

DIVERSIFICACIÓN EVOLUTIVA DE LOS GENES DE CATEPSINAS EN VERTEBRADOS

por CARMEN GÓMEZ-VERGARA, GUILLERMO THODE

ÁREA DE GENÉTICA, UNIVERSIDAD DE MÁLAGA, 29071 - MÁLAGA (ESPAÑA)

CARMENGOMVER@GMAIL.COM

Palabras clave: genes parálogos, evolución molecular, duplicación, filogenia de catepsinas

Enviado: 19/03/2019

Keywords: paralog genes, molecular evolution, duplication, phylogeny of cathepsins

Aceptado: 23/03/2019

Los genes parálogos son un claro ejemplo de duplicaciones y fuente de diversificación en la evolución de los genomas de los seres vivos. Como muestra de ello, se ha llevado a cabo un estudio de los genes *CTS*, que dan lugar a las enzimas proteolíticas llamadas catepsinas. Tras la búsqueda en bases de datos bibliográficas y de secuencias moleculares, se ha intentado clarificar la confusa información sobre estos genes y analizar las relaciones filogenéticas entre las secuencias aminoacídicas de estas proteínas con el fin de identificar los procesos que han podido intervenir en su diversificación evolutiva desde el origen de los vertebrados terrestres hasta la actualidad. Las especies incluidas en este estudio pertenecen principalmente a algunos taxones concretos como mamíferos (Primates, Rodentia, Ungulata, Carnivora y Marsupialia), aves y peces óseos. En términos generales, puede decirse que, de los 32 genes encontrados en vertebrados, al menos 4 de ellos han surgido por duplicaciones posteriores a la colonización del medio terrestre y 2 de ellos parecen haberse perdido en el linaje de las aves. Otras modificaciones observadas son debidas principalmente a mutaciones como las sustituciones de aminoácidos o las inserciones/deleciones (indels).

Paralogous genes are a clear example of duplications and source of diversification in the evolution of the genomes of living beings. Here we addressed CTS genes that give rise to the proteolytic enzymes called cathepsins. After a thorough search in bibliographic databases and molecular sequences we found information about these genes is confusing. In order to identify the processes involved in their evolutionary diversification since the origin of terrestrial vertebrates to the present we analysed the phylogenetic relationships between the amino acid sequences of cathepsins. The species included in this study belong mainly to some specific taxa such as Mammals (Primates, Rodents, Ungulates, Carnivora and Marsupials), Birds and Bony Fishes. In general terms, among the 32 genes found in vertebrates, at least 4 of them have arisen due to duplications subsequent to the colonization of the terrestrial environment and 2 of them seem to have been lost in the lineage of the Birds. Other modifications observed are mainly due to mutations such as amino acid substitutions or insertions/deletions (indels).

Introducción

Es conocido que la Bioinformática es una disciplina basada en la recopilación, almacenamiento, organización, manejo, análisis y transmisión de información molecular referente a datos biológicos. La creciente acumulación de secuencias nucleotídicas de genes y genomas estimula, también de forma creciente, la creación de nuevas aplicaciones con utilidad en la investigación biológica. Actualmente, las bases de datos moleculares contienen gran cantidad de información genética acerca de numerosas especies de animales, plantas y microorganismos; y el análisis de grupos de genes, relacionados entre sí por sus características estructurales y funcionales, permite establecer relaciones de parentesco evolutivo entre ellos y entender algo acerca de la forma en que han ido evolucionando, llegando incluso a datar acontecimientos evolutivos de índole molecular no percibidos por otros medios, como es el caso de las

globinas^[1,2] o el de los genes homeóticos^[3]. A los genes que componen cada uno de esos grupos se les califica de «parálogos», al ser originados por duplicación a partir de un gen ancestral y su posterior diferenciación por mutaciones, y son un claro ejemplo de los fenómenos del mecanismo de amplificación de la información genómica en la evolución de los seres vivos.

Un grupo de este tipo de genes duplicados, aún poco estudiados desde esa perspectiva, es el de los *CTS*, que codifican ciertas enzimas con función proteolítica denominadas catepsinas, cuya característica principal es la presencia de un dominio proteasa hacia el final de su secuencia aminoacídica. Su interés radica en que presentan funciones en diferentes procesos celulares y en que su ausencia o mal funcionamiento provocan ciertas patologías o enfermedades^[4,5].

