

doi: 10.13241/j.cnki.pmb.2014.08.001

· 基础研究 ·

动物组织蛋白酶 L-like 基因家族的进化分析 *

张耀洋^{1,2} 王 博^{1,2} 姚蜜蜜^{1,2} 蔡中华^{2Δ}

(1 清华大学生命科学学院 北京 100084;

2 清华大学深圳研究生院 深圳海洋微生物资源筛选与利用公共服务平台 广东 深圳 510000)

摘要 目的:组织蛋白酶 L-like 家族是在溶酶体中发现的一类非常重要的半胱氨酸组织蛋白酶。其主要功能为催化各种蛋白质的水解,并通过水解蛋白质参与到许多的生理调节过程当中。根据序列比对分析和传统的功能分类,在动物中,组织蛋白酶 L-like 家族成员包括组织蛋白酶 L、V、S、K、H 和 F。但是这些家族成员之间的进化关系仍然没有详细研究分析清楚。本课题主要研究组织蛋白酶 L-like 家族成员之间的进化关系。**方法:**本研究通过搜集整理 22 个物种的 177 条组织蛋白酶 L-like 家族蛋白的序列,并构建系统发育进化树来分析组织蛋白酶 L-like 家族各成员之间的进化关系。**结果:**序列数据结果显示,串联重复在组织蛋白酶 L-like 家族的进化过程中发生。斑马鱼的组织蛋白酶 L,爪蟾的组织蛋白酶 S 和 K,大鼠和小鼠的组织蛋白酶 L 都发生过明显的串联重复事件。进化树结果显示了组织蛋白酶 H、S 和 K、L 和 V 之间的进化关系,组织蛋白酶 S 和 K 在脊椎动物出现的进化过程中,从组织蛋白酶 L 中分化出来,与他们在脊椎动物体内的特异性功能,以及脊椎动物在进化过程中分化产生的特异性功能相对应。**结论:**在物种进化的过程中,组织蛋白酶 L-like 家族成员 F、H、S 和 K、L 和 V 按时间顺序分化,这表明组织蛋白酶 L-like 基因家族结构和功能的分化与新的物种和新的功能出现密切相关。

关键词:组织蛋白酶;串联重复;系统进化

中图分类号:Q593, Q78, Q95-3 **文献标识码:**A **文章编号:**1673-6273(2014)08-1401-06

Phylogenetic Analysis of Cathepsin L-like family in Animal*

ZHANG Yao-yang^{1,2}, WANG Bo^{1,2}, YAO Mi-mi^{1,2}, CAI Zhong-hua^{2Δ}

(1 Department of Biological Science and Biotechnology, Tsinghua University, Beijing, 100084, China; 2 Shenzhen Public Service Platforms of Marine Microbial Resource Screening and Exploitation, Graduate School of Tsinghua University, Shenzhen, Guangdong, 510000, China)

ABSTRACT Objective: The cathepsin L family is an important cysteine protease found in lysosome. Its main function is catalyzing proteolysis, and participating in numerous biological processes by proteolysis. The family contains cathepsin L, V, S, K, H and F in vertebrates according to sequence alignment and traditional function classification. The evolutionary relationship of this family members has not been clearly elucidated. Our study focused on the phylogeny of Cathepsin L-like family members. **Methods:** 177 cathepsin L-like gene and protein sequences of 22 species were sourced on database. Phylogenetic tree was constructed to trace the evolutionary relationship of the cathepsin L family genes with these sequences. **Results:** Tandem repeat duplication happened in the evolutionary history of cathepsin L family. Cathepsin L in zebrafish, cathepsin S and K in xenopus, cathepsin L in mouse and rat experienced obvious tandem repeat event. The results of phylogeny trees showed the evolution relationship of cathepsin H, S and K, L and V. Cathepsin S and K differentiated from cathepsin L in the appearance of vertebrates, which coincide with their special function role in the differentiation of vertebrates. **Conclusions:** Cathepsin L-like family members F, H, S and K, L and V diverged chronologically in the evolution of species. This indicates that the structure and function divergence of cathepsin L-like family and the appearance of new species and new function are closely related.

Key words: Cathepsin; Tandem repeat; Phylogeny

Chinese Library Classification(CLC): Q593, Q78, Q95-3 **Document code:** A

Article ID:1673-6273(2014)08-1401-06

前言

组织蛋白酶是一类在溶酶体中被发现的蛋白质水解酶类。

在溶酶体中,蛋白质降解是各种不同蛋白酶(或肽酶)随机组合发生的一种限制性的催化作用过程。为了使溶酶体中的生物大分子能有效降解,溶酶体包含了大量不同的酶类,包括蛋白酶,

* 基金项目:国家重点基础研究发展计划“973”前期项目(2012CB426504);国家海洋局公益性项目(201205020-09;20130502)

作者简介:张耀洋(1984-),男,硕士研究生,主要研究方向:生物信息学

Δ通讯作者:蔡中华(1966-),博士后,主要研究方向:鱼类免疫学、海洋动物病害控制等

(收稿日期:2013-10-11 接受日期:2013-10-30)

淀粉酶,脂肪酶和核酸酶。在约 50 种已知的溶酶体水解酶中,天冬氨酸组织蛋白酶,丝氨酸组织蛋白酶和半胱氨酸蛋白酶显得尤为重要^[1]。而其中半胱氨酸组织蛋白酶在溶酶体蛋白酶中起关键作用。他们属于半胱氨酸组织蛋白酶家族中的 CA 家族,在生物体中广泛分布,并且是研究得最透彻的酶类家族之一。更确切的说,它们是半胱氨酸蛋白水解酶中,最大的也是研究得最多的木瓜蛋白酶类 C1 家族中的成员,包括木瓜蛋白酶以及相关的植物酶类,蠕虫以及寄生虫体内的蛋白酶,昆虫中木瓜蛋白酶的类似物,吸虫半胱氨酸蛋白酶,病毒蛋白酶,和溶酶体半胱氨酸组织蛋白酶^[2]。

