

蛋白质超家族分子进化研究的一种新方法

胡胜民^{①*} 王 钰^① 贺福初^② 孙之荣^{①†}

(^①清华大学生物科学与技术系生物信息研究所, 北京 100084; ^②军事医学科学院放射医学研究所, 北京 100850.

* 现在地址: 中国科学院福州物质结构研究所, 福州 350002. † 联系人, Email: sunzhr@mail.tsinghua.edu.cn)

摘要 选取丝氨酸蛋白酶超家族作为研究对象, 采用 SSAP 算法从结构比较的层次上进行蛋白质超家族分子进化的研究. 分析了相同或相似物种中酶结构和功能的演化、鼠类胰蛋白酶突变体结构的聚类关系以及不同抑制剂对牛 β -胰蛋白酶天然结构的影响, 探索了结构比较对于蛋白质工程中定点突变改造蛋白质结构与功能以及蛋白质抑制剂设计等方面的指导作用. 基于结构比较的方法是蛋白质分子进化研究的一种新方法.

关键词 丝氨酸蛋白酶超家族 结构比较 SSAP 算法 分子进化

生命是由基本组成物质(蛋白质、核酸等)经过漫长的进化演变形成的复杂系统, 因此生物大分子的进化研究对揭示生命的起源与进化机制有重要意义. 以前的蛋白质进化研究多数采用序列比较分析的手段, 虽然重视了序列变化与其功能的相互关系, 但未能考虑其空间结构基础. 蛋白质结构与功能的统一性、功能对结构的依赖性在自然界是普遍存在的规律, 因此蛋白质的结构与功能的进化是密不可分的. 近年来的研究表明, 空间结构比较与序列比较相比可提供更科学的蛋白质分类信息. 此外, 由于现有序列比较方法对同源性低于 20% 的蛋白质分子缺乏有效的分析手段, 因此使得现有分子系统分析方法在蛋白质超家族以上层次遇到了困难, 蛋白质分子进化研究需要发展新方法^[1,2].

结构比较可用于分析蛋白质结构和功能的演化、分析天然蛋白突变体结构的聚类关系以及不同抑制剂对蛋白质天然构象的影响. 这对于加深对蛋白质结构与功能的认识、通过定点突变改造天然蛋白质的结构与功能、利用蛋白质工程设计蛋白质抑制剂都有重要的指导作用. 结构比较还可用于分析蛋白质活性中心的结构保守性. 对于活性中心保守区域的寻找与定位,

定位, 结构比较也比序列比较具有更大的优越性.

本文选取丝氨酸蛋白酶超家族作为研究对象, 采用 SSAP 算法从空间结构比较的层次上进行蛋白质超家族分子进化的研究, 这是一个分子进化研究的新方法, 是国际上崭新的前沿课题. 研究表明, 该方法和以前的序列比较方法具有一致性、相关性, 同时又具有很多优越性, 可以提供与物种进化更为一致的进化关系, 从而为在分子水平上研究物种进化提供了新手段.

1 材料与方法

(i) 材料的选取. 选取物种跨度大、家族成员多且功能明确的丝氨酸蛋白酶超家族作为研究对象. 该超家族成员属于蛋白水解酶类, 都有一个独特易反应的活性丝氨酸残基. 在结构上丝氨酸蛋白酶超家族为全 β 蛋白, 它的核心结构由 β 桶构成. 这些蛋白质有两个结构域, 每个结构域均含有 6~8 个 β -折叠片, 构成一个 β -折叠桶, 活性部位位于两个 β 桶之间, 丝氨酸是活性部位的必需基团. 丝氨酸蛋白酶具有重要的生物学功能, 该超家族的进化研究对蛋白质功能分化的确定有重要的意义.

在 PDB 蛋白质结构数据库中, 已收录了 240 多个(截至 1998 年 2 月)丝氨酸蛋白酶的结构数据, 数量较多, 为实验研究提供了足够的样本. 根据酶的系统命名法将它们按功能分为 26 个谱系. 由于同一谱系内蛋白质结构相似性非常高, 所以从每个谱系内不同物种中各挑选一个酶作为样本. 挑选的原则如下: 从 PDB 库中尽量挑选天然蛋白的晶体结构; 如果只有在与抑制剂相结合的复合物状态下测定的结构, 那么挑选结晶条件最接近正常生理状态且分辨率最高的晶体结构. 这样共挑选出 33 个丝氨酸蛋白酶样本进行研究.