En este artículo, se pretende clarificar la, a veces confusa, información sobre estos genes y profundizar

en el conocimiento de los mecanismos que han podido participar en su diversificación, con el fin de encontrar un modelo evolutivo que la explique desde el origen de los vertebrados terrestres^[6]. Para ello, se ha realizado una búsqueda en bases de datos bibliográficas y de secuencias moleculares y se ha llevado a cabo su análisis filogenético mediante la generación de alineamientos múltiples de secuencias aminoacídicas y de las correspondientes filogenias para algunas de las especies más representativas de ciertos taxones. Este es un ejercicio que ha permitido además comprobar la utilidad actual de los recursos bioinformáticos para un biólogo molecular.

Métodos

En relación con las catepsinas conocidas en la especie humana, se han buscado en las bases de datos del National Center for Biotechnology Information (NCBI) genes «homólogos» (genes provenientes de ancestros comunes) a cada uno de los parálogos en especies pertenecientes a taxones bien diferenciados a nivel genómico y filogenético: primates, roedores, ungulados, carnívoros, marsupiales, aves, anfibios y peces óseos. Adicionalmente, se anotó la ubicación cromosómica de cada gen de catepsina para las especies analizadas (tabla 1).

Con las más de 400 secuencias aminoacídicas de diferentes especies encontradas, se realizaron alineamientos múltiples y árboles filogenéticos mediante el uso del software Seaview, con el fin de depurar la muestra a analizar y seleccionar las secuencias más representativas, excluyendo todas aquellas que mostraban errores inaceptables de distinta naturaleza.

Análisis de los resultados y discusión

Entre las secuencias halladas, destaca cierta problemática en su sinonimia ($C=DPP$, $H=I$, $L=L1$, $L2=V=U$, $1=7$, $2=8$, $P=J$...). Lo más interesante aquí es la ubicación sinténica de varias catepsinas (tabla 1), lo que sugiere una posible relación funcional, así como un alto grado de conservación evolutiva y una mayor proximidad filogenética. También destacan catepsinas aparentemente ausentes (quizás por delección o pérdida tras la duplicación) o no expresadas (por afuncionalización), principalmente en aves y peces o como el caso de la L2, ausente en roedores, y la G, en marsupiales; y catepsinas más o menos exclusivas de ciertos taxones, como las *placentally expressed cathepsins* (PEC) específicas de roedores, y otras adicionales (ver tabla 1) en peces óseos^[7,8,9,10]. En este último taxón, no se puede ser concluyente por la escasez de datos.

Tabla 1. Ubicación cromosómica de los genes de catepsinas en vertebrados. En los encabezados (fondo gris), los nombres de taxones, de las especies mejor estudiadas y de las diferentes catepsinas. PECs = catepsinas expresadas placentariamente. Los datos numéricos representan a los cromosomas implicados, apareciendo en celdas de un mismo color los indicativos de 'sinténias' (coincidencias en ubicación) confirmadas. Las casillas vacías indican ausencia o falta de datos en esas especies. Las casillas con '+' indican la presencia de dicha catepsina sin conocimiento de su ubicación cromosómica. Datos extraídos de la base de datos *Gene* del NCBI (ncbi.nlm.nih.gov).