在人类的基因组序列中一共发现有 11 种半胱氨酸组织蛋白酶,它们是组织蛋白酶 B,C,F,H,K,L,O,S,V,X 和 W。通过生物信息学分析人类基因组序列草图也得出同样的结论^[3]。溶酶体组织蛋白酶需要一个还原的,微酸的环境(如溶酶体中的环境),来保持最佳的活性状态。因此,半胱氨酸组织蛋白酶最初被认为是一种细胞内消化酶,参与微酸环境下的胞内体或溶酶体中的非特异性的,大量的内源性和外源性的蛋白质水解降解^[4]。然而,近年的研究表明,半胱氨酸组织蛋白酶还参与了大量的特异性的生理活动。

主要的组织蛋白酶在人类组织中呈遍在表达,如组织蛋白酶 B,H,L,C,X,F 和 O。他们的表达特点表明这些酶与细胞内的正常蛋白的降解和转化有关。而一些组织蛋白酶,如组织蛋白酶 K 和 S 表现出限制性的细胞和组织分布,表明它们参与了特殊的生理过程。例如,组织蛋白酶 K 在破骨细胞,大多数的上皮细胞和风湿性关节炎关节处的滑膜成纤维细胞中高度表达。在基质降解酶中,已有明确的证据表明,组织蛋白酶 K 是在鼠和人体的骨吸收作用过程中起关键作用的唯一的酶^[5]。组织蛋白酶 S 在抗原呈递细胞(APCs)中,例如树突状细胞和 B 细胞,呈显著性表达^[6]。另外,组织蛋白酶 V(也被称为 L2)虽然与组织蛋白酶 L 高度同源,但与组织蛋白酶 L 遍在表达不同的是,组织蛋白酶 V 只在胸腺及睾丸中表达^[7]。然而,最近的研究发现,在其他的细胞结构中也有活跃的组织蛋白酶存在,如细胞核,细胞质和质膜中。具有催化活性的组织蛋白酶 L 的变体在细胞核中被定位出来,其对组蛋白 H3 尾巴的 N 末端的酶水解功能,使其在细胞周期的进程中扮演一定的角色^[8]。另外,有研究发现,细胞凋亡前分子 Bcl-2 的同类蛋白 Bid,能被溶酶体的提取物分解,从而引发了细胞色素 c 从线粒体的释放和细胞凋亡^[9]。在体外实验条件下,多种组织蛋白酶能够切割 Bid,并且在无细胞体系下,组织蛋白酶 B,K,L 和 H 都能有效的处理 Bid,得到其凋亡前形式^[10]。这一发现在不同的细胞模型下都被证实。而且,除了 Bid 之外,组织蛋白酶同时还降解抗细胞凋亡因子 Bcl-2 家族蛋白,如 Bcl-2,Bcl-xl,Mcl-1 和 XIAP,从而以协同作用的方式调控细胞凋亡^[11]。除此之外,在许多的生理过程中,组织蛋白酶均扮演重要的角色,但他们的许多底物目前还未鉴别出来。

尽管组织蛋白酶类有类似的蛋白质分解功能,但他们前导肽链的氨基酸序列却区别很大。这与他们在剪切掉前导肽链后的成熟酶形式时,大多数所表现出的大体上的相似性形成鲜明的对比。这可以解释为,在将其酶原传输到内吞溶酶体小隔间及后续的处理激活降解功能的过程中,需要选择性的抑制其

同源酶的催化功能。

木瓜蛋白酶家族详细的序列分析表明组织蛋白酶主要分为两大类,组织蛋白酶 L-like 家族和组织蛋白酶 B-like 家族,通过前导肽和成熟酶的形式均能加以区别^[12]。这两个亚家族之间主要的区别在于未剪切过前导肽的序列和长度^[13]。组织蛋白酶 L-like 家族(组织蛋白酶 L,V,K,S,H)包括一个大约有 100 个氨基酸长度的前导肽,其中有两个保守的共同基序:ERFNIN 基序和 GNFD 基序。而组织蛋白酶 B 家族成员缺少 ERFNIN 基序。

组织蛋白酶 B 和 B-like 家族成员在植物中被发现,证明组织蛋白酶 B 是个非常古老的基因^[14]。而组织蛋白酶 L 及 L-like 家族成员则未在植物中发现,证明组织蛋白酶 L-like 家族的起源要晚于组织蛋白酶 B-like 家族。

在组织蛋白酶 L-like 家族中,又可以分成一些小的亚组,如组织蛋白酶 L 和 V 相近,组织蛋白酶 K 和 S 相近。但组织蛋白酶 L-like 家族成员之间系统进化关系依然没有搞清楚。搞清楚组织蛋白酶 L-like 家族各成员之间的进化关系,对于组织蛋白酶这一类酶的起源发生及系统进化研究,意义重大。而组织蛋白酶的系统进化研究,能够从另一个侧面解释组织蛋白酶在物种中的功能变化发展,与物种对环境的适应,物种的进化之间的内在联系,提供更多的线索与证据。

本研究利用生物信息学软件等方法和网络生物信息数据库,对动物的组织蛋白酶 L-like 家族的系统进化进行了深入的研究,主要内容:组织蛋白酶 L-like 家族基因在染色体上的定位和排布信息、扩增基因及其扩增模式的监测、基因间的系统进化关系及发育树重建、组织蛋白酶 L-like 家族成员间的关系定位。通过以上方法,揭示动物中组织蛋白酶 L-like 家族各成员之间的进化历史及其相互关系。