针对不同的研究目标, 还进一步建立了相应的样本子集. 包括猪和牛的 8 个丝氨酸蛋白酶晶体结构, 鼠类胰蛋白酶的 8 个突变体结构, 牛 β -胰蛋白酶以及在与不同抑制剂结合的复合物状态下测定的结构.

(ii) 方法. 结构比较在 SSAP(structure and sequence alignment program)算法的基础上发展, SSAP 算法是一种新的蛋白质结构比较算法. 该算法执行两级动态寻优过程, 首先是初级水平, 即距离水平的动态寻优, 寻找相比较的两个残基的最佳等价距离; 然后是高级水平, 即残基水平的动态寻优, 寻找相比较的两个序列的最佳等价残基. 由于该方法采用在蛋白质序列比较中广泛应用的动态规划最优化技术, 因此它使蛋白质结构和序列比较技术统一化^[3, 4]. 结构分子进化树构建时采用 Philip 软件包中的 NEIGHBOR 程序. 该方法是基于距离矩阵进行进化树构建的.

2 结果与讨论

2.1 进化整体趋异

研究结果表明, 丝氨酸蛋白酶超家族的纵向进化整体上是趋异进化^[5]. 进一步的工作将研究蛋白质超家族的基因起源机制, 进行横向进化和纵向进化的比较研究.

2.2 相关性分析及结构进化的保守性

对选取的 33 个丝氨酸蛋白酶分别进行结构比较、蛋白质序列比较和基因编码区序列比较, 得到相应的距离矩阵. 用曲线相关的方法对两个蛋白质之间的 3 种距离进行相关性分析. 由计算可知, 序列比较距离与编码区序列比较距离之间的相关系数 $R_1 = 0.92$; 结构比较距离

与序列比较距离之间的相关系数 $R_2 = 0.94$ ；结构比较距离与编码区序列比较距离之间的相关系数 $R_3 = 0.88$ 。相关系数 R 是表示两个变量相关密切程度及其性质的统计量, R 越接近于 1, 相关越密切; 越接近于 0, 越可能无相关。这说明用 3 种独立的方法从 3 种不同的角度, 即结构比较、蛋白质序列比较和编码区序列比较得到的 3 种进化距离之间是彼此相关的, 它们包含着相关的生物大分子进化的信息。如图 1 仅以人的胰蛋白酶作为参考的 3 种距离曲线为例, 3 条曲线的趋势是基本一致的, 与微生物类相比较的距离较大(lhav ~ ltry), 与高等生物相比较的距离较小(2tbs ~ 1fuj), 最低值都出现在与人最接近的物种的胰蛋白酶处(1slu 鼠胰酶, 2ptn 牛胰酶)。由图 1 还可看出, 两个相比较的蛋白质, 序列间的距离远大于结构间的距离。以高等生物的各丝氨酸蛋白酶与人的胰蛋白酶相比, 进化过程中在序列上有 50% ~ 70% 的残基不同, 而在残基构象上只有 10% 左右的不同。这表明在丝氨酸蛋白酶中, 三级结构比一级结构更为保守。

