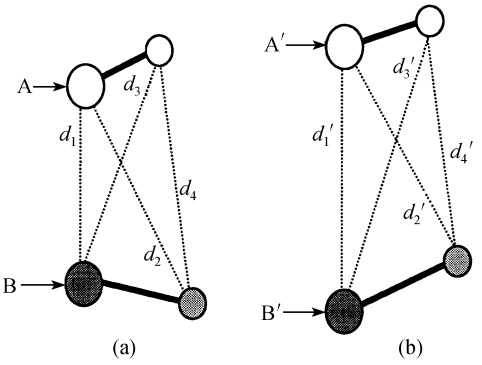

图 2　三维功能基团模体示意图

(a) 定义的功能基团三维模体; (b) 一个与功能基团三维模体匹配的局部结构. 图中虚线表示功能基团组成原子之间的距离

$$\mathrm{RMSD_{dm}} = \sqrt{\sum_{i=1}^{n} (d_i - d_i')^2 \,/\, n}, \tag{1}$$

其中 n 为模体组成基团所包含的原子之间形成的距离数目, 该例 $n = 4$.

(3) 功能基团三维模体在蛋白质结构库中的随机分布. Wallace 等[6]的工作表明, 有活性的三维模体和具有相似或相同基团组成但无功能的局部结构之间在立体结构上存在一定的差异, 为对三维模体搜索中找到的匹配进行评估提供了一定的依据. 给定一个功能基团三维模体, 依次对非冗余蛋白质结构库中各蛋白质结构进行搜索来统计该模体组成基团在蛋白质结构空间中的分布情况. 以这个分布作为参照, 可以对搜索到的匹配进行评估. 首先将各蛋白质拆分为功能基团并按基团类型聚类; 采用深度优先策略, 依次搜索其中与给定模体相匹配的基团组合 (这里没有结构相似性要求), 每得到一个完整的功能基团组合即用(1)式计算此局部结构与给定模体之间的 $\mathrm{RMSD_{dm}}$. 搜索结束后将所有的 $\mathrm{RMSD_{dm}}$ 进行排队, 计算得累积概率值 $P(x)$, 用于表示一个局部结构与给定模体随机匹配导致 $\mathrm{RMSD_{dm}} < x$ 的概率. 在 "三联体和氧穴"这个功能基团三维模体例子中, 对 $x < 3.0$(认为 RMSD>3.0 为无意义的匹配)进行 $\log(P(x))$ 与 $1/x$ 线性拟合((2)式), 相关系数大于 0.94, 表明两者之间成很好的线性关系.

$$\log(P) = a + \frac{b}{\mathrm{RMSD_{dm}}}. \tag{2}$$

每个功能三维模体有相应的一个 a, b 值, 经拟合的 a, b 参数代入上述公式可用于概率的计算.

(4) 功能基团三维模体的搜索. 功能基团三维模体的搜索采用深度优先算法, 并加以一定的距离约束, 以加快搜索. 记一个三维模体由 n 个功能基团 M_a M_b,…, M_n 组成, 在目标蛋白中

搜索得到一个匹配结构为 T_a, T_b, …, T_n, 它们之间的距离矩阵均方根偏差为 $\mathrm{RMSD}_{\mathrm{dm}}(M_i, T_i, 1 \leqslant i \leqslant n)$, 根据该三维模体相应的 a, b 参数和 $\mathrm{RMSD}_{\mathrm{dm}}$, 利用公式(2)得概率值 $P(\mathrm{RMSD}_{\mathrm{dm}}(M_i, T_i, 1 \leqslant i \leqslant n))$. 另外需考虑不同蛋白质的基团组成情况, 因为目标蛋白质中与三维模体中的基团相匹配的基团越多, 在目标蛋白质中找到一个与该三维模体匹配的局部结构的期望值也会越大. 记目标蛋白中与 M_i 相匹配的基团有 $N(M_i)$个, 所有与三维模体相匹配的基团组合数记为 $\mathrm{comb}(N(M_i), 1 \leqslant i \leqslant n)$. 则该目标蛋白中出现一个局部结构与功能基团三维模体之间的距离矩阵均方根偏差为 $\mathrm{RMSD}_{\mathrm{dm}}(M_i, T_i, 1 \leqslant i \leqslant n)$的期望值为