1 材料与方法

1.1 氨基酸与 cDNA 序列的搜集

通过多个数据库网站搜索下载,获得不同物种组织蛋白酶 L 家族的氨基酸与 cDNA 序列,并整理。首先在 NCBI 网站的蛋白质数据库中获得人类组织蛋白酶 L 的氨基酸与 cDNA 序列^[15]。然后通过 Ensemble 的 BLAST/BLAT 程序(<http://asia-ensembl.org/Multi/blastview>) 和 UCSC 的 BLAT 程序(<http://genome.ucsc.edu/cgi-bin/hgBlat>)进行全基因组序列搜索,从而得到相应的全长 cDNA 与蛋白质氨基酸序列^[16,17]。部分序列通过数据库 JGI 获得。部分序列通过 EMBOSS Transeq 程序进行翻译成氨基酸序列。然后将这些序列在 NCBI 网站上使用 BLAST 程序,在 GenBank 数据库中进行比对验证,确认数据的可靠性。最后,将所有可靠序列统一整理成 FASTA 格式。

1.2 基因组定位与基因结构分析

利用生物信息学网站 Ensembl 的 BLAST/BLAT 程序(<http://www.ensembl.org/Multi/blastview>) 和 UCSC 的 BLAT 程序(<http://genome.ucsc.edu/cgi-bin/hgBlat>),对 22 个已完成全基因组序列测定的物种的组织蛋白酶 L-like 家族蛋白在基因组中的定位进行查看并整理,同时获得相应基因的外显子内含子数量及分布情况。

1.3 系统发育进化树的建立

来自 22 个物种的 177 条组织蛋白酶 L 家族蛋白的氨基酸序列通过 Clustal X 1.83 软件进行多序列比对,并用手动调整至最佳最佳参数设置,同时将处理的比对结果保存为 Aln、Phylip 和 Nexus 的文件格式^[18]。

本研究使用邻接法(neighbor joining, NJ)、最小进化法(minimum evolution, ME)、贝叶斯法(Bayesian)和最大似然法(maximum likelihood, ML)四种方法建立进化树^[19]。邻接法和最小进化法主要通过 MEGA 4.0 程序建树^[20],将 Clustal X 程序多序列比对结果生成的 Aln 格式文件通过 MEGA 4.0 程序转换成 Mega 格式,再分别在 MEGA 4.0 程序下使用这两种方法建树,参数设定主要选择程序默认值。贝叶斯法和最大似然法首先依据 Prottest 程序(http://darwin.uvigo.es/software/prottest_server.html)分析,得到建议建树模型为 WAG+I 替代模型^[21]。贝叶斯法使用 MrBayes3.1.2 程序构建系统进化树,将 Nexus 格式的氨基酸序列文件上传到 BioHPC 服务器(<http://cbsuapps.tc.cornell.edu/mrbayes.aspx>)进行计算^[22]。贝叶斯法通过 MC3(Metropolis-coupled Monte Carlo-Markov Chain method)估算进化枝的后验概率。后验概率通过包含不均等氨基酸频率的不确定信息先验概率进行估算。贝叶斯法的主要参数为:两个平行且各种独立的随机树,每个分析使用 4 条链,包括 3 条热链和 1 条冷链,各跑 106 代,先验概率密度设为 1.0。进化树每 100 代取样一次,并且随后每 2500 代老化一次。最后生成一个 95%主要规则一致的树来计算后验概率的值。ML 法使用 PhyML 程序(<http://www.atgc-montpellier.fr/phyml/>)^[23]计算获得,将 Phylip 格式文件上传至服务器(<http://www.atgc-montpellier.fr/phyml/>),替代模型仍选择 WAG+I 模型。

上述程序建立的系统进化树,均在 MEGA 4.0 程序中显示。

2 结果

2.1 组织蛋白酶 L-like 家族的序列数据

通过搜索和整理,得到 22 个物种的 177 条组织蛋白酶 L-like 家族的蛋白质序列及 cDNA 序列,这 22 个物种包括:人(Human, *Homo sapiens*)、黑猩猩(*Chimpanzee*, *Pan troglodytes*)、猕猴(*Macaque*, *Macaca mulatta*)、小鼠(Mouse, *Mus musculus*)、大鼠(Rat, *Rattus norvegicus*)、豚鼠(*Guinea pig*, *Cavia porcellus*)、家兔(Rabbit, *Oryctolagus cuniculus*)、猪(Pig, *Sus scrofa*)、牛(Cow, *Bos taurus*)、狗(Dog, *Canis familiaris*)、负鼠(*Opossum*, *Monodelphis domestica*)、鸭嘴兽(*Platypus*, *Ornithorhynchus anatinus*)、鸡(Chicken, *Gallus gallus*)、蜥蜴(Lizard, *Anolis carolinensis*)、爪蟾(Frog, *Xenopus tropicalis*)、河豚(Fugu, *Takifugu rubripes*)、青鳉(Medaka, *Oryzias latipes*)、刺鱼(Stickleback, *Gasterosteus aculeatus*)、斑马鱼(Zebrafish, *Danio rerio*)、七鳃鳗(Lamprey, *Petromyzon marinus*)、文昌鱼(Lancelet, *Branchiostoma floridae*)、海鞘(Ciona, *Ciona intestinalis*)。这些物种的序列,涵盖了从最低等的脊索动物,到最高等的脊椎动物主要纲次的代表物种,对这些序列进行进化发育关系重建,可以大致反应在脊椎动物中,组织蛋白酶 L-like 家族的成员在发育过程中的进化关系。另外还有 2 条序列作为外群:人的组织蛋白酶 F 的氨基酸序列、斑马鱼的组织蛋白酶 F 的氨基酸序列。详见表 1。