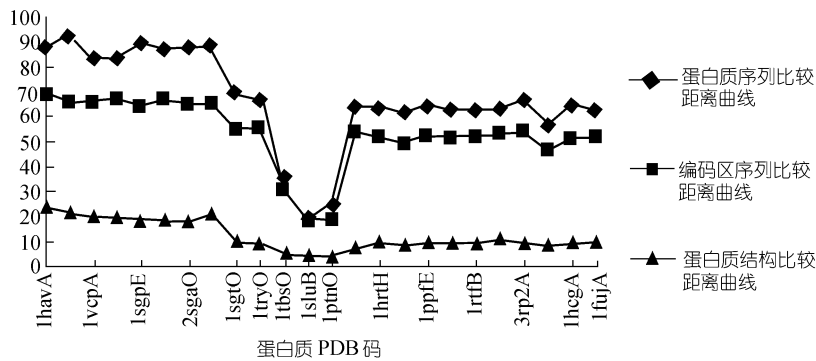


图 1 蛋白质序列比较、编码区序列比较、蛋白质结构比较距离曲线

2.3 相同或相似物种中不同酶结构和功能的差异

选取猪和牛的 8 个丝氨酸蛋白酶构建结构进化树, 如图 2 所示。

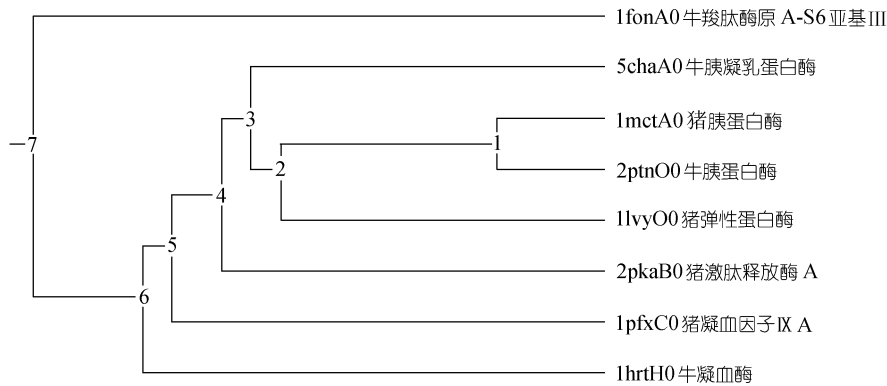


图 2 猪和牛的 8 个丝氨酸蛋白酶的结构进化树

可以看到, 猪和牛的胰蛋白酶、猪弹性蛋白酶、牛胰凝乳蛋白酶首先聚类, 并且结构产生分化的时间是最晚的, 这表明 3 种酶的三级结构具有明显相似性。3 种酶的多肽主链基本上可重叠, 三者结构上的主要区别在于结合底物氨基酸侧链的口袋的一些变化。正是这微小

的结构上的变化使得 3 种酶具有不同的专一性, 而专一性的不同是这 3 种蛋白水解酶功能上的主要差别. 胰凝乳蛋白酶催化由具有大的疏水侧链芳香族氨基酸(苯丙氨酸、酪氨酸以及色氨酸)的羧基所形成的肽键水解, 胰蛋白酶催化由碱性氨基酸(赖氨酸和精氨酸)的羧基所形成的肽键水解, 而弹性蛋白酶催化由具有小的疏水侧链氨基酸(如丙氨酸)的羧基所形成的肽键水解. 与功能上的专一性相适应的结构如图 3 所示.

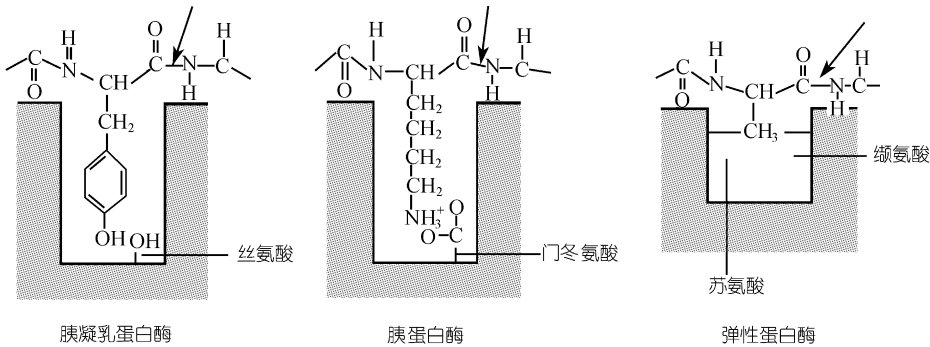


图 3 3 种胰脏蛋白酶的底物结合部位示意图

结构进化树还表明这 3 种酶与激肽释放酶 A, 凝血因子 IXA, 凝血酶和羧肽酶原 A-S6 亚基 III 在结构和功能专一性上都有较大的差异. 激肽释放酶 A 催化从血浆球蛋白生成激肽, 凝血因子 IXA 和凝血酶参与血液凝固的生化过程, 凝血酶的作用专一性是只断裂某些 Arg-Gly 键. 所以通过构建结构进化树, 分析不同酶的聚类关系可用于分析酶结构和功能的差异.