$$E = \mathrm{comb}(N(M_i), 1 \leqslant i \leqslant n) \times P(\mathrm{RMSD}_{\mathrm{dm}}(M_i, T_i, 1 \leqslant i \leqslant n)), \quad (3)$$

此距离矩阵概率和期望值小于一定阈值(对于"三联体和氧穴"功能基团三维模体, 本文概率阈值取 1.0×10^{-10}, 期望值阈值取 1), 表明在该目标蛋白质中随机出现这种局部结构的概率非常小, 因而这种匹配的统计显著性成立. 如果在一个未知功能的目标蛋白质中搜索到有显著意义的功能基团三维模体的匹配, 即可预测目标蛋白质可能具有该模体代表的功能.

2 结果

下面以丝氨酸蛋白酶的"催化三联体和氧穴"定义的功能基团三维模体为例, 统计其相应的 a, b 参数, 并在多个蛋白质测试集中进行搜索, 用于检验基于功能基团三维模体的方案的有效性和说明其在功能预测等方面的应用.

2.1 功能基团三维模体的定义

从胰蛋白酶(PDB 代码 1mct) 提取亲核基团(Ser195 的羟基, 记为 OH), 碱性基团 (His57 的咪唑基, 记为 GH), 酸性基团(Asp102 的侧链羧基, 记为 GD)以及两个组成氧穴的主链氨基(Ser195, 以及 Gly193 的主链氨基, 记为 NH)(见图 3(a)), 构建功能基团三维模体. 每个组成基团分别定义一个相似性匹配分数, 即其中咪唑基和羧基只允许相同的基团与之匹配, 羟基允许相同基团和巯基与之匹配, 而氨基允许其他能提供氢键供体的基团. 考虑到其他的酶利用主链羰基也可以执行酸的功能[13,14], 分别利用 Asp102 侧链羰基 (记为 EO^1, EO^2)构建了另外两个模体, 酸性基团允许的匹配基团包括 Asp 和 Glu 的侧链羰基 (记为 EO), 主链羰基 (记为 CO), Asn, 以及 Gln 酰氨基中的羰基(记为 QO), 其他基团相同. 分别进行上述 3 个模体距离矩阵均方根偏差分布统计, 求得各自 a, b 参数, 结果见表 1. 最终定义两个"三联体和氧穴"功能基团三维模体, 一个是酸性基团只能是羧基 GD(记为模体 1), 另外一个模体中酸性基团可以是完整的羧基 GD, 也可以是羰基 EO^1 或 EO^2(记为模体 2).

表 1 胰蛋白酶功能基团模体在蛋白质结构中的随机分布

功能基团模体	a	b	相关系数
NH(Ser195) NH(Gly193) GH(His57) GD (Asp102) OH(Ser195)	−10.050	−9.779	0.968
NH(Ser195) NH(Gly193) GH(His57) EO^1(Asp102) OH(Ser195)	−10.986	−8.530	0.941
NH(Ser195) NH(Gly193) GH(His57) EO^2(Asp102) OH(Ser195)	−8.595	−12.273	0.991