表 1 22 个物种的 177 条组织蛋白酶 L-like 家族基因及蛋白序列数
Table 1 177 cathepsin L-like family gene and protein sequences of 22

species		
物种类别	名称	序列数
Classification	Species	Sequences
哺乳类 Mammalia	人类 Human(Hsa)	5
	黑猩猩 Chimpanzee (Ptr)	5
	猕猴 Macaque(Mmul)	6
	小鼠 Mouse(Mmu)	15
	大鼠 Rat(Rno)	15
	豚鼠 Guinea pig(Cpo)	4
	家兔 Rabbit(Ocu)	6
	猪 Pig(Sus)	7
	牛 Cow(Bta)	5
	狗 Dog(Cfa)	6
	负 Opossum(Mdo)	8
	鸭嘴兽 Platypus(Oan)	6
	鸡 Chicken(Gga)	6
	蜥蜴 Lizard(Aca)	6
爬行类 Reptilia	蛇 蛇(Liza)	6
	爪蟾 Frog(Xtr)	16
两栖类 Amphibia	河豚 Fugu(Tru)	6
	青鳉 Medaka(Ola)	6
鱼类	刺鱼 Stickleback(Gac)	8
	斑马鱼 Zebrafish(Dre)	21
Actinopterygii	七鳃鳗 Lamprey(Pma)	6
无颌类 Agnatha	文昌鱼 Lancelet(Bfl)	9
头索动物 Cephalochordata	海鞘 Ciona(Cin)	5
尾索动物 Urochordata		
总计 Total	22	177

2.2 基因的定位与分布特点

基因的染色体定位(表 2)发现,大多数物种的组织蛋白酶 L-like 家族蛋白主要包含 5 条,即组织蛋白酶 L、V、K、S、H 各一条,同时组织蛋白酶 L 和 V,组织蛋白酶 K 和 S,组织蛋白酶 H,分别定位在三条不同的染色体上,如人的组织蛋白酶 L 和 V 位于 1 号染色体上,而人的组织蛋白酶 K 和 S 位于 9 号染色体上,人的组织蛋白酶 H 则位于 15 号染色体上,这与进化树显示的分化关系基本一致。组织蛋白酶 H 通常为 12 个外显子,而组织蛋白酶 L、V、K、S 通常为 8 个外显子,这可以成为组织蛋白酶 H 与组织蛋白酶 L、V、K、S 的分化可能发生在较早的时候又一证据。在收集的绝大部分物种中,组织蛋白酶 L 和 V 通常分别位于同一条染色体上,位置相近的两条正负链上,而组织蛋白酶 K 和 S 通常位于同一条染色体上,位置相近的同一条链上,显示了他们在基因的扩增分化上,在进化发生的过程中的不同方式。

2.3 部分物种组织蛋白酶 L-like 家族的串联重复扩张

由表 1 可以看出在搜集的物种中,有四个物种的组织蛋白酶 L-like 家族基因发生了非常明显的扩增,他们是:小鼠 15 条、大鼠 15 条、爪蟾 16 条、斑马鱼 21 条。根据染色体定位,可以看出,小鼠位于 13 号染色体上的组织蛋白酶 L 发生了明显的基因扩增,大鼠位于 17 号染色体上的组织蛋白酶 L 发生了明显的基因扩增,爪蟾的组织蛋白酶 S 发生了明显的基因扩

增,以及斑马鱼位于12号染色体上的组织蛋白酶L发生了明显的基因扩增。这些扩张的基因,基本都呈线性排列在同一条染色体上,提示了组织蛋白酶基因在这些物种上的扩增方式主要是串联重复。

2.4 系统进化树

通过邻接法(NJ)、最小进化法(ME)、最大似然法(ML)和贝叶斯法(Bayes)构建出的进化树,拓扑结构相似,进化主线大致相同。从图1可以看出,组织蛋白酶L-like基因家族的进化主要分成三大部分,首先是组织蛋白酶H从组织蛋白酶L祖先中分化出来,图1中用青色表示。然后从鱼类开始,组织蛋白酶

K和S从组织蛋白酶L家族中分化出来,图1中用蓝色表示。同时,组织蛋白酶L也在哺乳动物中发生了进一步的分化,图1中用绿色表示。

邻接法和最小进化法的后验概率较低,普遍在60-80,有些甚至低于60,而贝叶斯法和最大似然法的后验概率较高,普遍在90以上,其中又以贝叶斯法所建树的所有分支都具有最高的后验概率支持。与贝叶斯建树相比,其他三种方法的系统进化树存在个别分支自展支持很低的情况,因此,建树结果以贝叶斯法最为可靠,后续研究都在贝叶斯建树结果上展开。