2.4 鼠类胰蛋白酶突变体结构的聚类分析

鼠类的 8 个胰蛋白酶均为突变体, 在结构进化树中聚类成为两组. 在第 1 组中, 1bra, 1brb 与 1brc 突变位点相同而首先聚类, 1trm 与它们突变位点不同, 但突变后结构相似而聚为一组. 1trm 与 1bra 突变位点不同而与相同抑制剂结合, 但并未先聚类. 第 2 组的 4 个突变体定点突变位点相同而与不同抑制剂结合. 这些结果表明主链残基突变引起的结构变化是最主要的, 而抑制剂对结构的影响是次要的(图 4).

突变体结构的聚类分析为利用定点突变的方法改造天然蛋白质的结构与功能提供了一定的指导作用.

2.5 不同抑制剂对牛 β -胰蛋白酶天然结构的影响

图 5 是与不同抑制剂结合的牛 β -胰蛋白酶的结构进化树. 在与胰脏胰蛋白酶抑制剂结合的复合物状态下测定的结构 2ptc 最先分化出来, 然后两个天然体 1tld 和 1btp 与在结合其他抑制剂的复合物状态下测定的结构分别聚类, 在与其他抑制剂结合的复合物状态下测定的结构中, 1may 先与 1bty 聚类, 再与 1tp, 1ntp, 1tyn 依次聚类. 经分析可知, 胰脏胰蛋白酶抑制剂对 β -胰蛋白酶天然构象的改变是最大的, 它是一种天然产生的多肽抑制剂, 与胰蛋白酶结合得非常牢固, 因为它被密封在酶的一个与正常柔性底物呈现结合状态的构象中. 其他抑制剂对天然构象都有不同程度的改变. 不同抑制剂对天然蛋白构象改变的分析对蛋白质工程中蛋白质抑制剂的设计有重要的启示.