2.2 Russell 蛋白质测试集[7]的模体搜索结果

用模体 1 搜索 Russell 蛋白质测试集[7], 详细结果见网页 1), 总的来看, 用胰蛋白酶功能基

1) http://dna.sibc.ac.cn/ ~ ye/test 1 table. html

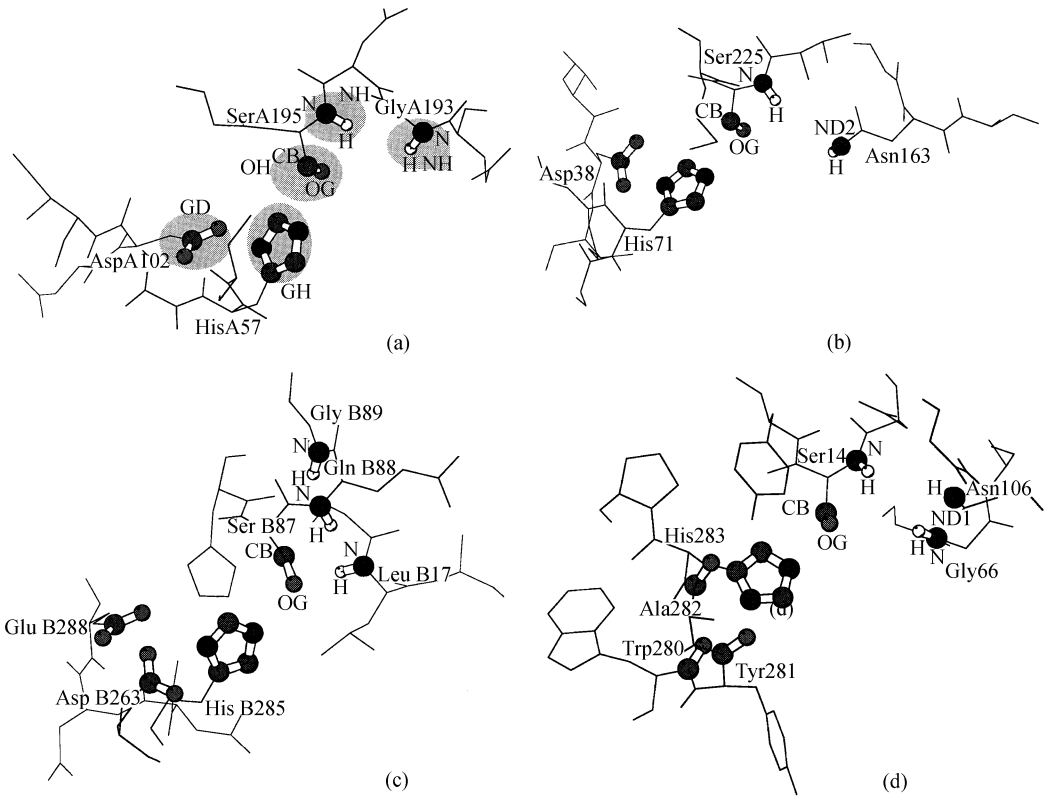

图 3　“催化三联体和氧穴”功能基团三维模体

各图仅显示提供功能基团的残基及其相邻残基的线型模型，其中功能基团用球棍模型突出表示，用 MOLSCRIPT **软件**[17] **制作**. (a) **猪胰蛋白酶** (PDB **代码** 1mct), **用于功能基团模体的构建，其中** 5 **个组成功能基团由填充圆突出标识**. (b) **枯草杆菌蛋白酶**(PDB **代码为** 1thm). (c) α/β **水解酶** (PDB **代码** 1tah), **该蛋白质具有一个替代三联体**. (d) **疥疮病链霉菌酯酶** (PDB **代码** 1ese), **主链羰基在该蛋白质中起酸性基团的作用**