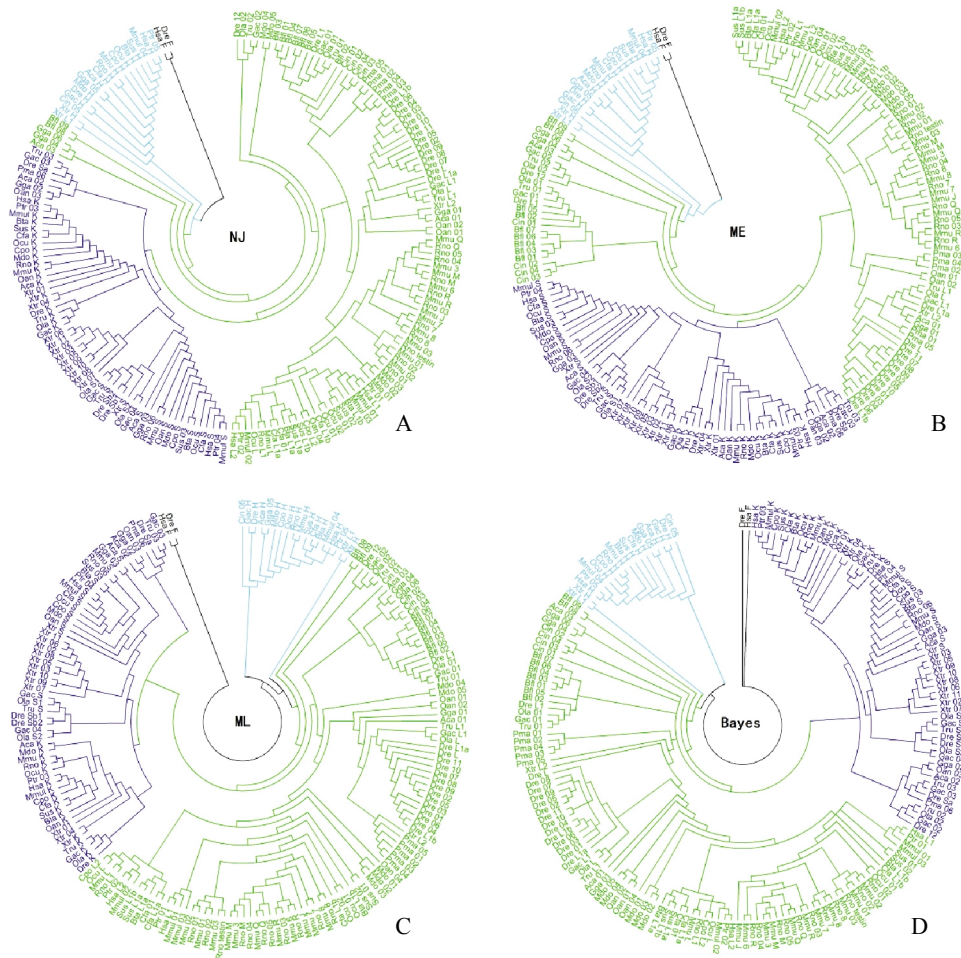


图1 四种方法对177条动物组织蛋白酶L-like家族序列建立进化树

Fig.1 Phylogeny trees built with four methods for 177 sequences of cathepsin L-like family members

A 邻接法进化图(NJ); B 最小进化法(ME); C 最大似然法(ML); D 4 贝叶斯法(BYS)

在图2贝叶斯法建立的系统进化树中,组织蛋白酶H,组织蛋白酶L(或L1)和V(或L2),组织蛋白酶K和S,分别聚成三大簇,显示他们在进化上的亲缘关系,即组织蛋白酶H在较早之前,可能在脊索动物出现以前,已经从组织蛋白酶L-like家族中分化出来。而组织蛋白酶L、V、K、S则在脊索动物出现后才分化,并在组织蛋白酶H分化以后,从组织蛋白酶L的主枝中分化出组织蛋白酶K和S,然后各自成簇,而组织蛋白酶L则分化出L和V。

2.5 组织蛋白酶K和S

由系统进化树可以看出组织蛋白酶K和S在真正的脊椎动物出现以后才分化出来。通过染色体定位发现,所有物种的

组织蛋白酶K和S都出现在同一条染色体上的相近位置,连锁遗传。并且几乎所有物种的组织蛋白酶K和S的基因都在同一条DNA链上(表2),表明组织蛋白酶K和S可能起源于同一基因,并由这个基因复制一次后产生两个拷贝,分别进化成为组织蛋白酶K和S。在玻璃海鞘以及更低等的生物中,没有组织蛋白酶K和S以及类似物被发现出来。在最低等的脊椎动物无颌类七鳃鳗中发现一条与组织蛋白酶K和S的类似物Pma06,在图2中,Pma06在系统发育进化树中与组织蛋白酶K和S的进化枝聚到一簇,因此推断,组织蛋白酶K和S可能起源于脊椎动物的分化。