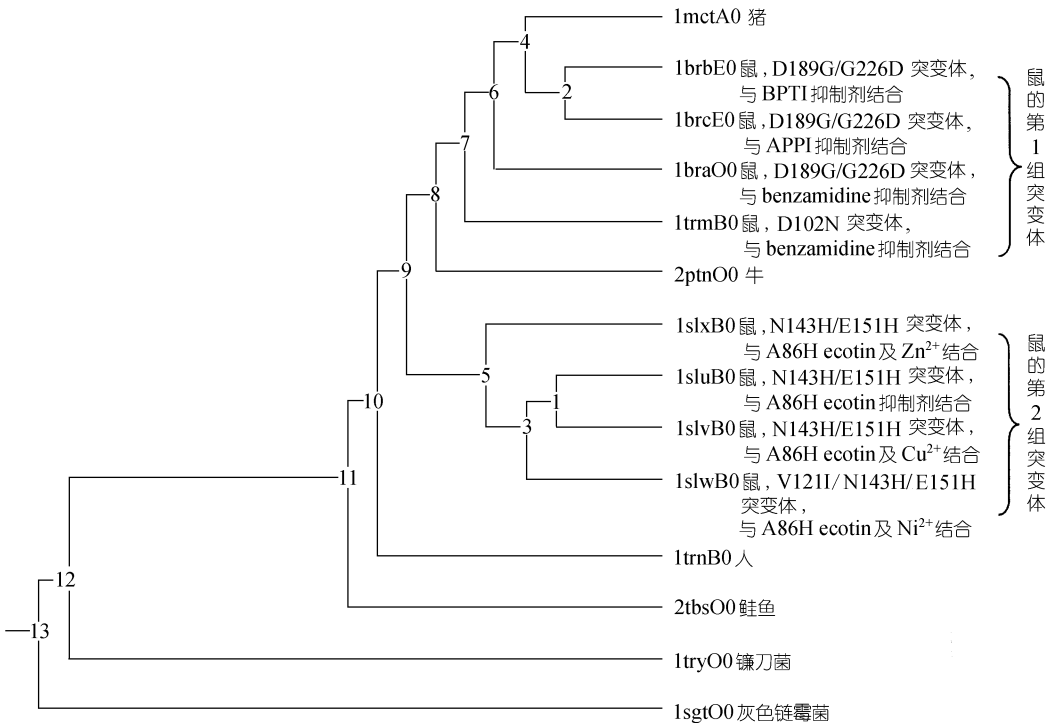


图 4 不同种属间胰蛋白酶(包括鼠的 8 个突变体)的结构进化树

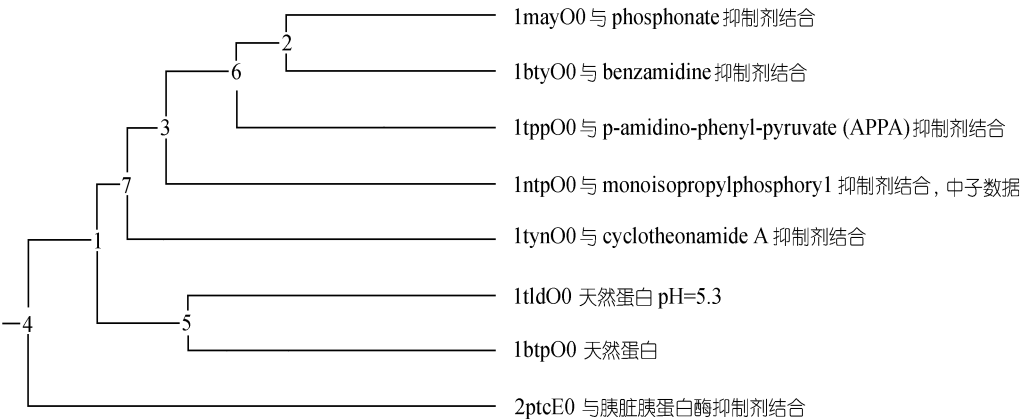


图 5 与不同抑制剂结合的牛β-胰蛋白酶的结构进化树

2.6 丝氨酸蛋白酶活性中心的结构保守性

结构比较还可用于分析蛋白质活性中心的结构保守性的情况. 丝氨酸蛋白酶都有一个独特易反应的丝氨酸残基. 图 6 是 30 个丝氨酸蛋白酶活性中心丝氨酸附近的残基拓扑结构比较图(仅以与人的胰蛋白酶结构比较为例). 可以看到, 活性丝氨酸最邻近的 5 或 6 个残基在空间结构上均为拓扑等价残基, 见图中阴影部分. 通过与进行结构比较的两个蛋白质其他等价残基的相似程度作比较, 发现活性中心的等价残基在拓扑结构上具有更高的相似程度. 这表明