团模体可以在胰蛋白酶家族 (**如** 1ppfE, **概率为** 2.22×10^{-17}, **期望值为** 6.61×10^{-10}), **枯草杆菌蛋白酶家族**(**如** 1thm, **概率为** 3.26×10^{-12}, **期望值为** 4.84×10^{-4}, **见图** 3(b)), **以及**α/β**水解酶**(**如** 1tahB, **概率为** $1.71e\times10^{-12}$, **期望值为** 1.25×10^{-3}, **见图** 3(c))**中搜索到相似的匹配基团**. **一般情况下，在胰蛋白酶和**α/β**水解酶中由两个主链氨基形成氧穴，而在枯草杆菌蛋白酶中由** Asn **的侧链酰氨基中的氨基和一个主链氨基参与形成，这与实验结论是符合的**[15]. **搜索结果还表明提供主链氨基的残基本身可以不具备残基特异性，只需在立体结构上合适即可**. **另外，我们的方法能将** Russell **搜索方法**[7]**得到的** 13 **个** (PDB **代码分别为** 1vnc, 1tlcA, 1ribA, 1fatA, 1mioA, 1amp, 1celA, 1jbc, 1vhh, 2cpl, 1pta, 2rmcA **和** 1cynA)**未在试验上所证实的蛋白质与其他已知酶区分开来，即提高了灵敏度**. **这归因于使用了更加详细的三维模体，包括三联体和氧穴**. **需指出，胰蛋白酶** 1ton **的晶体结构中由于** Zn^{2+}**结合于活性部位导致活性部位结构的变形**[18]**，故在该蛋白质中没有找到匹配**.

对三酰甘油水解酶 (PDB **代码** 1tahB), **这里不仅找到真实的三联体** Asp263-His285-Ser87, **还找到“替代”三联体** Glu288-His285-Ser87 (**见图** 3(c)), **与实验报道相符**[19]. **另外，在辛德比斯病毒** (*sindibis virus*) **衣壳蛋白** (PDB **代码** 1kxf) **中除了三联体** Asp163-His141-Ser215, **还搜**

索到“替代”三联体 Glu263-His141-Ser215，预测该蛋白的 Asp163 发生突变后仍具有酶活性.

对于硫酯酶(PDB 代码 1tht)活性位点的研究存在争议，Ferri[20]等人认为 Ser71 是亲核基团，而 Lawson 等人[21]的突变试验等结果表明，Ser114, His241, Asp211 构成了该蛋白的三联体，还指出 Leu115 的主链氨基参与形成氧穴，但是没找到另外一个氢键提供者. 我们的结果支持了后者的结论，并且认为 Ser116 的主链氨基也参与形成氧穴.

2.3 a/b水解酶的模体搜索结果

α/β水解酶是一类序列上差异很大，但是均具有核心的α螺旋/β折叠的结构，Heikinheimo 等[22]对α/β水解酶超家族蛋白质进行了整理与分类. 我们以 ESTHER 网站整理的α/β水解酶家族蛋白质为对象，去除仅有 CA 坐标的蛋白质 (PDB 代码分别为 1tia, 1tic, 1tgl 和 5tgl)，只有部分结构的模建结构的蛋白质 (PDB 代码为 3ace 和 4 ace)以及突变体(Ser 120Ala, PDB 代码 1cui)，用模体 1 进行搜索. 在 13 个卤烷脱卤素酶 (haloalkane dehalogenase, 相应亲核基团位置的残基为 Asp) 以及 29 个锌依赖外肽酶 (zinc-dependent exopeptidase) 中均没找到匹配，而实际上这两类蛋白质并不具备经典三联体[22]. 其余 115 个α/β 水解酶中大部分 (99 个蛋白质)均找到与胰蛋白酶的三联体和氧穴模体相匹配的局部结构，具体见表 2.