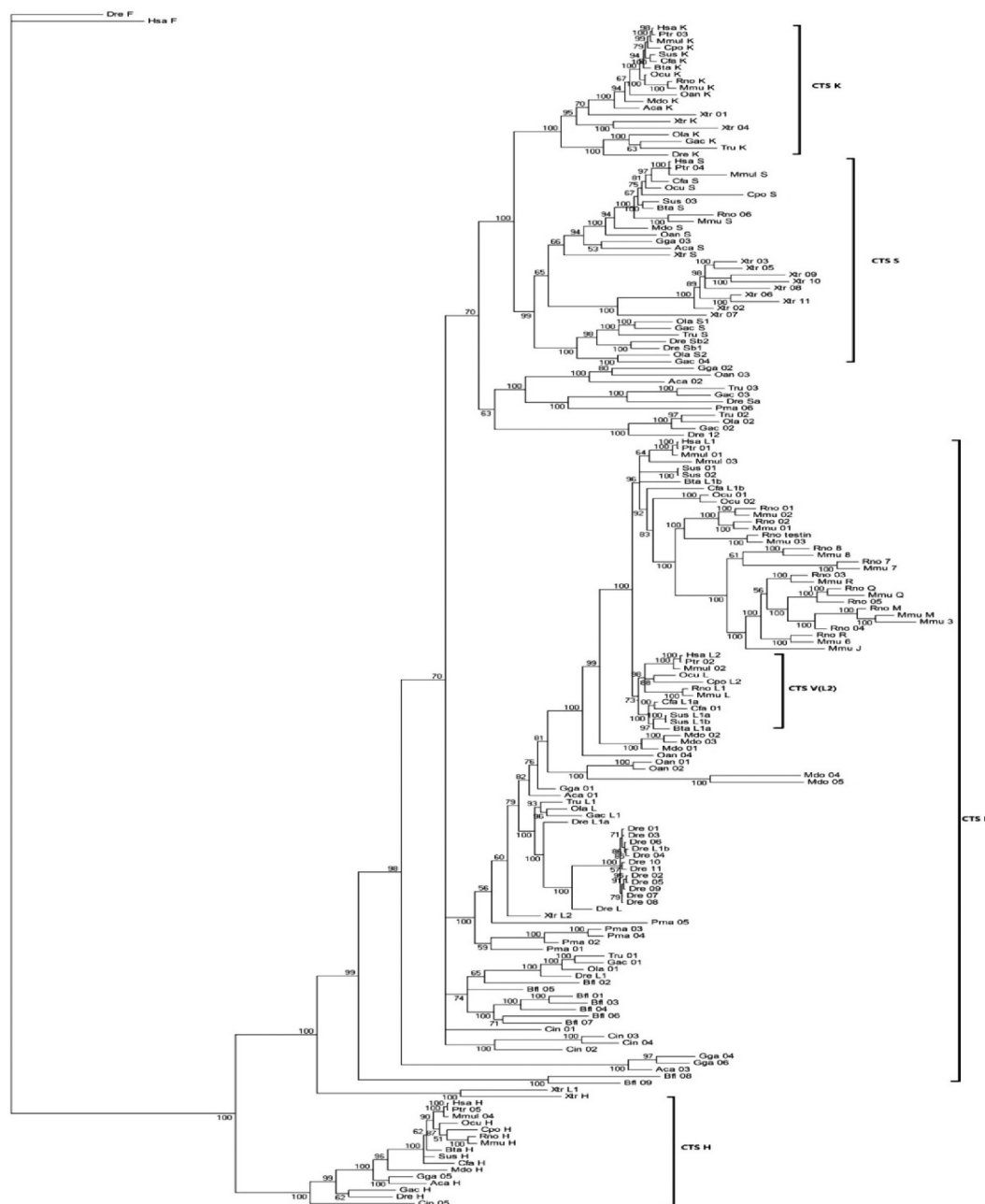


图2 动物组织蛋白酶 L-like 家族进化树:使用 MrBayers 对 177 条组织蛋白酶 L-like 家族成员和两条外群蛋白序列进行构建进化树,节点数字表示贝叶斯法计算的后验概率。

Fig.2 Phylogeny tree of cathepsin L-like family: MrBayes was used to build phylogeny tree of 177 cathepsin L-like family members and two sequences as outgroup, numbers at nodes are posterior probabilities from MrBayes.

3 讨论

通过对酿酒酵母 (*Saccharomyces cerevisiae*) 全基因组 BLAST, 没有找到 cathepsin L 的同源序列, 这表明 cathepsin L 家族成员是在真核生物出现之后才出现的。丝氨酸蛋白酶家族成员之间的序列比对和系统分析结果表明, cathepsin L 家族从 cathepsin B 家族分歧出来的时间甚至早于植物和真菌 (*Dictyostelium discoideum* 1 家族) 分化的时间^[24]。这些证据表明, 木瓜蛋白酶超家族的丝氨酸蛋白酶可以上溯到出现在真核生物进化史, 可能会早于真核生物与原核生物分化的时间。cathepsin L 类家族成员可分为 cathepsin L、V、S、K、H 和 F。虽然

cathepsin F 前肽比 cathepsin L 家族其他成员都长, 但其成熟蛋白酶体却很相似。将线虫 (*C. elegans*) 和果蝇 (*Drosophila melanogaster*) 全基因组与人 (*Homo sapiens*) 和斑马鱼 (*Danio rerio*) cathepsin L 和 F 进行 BLAST, 结果说明 cathepsin L 和 F 在这些无脊椎有不同的对应基因。在线虫中 cathepsin L 对应基因是位于 5 号染色体的 T03E6.7.2 (蛋白 ID, ENSEMBL), cathepsin F 对应基因是位于 5 号染色体的 F41E6.6.1 (蛋白 ID, ENSEMBL)。有趣的是, 当用 cathepsin L、V、S、K 和 H 与线虫基因组进行 BLAST 时, 得到的对应基因总是 T03E6.7.2。在果蝇中得到的结果是类似的。以上结果表明 cathepsin F 从 cathepsin L 分歧出来的时间早于线虫, 而 cathepsin H、V、S 和 K 分歧出来

的时间晚于此。

对 cathepsin L 家族成员进行系统分析,结果表明 cathepsin H、S、K 和 V 从 cathepsin L 分歧出来的时间是按照时间先后的。数据中获得的 cathepsin H 有 12 个外显子,而 cathepsin S、K、L 和 V 只有大约 8 个。海鞘中(*C. intestinalis*)有类似 cathepsin H 的基因 Cin05, 这说明 cathepsin H 至少是开始于脊索动物中从 cathepsin L 中分歧出来,甚至更早。Cathepsin S 和 K 可能是随着脊椎动物的出现而从 cathepsin L 分歧出来的。通过对尾索动物、头索动物和更低等动物(*Strongylocentrotus purpuratus* 和 *Pinctada fucata* 等)测序,未发现 cathepsin S 和 K 相应的基因和蛋白。这可能说明 cathepsin S 和 K 是在脊椎动物出现后才出现的。Cathepsin S 和 K 染色体定位的结果表明它们起源于 cathepsin L 类家族基因的串联重复或基因复制事件。从进

化树的结果中, 我们可以看出 cathepsin S 和 K 是在脊椎动物出现之后从 cathepsin L 中分歧出来的。详细地说, 一个七鳃鳗 (*Petromyzon marinus*) 的基因 Pma06 与 cathepsin S 和 K 基因枝, 这说明 cathepsin S 和 K 可能是起源于脊椎动物的祖先 -- 无颌类脊椎动物。Cathepsin K 在破骨细胞中高表达^[2],是唯一的一个在骨吸收过程中扮演重要作用的蛋白酶,这在人和小鼠中已经得到证实^[3]。Cathepsin S 主要在抗原呈递细胞(APCs)中表达,比如 B 细胞核树突细胞(DCs),这些细胞参与到获得性免疫过程中,比如 MHC-II 介导的抗原呈递作用^[4]。内骨骼和免疫系统的形成来源于脊椎动物的祖先。对 cathepsin S 和 K 进行功能分析, 我们可以看到 cathepsin S 和 K 的出现与脊柱动物形体结构的形成是相互关联的。这是另外一个正面 cathepsin S 和 K 的起源的证据。