TAXONES	ESPECIES		TINAG	TINAGL1	A	B	C	D	E	F	G	H	K	L	L2	O	S	W	Z			
Primates	Homo sapiens	Humano	6	1	20	8	11	11	1	11	14	15	1	9	9	4	1	11	20			
	Pan troglodytes	Chimpance	6	1	20	8	11	11	1	11	14	15	1	9	9	4	1	11	20			
	Gorilla gorilla	Gorila	6		20	8	11	11	1	11	14	15	1	9	9	4	1	11	20			
	Nomascus leucogenys	Gibón	22	12	13	4	15	4	5	4	22	6	12	1	1	7	12	4	13			
	Macaca mulatta	Macaco	4	1	10	8	14	14	1	14	7	7	1	15	15	5	1	14	10			
	Chlorocebus sabaeus	Mono Verde	17	20	2	8	1	1	25	1	24	26	+	12	12	7	+	1	2			
	Calithrix jacchus	Titi	4	7	5	13	11	11	19	11	10	10	18	1	1	3	18	11	5			
	Microcebus murinus	LemurR	6	2	18	20	5	5	27	5	9	2				14	2	5	18			
	Mus musculus	Ratón	9	4	2	14	7	7	19	14	9	+	3	13	3	3	19	2	13			
Roedores	Rattus norvegicus	Rata	8	5	3	15	1	1	13	1	15	8	2	17	2	2	1	3				
	Microtus ochrogaster	Topillo	5	10	8	17	22	8	6	8	+	+	21	16	+	1	23	8	8			
	Capra hircus	Cabra	23		13	8	29	29	+	29	21	21	3	8	8	17	3	29	13			
Ungulados	Bos taurus	Vaca	23	2	13	8	29	29	+	29	21	21	3	8	8	17	3	29	13			
	Bubalus bubalis	Bufalo de agua	2	2	14	3	5	5	+	5	20	20	6	3	3	17	6	5	14			
	Sus scrofa	Cerdo	7	6	17	14	9	2	+	2	7	7	4	10	10	8	4	2	17			
	Equus caballus	Caballo	20	2	22	2	7	12	5	12	1	+	5	23	23	2	5	12	22			
Carnívoros	Canis lupus familiaris	Perro	12	2	24	25	21	18	38	18	8	3	17	1	1	15	17	18	24			
	Felis catus	Gato	802	C01	A03	B01	D01	D01	F01	D01	803	803	C01	D04	D04	801	C01	D01	A03			
Marsupiales	Monodelphis domestica	Zarigüeya	2	4	1	1	4	+	2	8	+	1	2	+	5	2	8	1				
	Meleagris gallopavo	Pavo		25	22	2	1	5	26			12	27	2	4	27	27	22				
Aves	Gallus gallus	Pollo	3	23	20	3	1	5	26			10	25	2	4	25	+	20				
	Coturnix japonica	Codorniz	3	23	20	3	1	5	26			10	2	4	25	+	20					
	Taeniopygia guttata	Pinzón	3		20	3	1	5	+			10	2	4	+	20	20	20				
	Ficedula albicollis	Papamoscas		3	20	3	1	5	26			10	2	+	+	20	20	20				
Anfibios	Xenopus tropicalis	Rana		2	10	+	2	4	2	4	1	3	8	3	1	+	8	4	+			
	Salmo salar	Salmón			22	1	20						5	113	5	2	5					
Peces óseos	Oncorhynchus mykiss	Trucha		18	7	4								17		14			17			
	Oreochromis niloticus	Tilapia		22	5		14	1				7	11			2			5			
	Astatotilapia calliptera	Ciclido			5		14	1				7				2			5			
	Danio rerio	Zebrita		19	6	17	15	18		14		18	16	5	14	16	6	22	5	12	17	20
	Oryzias latipes	Medaka		11	5	3	13					6	9		10	11			5			
	Iticulus punctatus	Siluro			11	9	17	12				26	1	22		8	1	8	12			
	Cynoglossus semilaevis	Lenguado		18	11	5	4	5				6	2			5			11			
	Esox lucius	Lucio		10	17	15	1	2					19			4			17			
	Leptosteus oculatus	Catán		6	18	+	3	27	3	3				2	4		28	18				

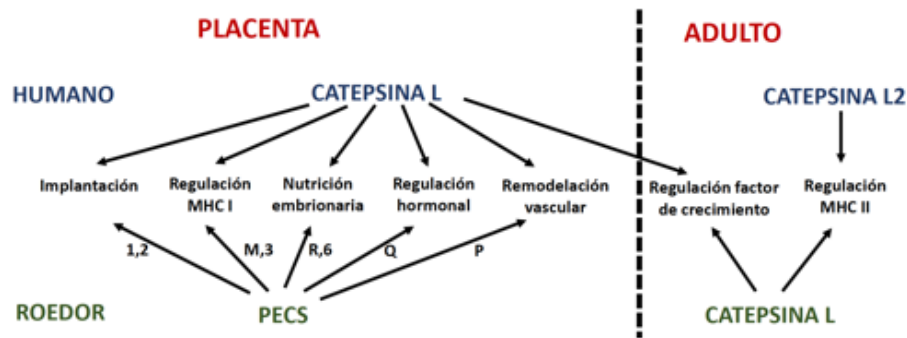


Figura 1: Posibles funciones de las catepsinas PEC en relación con las L y L2 (=V). En la placenta (izquierda), cada una de las PEC de los roedores cumple una de las funciones específicas entre las que posee la catepsina L de humanos. En el adulto (derecha), en cambio, la catepsina L de roedores adopta funciones de las L y L2 (V) en humanos (imagen redibujada de Manson, 2008^[9]).