表 2 α/β 水解酶超家族的模体 1 搜索结果

蛋白质家族	找到匹配的蛋白质	总数	未找到匹配的蛋白质	总数
乙酰胆碱酯酶 (acetylcholinesterase)	1acl, 1cfjA, 1eea, 1eve, 1maaA, 1oce, 1somA, 1mah, 1vot, 2ace, 2ack, 2dfpA,	12	1acj, 1amn, 1ax9, 2clj, 1fss	5
细菌脂肪酶 (bacterial lipase)	1cvl, 1oilA, 1tahB, 2lip, 3lip, 4lip, 5lip	7		0
羧肽酶 (carboxypeptidase)	1ac5, 1bcr, 1bcs, 1ivyA, 1whs, 1wht, 3sc2AB	7	1cpy, 1ysc	2
胆固醇酯酶 (cholesterol esterase)	1akn, 1aqlA, 2bce	3		0
角质酶 (cutinase)	1agy, 1cex, 1cua, 1cub, 1cuc, 1cudA, 1cue, 1cuf, 1cug, 1cuh, 1cuj, 1cus, 1cuu, 1cuv, 1cuwA, 1cux, 1cuy, 1cuz, 1ffa, 1ffc, 1ffd, 1ffe, 1xomA, 1xza, 1xzb, 1xzc, 1xze, 1xzf, 1xzg, 1xzh, 1xzi, 1xzj, 1xzkA, 1xzl, 1xzm, 2cut	36		1
双稀内酯水解酶 (dienelactone hydrolase)		0	1din	1
真菌羧肽酶 (fungal carboxylesterase lipase)	1cleA, 1crl, 1lpm, 1lpo, 1lps, 1thg, 1trh	7	1lpn, 1lpp	2
真菌三酰甘油脂酶 (fungal triacylglycerol lipase)	1lbs, 1lbt, 1lgyA, 1tca, 1tcbA, 1tccA, 1tib, 3tgl, 4tgl	9		0
卤素过氧化物酶 (haloperoxidase)	1a7uA, 1a88A, 1a8q, 1a8s, 1a8uA, 1broA, 1brt	7		0
激素敏感脂酶类蛋白 (hormone-sensitive lipase like)	1jkmA	1		0
羟腈裂解酶 (hydroxynitrile lyase)	1yas	1		0
PAF-乙酰水解酶 (PAF-Acetylhydrolase)	1jfr	1		0
胰脂肪酶 (pancreatic lipase)	1bu8, 1ethAB, 1lpbAB,	3	1gpl, 1hplA, 1lpaAB, 1rpl	4
脯氨酸亚氨基肽酶 (proline iminopeptidase)		0	1azm	1
脯氨酸内肽酶 (proline endopeptidase)	1qfmA, 1qfsA	2		0
假单胞杆菌羧肽酶 (pseudomonas carboxylesterase)	1auo, 1aur	2		0
硫酯酶 (thioesterase)	1thtA	1		0

不同的α/β蛋白水解酶的匹配情况有点差异. 角质酶 (cutinase) 家族成员大部分都找到了很好的三联体和氧穴, 概率和期望值也很小, 表明该酶的活性位点与胰蛋白酶的活性位点在立体结构上很相似, 包括一个突变体 (1cuj, Ser120Cys, 仍有功能), 搜索到的三联体 Asp175-His188-Cys120, 以及由 Gln121 主链氨基, Ser42 主链氨基和/或 Asn84 侧链氨基形成的氧穴. 7 个细菌脂肪酶和 7 个卤素过氧化物酶 (haloperoxidase) 家族成员均找到了匹配. 其他家族包括乙酰胆碱酯酶、羧肽酶、胆固醇酯酶、真菌三酰甘油脂酶等都有部分成员能够找到模体 1 的匹配. 对两类蛋白质酸性基团由 Glu 提供的乙酰胆碱酯酶和真菌三酰甘油脂酶, 我们的方法也能够找到正确的匹配. 如乙酰胆碱酯酶(PDB 代码为 1eve), 搜索得到的三联体是 Glu327-His441-Ser200, 以及由 Ala201 和 Gly119 主链氨基形成的氧穴. 但是在脯氨酸亚氨基肽酶 (proline iminopeptidase) 以及双稀内酯水解酶 (dienelactone hydrolase)中我们没有找到功能三维模体的匹配. 我们的搜索结果进一步证实了α/β水解酶具有和胰蛋白活性位点很相似的局部结构, 尽管它们的整体结构很不相似[22].