表 2 部分物种的组织蛋白酶 K 和 S 的染色体定位

Table 2 Chromosome localization of Cathepsin S and K from partial species

序列名称 Sequences	所属物种 Species	蛋白序列长度 Length of protein sequences	cdna 起始位置 Initial position	正负链 Chains	exon
Hsa_K	Human	329	Chromosome 1: 150,768,684-150,780,799	-	8
Hsa_S	Human	331	Chromosome 1: 150,702,672-150,738,433	-	8
Ptr_03	Chimpanzee	329	Chromosome 1: 128,975,594-128,987,818	-	8
Ptr_04	Chimpanzee	331	Chromosome 1: 128,913,629-128,947,089	-	8
Sus_K	Pig	330	Chromosome 4: 102,342,678-102,354,617	+	7
Sus_03	Pig	342	Chromosome 4: 102,368,163-102,393,283	+	8
Ocu_K	Rabbit	329	Chromosome 13: 41,346,878-41,360,274	+	8
Ocu_S	Rabbit	331	Chromosome 13: 41,370,897-41,392,664	+	7
Aca_S	Lizard	342	Scaffold GL343862.1: 18,139-36,017	+	7
Aca_K	Lizard	290	Scaffold GL343862.1: 1,635-8,701	+	6
Dre_K	Zebrafish	333	Chromosome 16: 31,753,228-31,762,840	+	8
Dre_Sb1	Zebrafish	330	Chromosome 16: 31,769,424-31,773,725	+	8
Ola_K	Medaka	331	Chromosome 16: 24,479,790-24,492,011	+	8
Ola_S1	Medaka	339	Chromosome 16: 24,494,714-24,503,215	+	8

组织蛋白酶 S 和 K 分化出来以后,在高等的哺乳动物中,组织蛋白酶 L 又分化为 L1(L)和 L2(V),并且其相应功能也发生了特异化。在搜集的组织蛋白酶的序列的定位和分布中,组织蛋白酶 S 和 K 基因通常在同一条染色体的同一位置,并且在同一条 DNA 链上,而组织蛋白酶 L 和 V 基因通常在同一条染色体同一位置的不同 DNA 链上。因此,可以推断组织蛋白酶 S 和 K 由同一个祖先基因,同时形成两个拷贝分化而来,而组织蛋白酶 V 则有可能是从组织蛋白酶 L 通过转座或串联重复的方式分化而来。串联重复,作为一种有效的基因扩增方式,在许多动物的组织蛋白酶基因中被发现,如斑马鱼的组织蛋白酶 L,爪蟾的组织蛋白酶 S 和 K,大鼠和小鼠的组织蛋白酶 L,均有明显的串联重复现象。这为组织蛋白酶的功能的分化与进化打下了基础,这些扩增的基因,其结构与功能的研究有待后续开展。

本研究发现,从系统进化树及相应的功能研究中,综合可以推出组织蛋白酶 L-like 家族大致的进化历史,真核生物出现以后,组织蛋白酶 L-like 家族出现,在低等动物中分出组织蛋白酶 F,线虫以后分出组织蛋白酶 H,在脊椎动物出现以后分化出组织蛋白酶 S、K 和 L、V。这些基因的结构和功能分化,在物种的进化过程中起到了重要的作用,对于物种进化及结构功

能的适应性的研究,意义重大。同时,本研究发现,在组织蛋白酶 L-like 家族进化的过程中,发生了许多基因扩增现象,这对于新基因产生及分化的研究,有一定的参考作用。这些通过不同方式扩增分化出来的新基因,可以进行进一步的详细的序列特异性位点分析,功能性分歧分析,以及验证其是否经历自然选择,而使其蛋白质功能得以在物种中积累的选择分析,为此家族蛋白的功能分化提供更多的分子证据。

参 考 文 献(References)

[1] Brix K. Lysosomal Proteases: Revival of the Sleeping Beauty [J]. *Eurekah Bioscience*, 2005, 1(3): 259-264

[2] Barrett A.J, N.D. Rawlings, J.F. Woessner. Handbook of proteolytic enzymes[M]. Vol. 2. 1998: Academic Press New York

[3] Rossi A, Deveraux Q, Turk B, et al. Comprehensive search for cysteine cathepsins in the human genome[J]. *Biological Chemistry*, 2004, 385 (5): 363-372

[4] Turk V. B. Turk, D. Turk. Lysosomal cysteine proteases: facts and opportunities[J]. *The EMBO Journal*, 2001, 20(17): 4629-4633

[5] Asagiri M. H. Takayanagi. The molecular understanding of osteoclast differentiation[J]. *Bone*, 2007, 40(2): 251-264

[6] Hsing L.C, A.Y. Rudensky. The lysosomal cysteine proteases in MHC class II antigen presentation [J]. *Immunological Reviews*, 2005, 207 (1): 229-241

(下转第 1427 页)