Un particular objetivo de estos estudios es encontrar posibles relaciones funcionales que concuerden entre genes «ortólogos» (genes homólogos resultado de la divergencia de especies o linajes), es decir, genes que mantengan la misma función en especies diferentes. Para ello, se ha comprobado la expresión de cada catepsina en distintos tejidos en el ejemplo de humanos. En este sentido, llama la atención el caso algo excepcional de las PEC de roedores. Las catepsinas PEC, debido a que se encuentran en el mismo cromosoma que la L y a que la L2 está ausente en este taxón, parecen ser producto de duplicaciones de uno de esos dos genes. Estas duplicaciones parecen haber promovido un reparto de funciones entre las PEC en relación con las catepsinas L y L2 humanas (figura 1). De hecho, las funciones de las PEC, L y L2 están relacionadas, formando un grupo separado de los otros genes.

Curiosamente, la relación filogenética entre estas catepsinas y las más semejantes a nivel de secuencia (K y S) refuerza un modelo de sucesivas duplicaciones de genes en el linaje de los roedores (figura 2), como ya expusieron Sol-Church y otros^[7]. Este caso podría ser un ejemplo de que las modificaciones que afectan a los genes duplicados deben producirse bajo una menor presión de selección y evolucionar derivando hacia funciones más bien aleatorias que faciliten su adaptación a las nuevas situaciones.

En términos generales, puede decirse que, de los 32 genes identificados en vertebrados, todos presentan un alto grado de semejanza a nivel de proteína (por conservación evolutiva). Sin embargo, en el extremo inicial de sus secuencias, algunas de ellas, como la catepsina B, aparecen más conservadas que otras, como Tinag y TinagL, como se puede apreciar al alinear sus secuencias (figura 3). Es muy probable que las diferencias en el grado de conservación evolutiva estén relacionadas con

la importancia funcional de las distintas proteínas.

La mayoría de las modificaciones que se ponen de manifiesto en los alineamientos múltiples de estas secuencias son sustituciones de aminoácidos (consideradas como mutaciones puntuales), pero son las deleciones las que revelan mejor su diferente grado de conservación evolutiva, sobre todo, en la región ajena al dominio peptidasa.

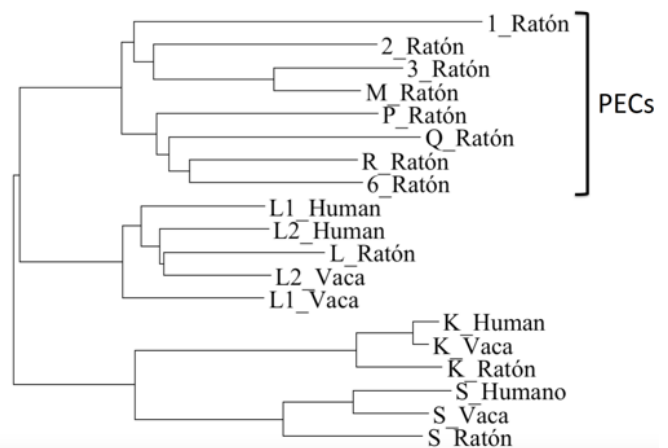


Figura 2: Filogenia simplificada de las PECs (1, 2, 3, 6, M, P, Q y R) en relación con las catepsinas más similares (L/L1, L2, K y S) de tres especies representativas (humano, ratón y vaca). El árbol fue obtenido por el método de distancias NJ (*Neighbour Joining*) a partir de un alineamiento múltiple realizado con el algoritmo *clustal*.

Por otra parte, el análisis filogenético de las principales catepsinas incluidas en este trabajo revela varios aspectos (figura 4), que sirven también como conclusión sobre el estudio de estas proteínas:

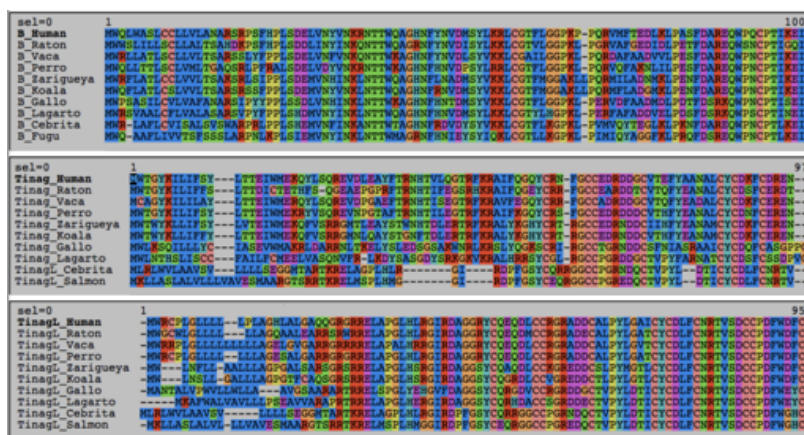


Figura 3: Sección inicial de los alineamientos múltiples de tres catepsinas. Sustituciones y deleciones muestran la variabilidad relativa de sus secuencias aminoácidas en especies representativas de los taxones analizados. Dado que Tinag no aparece en peces, se han utilizado las secuencias correspondientes al gen parálogo (TinagL) como grupo externo al resto de los taxones (alineamientos elaborados con el algoritmo *clustal*).