另外在三酰甘油脂酶 (PDB 代码为 1cvl) 中我们也发现了"替代"三联体现象, 除了三联体 Asp263- His285-Ser87 外, 还找到 Glu288- His285-Ser87.

2.4 变异三联体的结构-功能关系

自然界里存在一些蛋白质具有变异三联体, 这里只对酸性部位为主链羰基, 而非 Asp 或 Glu 羧基情况[14]进行研究(包括 1esc, 1esd, 1ese). 以上述定义的模体 2 在这 3 个蛋白质中进行搜索, 结果见表 3. 从结果看, Ser14, Gly66 主链氨基, 以及 Asn106 的侧链氨基都可参与形成氧穴, 这与 Wei 等人[14]的结论是一致的, 这里参与形成氧穴的氢键供体可以不止两个. 另外, Dodson[13]和 Wei[14]等人认为 Trp280 位的主链羰基起酸性基团的作用, 但我们的搜索结果表明起酸性作用的基团可以不止一个, 包括 Trp280, Tyr281, Ala282 三个残基的主链羰基从立体结构上看都可以较好地与碱性残基 (His) 作用(图 3(d) 示意 1ese 的活性位点), 我们提出的一种解释是, 虽然主链羰基酸性很弱, 但通过量变可以提高酸性.

表 3 疥疮病链霉菌酯酶的模体 2 搜索结果

PDB 代码	距离矩阵概率	基团组合数	期望值	与模体 2 匹配的基团组合				
1esc	5.13×10^{-12}	1.88×10^{9}	9.66×10^{-3}	NH(Ser14)	NQ(Asn106)	GH(His283)	CO(Tyr281)	OH(Ser14)
	1.11×10^{-11}	1.88×10^{9}	2.09×10^{-2}	NH(Ser14)	NQ(Asn106)	GH(His283)	CO(Trp280)	OH(Ser14)
	1.13×10^{-11}	1.88×10^{9}	2.13×10^{-2}	NH(Ser14)	NQ(Asn106)	GH(His283)	CO(Ala282)	OH(Ser14)
1esd	6.66×10^{-12}	1.88×10^{9}	1.26×10^{-2}	NH(Ser14)	NQ(Asn106)	GH(His283)	CO(Tyr281)	OH(Ser14)
	1.08×10^{-11}	1.88×10^{9}	2.03×10^{-2}	NH(Ser14)	NH(Gly66)	GH(His283)	CO(Trp280)	OH(Ser14)
	1.43×10^{-11}	1.88×10^{9}	2.70×10^{-2}	NH(Ser14)	NQ(Asn106)	GH(His283)	CO(Ala282)	OH(Ser14)
	1.45×10^{-11}	1.88×10^{9}	2.73×10^{-2}	NH(Ser14)	NH(Gly66)	GH(His283)	CO(Ala282)	OH(Ser14)
	2.48×10^{-11}	1.88×10^{9}	4.67×10^{-2}	NH(Ser14)	NH(Gly66)	GH(His283)	CO(Trp280)	OH(Ser14)
1ese	5.04×10^{-12}	1.88×10^{9}	9.49×10^{-3}	NH(Ser14)	NH(Gly66)	GH(His283)	CO(Ala282)	OH(Ser14)
	7.86×10^{-12}	1.88×10^{9}	1.48×10^{-2}	NH(Ser14)	NH(Gly66)	GH(His283)	CO(Tyr281)	OH(Ser14)
	7.86×10^{-12}	1.88×10^{9}	1.48×10^{-2}	NH(Ser14)	NQ(Asn106)	GH(His283)	CO(Ala282)	OH(Ser14)
	8.99×10^{-12}	1.88×10^{9}	1.69×10^{-2}	NH(Ser14)	NQ(Asn106)	GH(His283)	CO(Tyr281)	OH(Ser14)
	2.45×10^{-11}	1.88×10^{9}	4.62×10^{-2}	NH(Ser14)	NH(Gly66)	GH(His283)	CO(Trp280)	OH(Ser14)

2.5 功能基团三维模体在功能预测中的应用

利用已知功能的三维模体，在未知功能的蛋白质结构库中进行搜索，可以进行未知功能蛋白质的功能预测. 以非冗余蛋白质库 2 为检测蛋白质库，分别进行了模体 1 和 2 的搜索.