lhavA0	---AWRGKGEGLPCMC GG ALVS-----	SNQSIQNAILGIHVAGGNSILVAKLVTQEMF	病毒
larbO0	---EKRVLGQLHGC PPSS SATG-----	TNRSDQYGRVFTSWTGGGAAASRLSDWLDP	} 菌类
lsgpE0	-----GDSGGPLYSGTRAIG-----	LTSGGSGNCSSGGTTFFQPVTALVAYGVSV	
2sfaO0	-----GDSGGALFAGSTALG-----	LTSGG-SGNCRTGGTTFFQPVTALVAYGVSV	} 高等生物
2sgaO0	-----GDSGGSLFAGSTALG-----	LTSGG-SGNCRTGGTTFFQPVTALVAYGAT	
lhpgA0	-----GDSGGAHFAGSVALG-----	IHSGS-SGCSGTAGSAIHQPVTALVAYGVT	
lfujA0	---GICFGDSGGPLICDGI IQG ---	IDSFVIWGCATRLFPDFFTRVALYVDWIRST	
lppfE0	-----CFGDSGSLVLCNGLIHG-----	IASFVRGGCASGLYPDAFAPVAQFVNWIDSI	
3rp2A0	---FMGDSGGPLLCAGVAHG---	IVS---YGHPPAKPPAIFTRVSTVVPWINAV	
ldsuaO0	---CKGDSGGPLVCGGVLEG---	VVTSGSRVCGNRKKPGIYTRVASYAAWIDSV	
lfonA0	---CNGDSGGPLNC PAADGS WQVHGVTSFVSAFGCNTIKKPTVFTRVS AFID WIDET		
lcltO0	---CNGDSGGPLNCQVNG---SYVHGVTSFVSSSGCNASKKPTVFTRVSAYISWMNGI		
llvyO0	---CQGDSGGPLHLCLVNG---QYAVHGVTSFVSRLLGCNVTRKPTVFTRVSAYISWINNV		
lrtlB0	ANLHDACQGD SGG PLVCLNDG-RMTLVGIISWG---	LGCGQKDPVPGVYTKVNYLDWIRDN	} 高等生物
llmwB0	---DSCQGD SGG PLVCLSQ-RMTLTGIVSWG---	RGCALKDKPGVYTRVSHFLPWIRSH	
5chaA0	---CMGDSGGPLVCKKNG-AWTLVGIVSWG---	SSTCSTSTPGVYARVATLVN V WVQQT	
lsluB0	---DSCQGD SGG PVVCNGELQG---	IVSWGYG-CALPDNPGVYTKVCN YVD WIQDT	
lmctA0	---DSCQGD SGG PVVCNGQLQG---	IVSWGYG-CAQKNKPGVYTKVCN YVN WIQQT	
2ptnO0	---DSCQGD SGG PVVCNGQLQG---	IVSWGSG-CAQKNKPGVYTKVCN YV SWIKQT	
ltnB0	---DSCQGD SGG PVVCNGQLQG---	VVSWG DG -CAQKNKPGVYTKVNYV W KIKNT	
2tbsO0	---DSCQGD SGG PVVCNGELQG---	VVSWGYG-CAEPNPGVYAKVCFISD W LTST	
2pkaB0	---DTCMGDSGGPLICNGMWQG---	ITSWGHTPCGSANKPSIYTKLIFYLD W IDDT	
ltonO0	---DTCAGDSGGPLICDGLVQG---	ITSGGATPCAKPKT PAI YAKLIKFTSWIKKV	
lahtH0	---GDACEGDSGGPFVMKSPFN NRWY QMGIVSWGEG-CDRDGKYGFYTHV FRL KKW I QKV		} 高等生物
lhrtH0	---GDACEGDSGGPFVMKSPY NNRWY QMGIVSWGEG-CDRDGKYGFYTHV FRL KKW I QKV		
lhcgA0	---DACQGD SGG PHVTRFKDTY---FVTGIVSWGEGCARKGKYIYTKV TAFL KWID R SM		
lpfxC0	---DSCQGD SGG PHVTEVEGTS---FLTGIISWGEE-CAVGKGYIYTKVSR YVN WIKEK		
lsgtO0	---VDTCCGD SGG PMFRKDN ADE -WIQVGIVSWGEG-CARPGYPGVYEVSTFAS A IASA		
ltryO0	---DSCQGD SGG PIVDSNTLI---	GAVSWGNG-CARPNYSGVYASV GAL RSFIDTY	
lhy1A0	---PCFGDSGGPFVLSDKNLL-----	IGVVSFVSGAGCESGKPVGFSRVTSYMD W IQ	

图 6 30 个丝氨酸蛋白酶活性中心丝氨酸附近的肽链组成

蛋白质活性中心不仅其一级结构, 而且其三级结构在种系进化上也具有严格的保守性. 实际上, 在酶活性的表现上, 高级结构比一级结构更为重要, 因为只有高级结构才能形成活性中心. 活性中心的形成要求酶蛋白分子具有一定的空间构象. 经分析还发现, 这些活性中心以外相似程度很高的残基大多处于 β 折叠的构象状态下, 这表明 β 桶是该超家族活性中心形成及功能维持的结构基础.

致谢 本工作为国家自然科学基金资助项目(批准号: 39730310, 39980007).

参 考 文 献

- 1 Rufino S D, Blundell T L. Structure-based identification and clustering of protein families and superfamilies. J of Computer-aided Molecular Design, 1994, 8: 5 ~ 27
- 2 Orengo C A, Taylor W R. Protein superfamilies and domain superfolds. Nature, 1994, 372: 631 ~ 656
- 3 Taylor W R, Orengo C A. Protein structure alignment. J Mol Biol, 1989, 208: 1~22
- 4 Sali A, Blundell T L. Definition of general topological equivalence in protein structures. J Mol Biol, 1990, 212: 403~428
- 5 孙之荣, 王 钰, 胡胜民. 丝氨酸蛋白酶超家族分子结构进化研究. 生物物理学报, 1999, 15(3): 530 ~ 535

(2000-03-22 收稿, 2000-06-06 收修改稿)