- [3] Singh N, Agrawal M, Doré S. Neuroprotective properties and mechanisms of resveratrol in in vitro and in vivo experimental cerebral stroke models[J]. ACS Chem Neurosci, 2013, 4(8): 1151-1162
- [4] Tomé -Carneiro J, González M, Larrosa M, et al. Resveratrol in primary and secondary prevention of cardiovascular disease: a dietary and clinical perspective[J]. Ann N Y Acad Sci, 2013, 1290: 37-51
- [5] Shen M, Wu Rx, Zhao L, et al. Resveratrol attenuates ischemia/reperfusion injury in neonatal cardiomyocytes and its underlying mechanism[J]. PLoS One, 2012, 7(12): e51223
- [6] Cherubini A, Polldori MC, Bregnocebi M, et al. Antioxidant profile and early outcome in stroke patients[J]. Stroke, 2000, 31(10): 2295-2300
- [7] Love S. Oxidative stress in brain ischemia [J]. Brain pathol, 1999, 9(1): 119-131
- [8] Walle T, Hsieh F, DeLegge Mh, et al. High absorption but very low bioavailability of oral resveratrol in humans [J]. Drug Metab Dispos, 2004, 32(12): 1377-1382
- [9] Gruber J, Tang SY, Halliwell B. Evidence for a trade-off between survival and fitness caused by resveratrol treatment of *Caenorhabditis elegans*[J]. Ann N Y Acad Sci, 2007, 1100: 530-542
- [10] Baur JA, Pearson KJ, Price NL, et al. Resveratrol improves health and survival of mice on a high-calorie diet [J]. Nature, 2006, 444(7117): 337-342
- [11] Clark D, Tuor UI, Thompson R, et al. Protection against Recurrent Stroke with Resveratrol: Endothelial Protection [J]. PLoS One, 2012, 7(10): e47792
- [12] Wang LM, Wang YJ, Cui M, et al. A dietary polyphenol resveratrol acts to provide neuroprotection in recurrent stroke models by regulating AMPK and SIRT1 signaling, thereby reducing energy requirements during ischemia[J]. Eur J Neurosci, 2013, 37(10):1669-1681
- [13] Zhu JR, Tao YF, Lou S, et al. Protective effects of ginsenoside Rb(3) on oxygen and glucose deprivation-induced ischemic injury in PC12 cells[J]. Acta Pharmacol Sin, 2010, 31(3): 273-280
- [14] Zhou XQ, Zeng XN, Kong H, et al. Neuroprotective effects of berberine on stroke models in vitro and in vivo [J]. Neurosci Lett, 2008, 447(1): 31-36
- [15] Sharma A, Gautam V, Costantini S, et al. Interactomic and pharmacological insights on human sirt-1[J]. Front Pharmacol, 2012, 3: 40
- [16] Raval AP, Dave KR, Pé rez-Pinzó n MA. Resveratrol mimics ischemic preconditioning in the brain [J]. J Cereb Blood Flow Metab, 2006, 26(9): 1141-1147
- [17] Raval AP, Lin HW, Dave KR, et al. Resveratrol and ischemic preconditioning in the brain [J]. Curr Med Chem, 2008, 15(15):1545-1551

(上接第 1406 页)

- [7] Santamaría I, Velasco G, Cazorla M, et al. Cathepsin L2, a novel human cysteine proteinase produced by breast and colorectal carcinomas[J]. Cancer Research, 1998, 58(8): 1624-1630
- [8] Santos-Rosa H, Kirmizis I A, Nelson C, et al. Histone H3 tail clipping regulates gene expression[J]. Nature Structural & Molecular Biology, 2008, 16(1): 17-22
- [9] Stoka V, Turk B, Schendel Sharon L, et al. Lysosomal protease pathways to apoptosis cleavage of Bid, not pro-caspases, is the most likely route[J]. Journal of Biological Chemistry, 2001, 276(5): 3149-3157
- [10] Cirman, T, Oresic K, Mazovec G. D, et al. Selective disruption of lysosomes in HeLa cells triggers apoptosis mediated by cleavage of Bid by multiple papain-like lysosomal cathepsins [J]. Journal of Biological Chemistry, 2004, 279(5): 3578-3587
- [11] Droga-Mazovec G, Bojic L, Petelin A, et al. Cysteine cathepsins trigger caspase-dependent cell death through cleavage of bid and antiapoptotic Bcl-2 homologues [J]. Journal of Biological Chemistry, 2008, 283(27): 19140-19150
- [12] Cygler M, J. Mort. Proregion structure of members of the papain superfamily. Mode of inhibition of enzymatic activity [J]. Biochimie, 1997, 79(11): 645-652
- [13] Turk B., D. Turk, V. Turk. Lysosomal cysteine proteases: more than scavengers [J]. Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology, 2000, 1477(1): 98-111
- [14] van der Hoorn R.A, Leeuwenburgh M. A, Bogoy M, et al. Activity profiling of papain-like cysteine proteases in plants [J]. Plant Physiology, 2004, 135(3): 1170-1178
- [15] Altschul S.F, Gish W, Miller W, et al. Basic local alignment search tool[J]. Journal of Molecular Biology, 1990, 215(3): 403-410
- [16] Birney E, Andrews T. D, Bevan P, et al. An overview of Ensembl[J]. Genome Research, 2004, 14(5): 925-928
- [17] Kent W.J, Sugnet C. W, Furey T. S, et al. The human genome browser at UCSC[J]. Genome Research, 2002, 12(6): 996-1006
- [18] Thompson J.D, Gibson T. J, Plewniak F, et al. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools [J]. Nucleic Acids Research, 1997, 25(24): 4876-4882
- [19] Sandra L. Phylogeny for the faint of heart: a tutorial [J]. Trends in Genetics, 2003, 19: 345-351
- [20] Tamura K, Dudley J, Nei M, et al. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0[J]. Molecular Biology and Evolution, 2007, 24(8): 1596-1599
- [21] Abascal F, R. Zardoya, D. Posada. ProtTest: selection of best-fit models of protein evolution[J]. Bioinformatics, 2005, 21(9): 2104-2105
- [22] Ronquist F, J.P. Huelsenbeck, Mr Bayes 3: Bayesian phylogenetic inference under mixed models[J]. Bioinformatics, 2003, 19(12): 1572-1574
- [23] Guindon S, O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood [J]. Systematic Biology, 2003, 52(5): 696-704
- [24] Berti P.J, A.C. Storer. Alignment/phylogeny of the papain superfamily of cysteine proteases [J]. Journal of Molecular Biology, 1995, 246(2): 273-283
- [25] Salminen-Mankonen H., J. Morko, E. Vuorio. Role of cathepsin K in normal joints and in the development of arthritis [J]. Current Drug Targets, 2007, 8(2): 315-323