模体 1 的搜索结果得到 77 个蛋白质含有匹配的局部结构，推断这些蛋白质可能具有蛋白质水解酶、酯酶或脂酶等相关活性. 经检索 PDB 文件说明和文献，其中至少有 58 个蛋白质确实具有此类功能. 包括胰蛋白酶的 Ser120Cys 突变体 1dpo，其三联体是 Asp102-His57-Cys195. 另外在三酰甘油水解酶(PDB 代码为 1cvl)发现上述提到的“替代”三联体现象，三联体 Asp263-His285-Ser87 和替代三联体 Glu288-His285-Ser87.

其他一些未经实验验证的匹配详见网页 [1]，这些蛋白质具有与真实的功能基团模体很相似的局部结构，需要综合其他一些判据进一步判别其是否具有功能基团模体所体现的功能. 当然，最终的判据应当是实验验证.

因模体 2 的特异性没有模体 1 的强，因而搜索得到更多的候选者，有 264 个，包含了更多的假阳性. 但是用该模体搜索，可以找到一些变异体，如 1esc(主链羰基执行酸的功能，见表 3)等. 从这点上看，模体 1 比模体 2 优越，具有更高的灵敏度.

3 讨论

本文发展了基于功能基团的三维功能模体的描述方案，以及其搜索与评估方法. 检验的对象是胰蛋白酶、枯草杆菌蛋白酶、α/β水解酶等来源不同的蛋白质家族. 它们在序列上和整体结构上均不存在相似性，但是经过趋同进化都形成类似的三联体和氧穴局域结构. 此外自然界还存在一些三联体的变异体. 通过本方案，在各家族成员中找到了正确的模体，并可识别出多种变异体，如主链羰基起酸作用的 1esc，Glu 起酸作用的 AchE 家族成员，以及 Cys 起亲核作用的 1cuj 等. 所得到的结果优于基于残基的三维模体方案的结果. 同时，我们还利用本实验室发展的进化印记方法[3]提取出三联体的特征序列，对整个 PDB 库作了序列水平的搜索，借以比较序列模体与结构模体的识别效率. 结果表明序列模体对经典的三联体是有效的，但难于像功能基团三维模体那样，辨认序列保守性很低或含变异三联体的目标蛋白质. 可见，三维功能模体是进化印记方法向结构水平的发展，具有实际的应用前景.

由于定义代表性的模体并不容易，蛋白质结构测定也有误差，如何更好地协调敏感性和特异性是一个值得探讨的问题. 如变异三联体中，有碱性部位是 Lys 的这种情况，需要蛋白质其他部分为其提供一定的理化环境，保持其脱质子状态以实施碱的作用[13]. 由于其难以用局部功能基团加以充分描述，本文没有识别这种变异体. 可考虑加入结合界面的特征，与其他活性位点是否部分重叠甚至序列的进化等信息来作进一步的甄别.

随着结构基因组计划的展开，蛋白质结构越来越多，为通过三维模体搜索进行功能预测和有理设计提供了广阔的发展前景. 蛋白质功能基团三维模体的定义与搜索方案为通过活性位点嫁接设计新的蛋白质提供了依据[10~12]，活性位点嫁接方法构建新的蛋白质又可对功能预测进行检验.

致谢 汤海旭博士和盛泉虎对本工作提出了有益的建议，谨此致谢. 本工作为国家“863”生物高技术项

1) http://dna.sibc.ac.cn/ ~ ye/test4table.html

目(批准号：863-103-03-03)、 国家自然科学基金重大项目(批准号：39990600-03)和上海市重大基础研究项目.

参 考 文 献

1 Murzin A G, Patthy L. Sequences and topology from sequence to structure to function. Curr Opin Struc Biol, 1999, 9: 359~362

2 Orengo C A, Todd A E, Thronton J M. From protein structure to function. Curr Opin Struc Biol, 1999, 9: 374~382

3 解 涛, 陈 洁, 丁达夫. 基因组功能预测的进化印记方法. 生物化学与生物物理学报, 1999, 31(4): 433~439

4 Artymiuk P J, Poirrette A R, Grindley H M, et al. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. J Mol Biol, 1994, 243: 327~344

5 陈 洁, 汤海旭, 丁达夫. 用于蛋白质分子设计的三维模体搜索. 生物物理学报, 1997, 13(4): 639~646

6 Wallace A C, Laskowski R A, Thronton J M. Derivation of 3D coordinate templates for searching structural databases: Application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. Protein Science, 1996, 5: 1001~1013

7 Russell R B. Detection of protein three-dimensional side-chain patterns: New examples of convergent evolution. J Mol Biol, 1998, 279: 1211~1227

8 Kleywegt G J. Recognition of spatial motifs in protein structures. J Mol Biol, 1999, 285: 1887~1897

9 Li Zhang, Godzik A, Skolnick J, et al. Functional analysis of the Escherichia coli genome for members of the a/b hydrolase family. Folding & Design, 1998, 3: 535~548

10 Iengar P , Ramakrishnan C. Knowledge-based modeling of the serine protease triad into non-protease. Protein Eng, 1999, 12(8): 649~655

11 Quēmēneur E, Moutiez M. Engineering cyclophilin into a proline-specific endopeptidase. Nature, 1998, 391(6664): 301~304

12 叶玉珍, 汤海旭, 丁达夫. 活性位点转移设计新的功能蛋白质. 生物化学与生物物理学报, 1999, 31(3): 303~308

13 Dodson G, Wlodawer A. Catalytic triads and their relatives. Trends Bioche Sci, 1998, 23(9): 347~352

14 Wei Y, Derewemda Z S. A novel variant of the catalytic triad in the streptomyces scabies esterase. Nat Struc Biol, 1995, 2: 218~222

15 Whiting A K, Peticolas W L. Details of the Acyl-enzyme intermediate and the oxyanion hole in serine protease catalysis. Biochemistry, 1994, 33(2): 552~561

16 Brooks B R, Bruccoleri R E, Olafson B D, et al. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J Comp Chem, 1983, 4: 187~217

17 Kraulis P J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J Appl Crystallogr, 1991, 24: 946~950

18 Fujinaga M, James M N. Rat submaxillary gland serine protease, tonin. Structure solution and refinement at 1.8Åresulution. J Mol Biol, 1997, 195(2): 373~396

19 Noble M E, Cleasby A, Johnson L N, et al. The crystal structure of triacylglycerol lipase from pseudomonas glumae reveals a partially redundant catalytic aspartate. FEBS Lett, 1993, 331(1-2): 123~128

20 Ferri S R, Meighen E A. A lux-specific myristoyl transferase in luminescent bacteria related to eukaryotic serine esterases. J Biol Chem, 1991, 266(20): 12852~12857

21 Lawson D M, Derewenda Z S. Structure of a myristoyl-ACP-specific thioesterase from Vibrio harveyi. Biochemistry, 1994, 33(32): 9382~9388

22 Heikinheimo P, Goldman A, Jeffries C, et al. Of barn owls and bankers: a lush variety of α/β hydrolases. Structure, 1999, 7: R141~R146

(1999-11-22 收稿，2000-04-24 收修改稿)