- Las duplicaciones son la base evolutiva de su diversidad, aunque no es posible, por el momento, determinar cuántos eventos y en qué períodos se han producido ni si abarcaron solo uno o más genes.
- Dado que los genes F y W aparecen sinténicos en mamíferos y que están presentes también en reptiles (datos no mostrados), cabe suponer que en el linaje de las aves su ausencia se deba más bien a una sola deleción de ambos genes.
- Los distintos tipos de catepsinas — serina (A), aspartato (B), cisteína (C) — citados en la figura 4 forman grupos monofiléticos y debieron diferenciarse funcionalmente tras las primeras duplicaciones en períodos muy primitivos de la evolución de esta familia de proteínas.
- Desde el origen de los vertebrados aparecen catepsinas que no estaban presentes en invertebrados (como las S y K) y más de 4 genes (Tinag, L2, W y las PECs) han surgido por duplicaciones posteriores a la colonización del medio terrestre.
- Otras modificaciones observadas son debidas principalmente a mutaciones del tipo pequeñas inserciones/deleciones (indels) y sustituciones de aminoácidos.

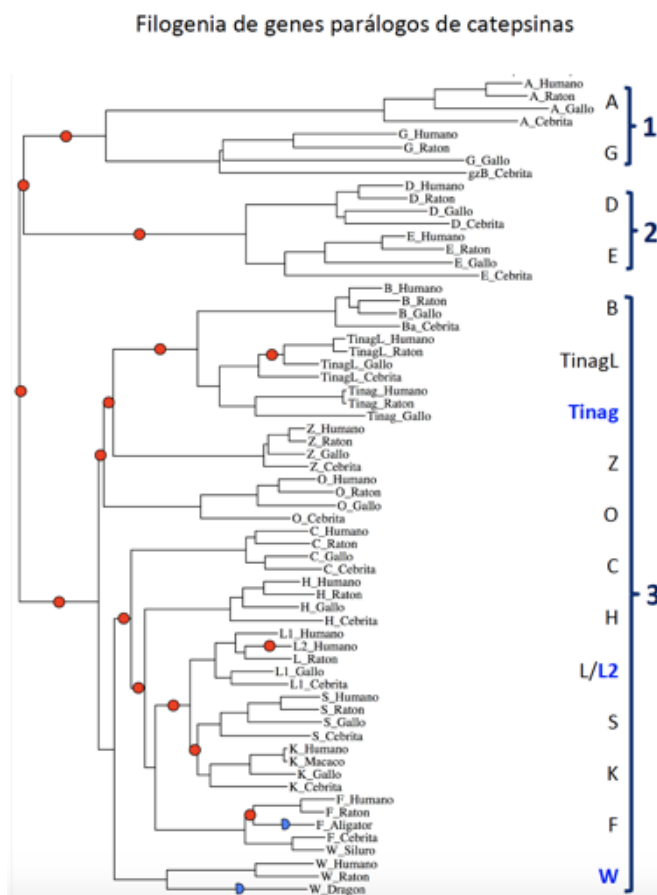


Figura 4: Árbol filogenético simplificado (obtenido por el método de distancias NJ) de las principales catepsinas, de una o dos especies de primates, roedores, aves/reptiles o peces. Círculos rojos: eventos de duplicación. Semicírculos azules: deleción de los genes F y W en aves. Letras en columna (derecha): nombres de catepsinas (en azul, las ausentes en peces). Los números representan respectivamente los tres tipos de catepsina establecidos, de serina (1), de aspartato (2) y de cisteína (3)^[12]. Destaca un ejemplo de sinonimia confusa en el gen F (aquí W) del Siluro.

Las futuras líneas de investigación permitirán concretar muchas de estas ideas, ampliar el espectro genético de especies, corregir errores de las secuencias moleculares conocidas y realizar estudios filogenéticos

específicos para cada uno de los genes. Mucho queda aún por conocer sobre las catepsinas en otras especies y sobre su diversidad funcional para llegar a conclusiones evolutivas más definidas, así como sobre sus posibles aplicaciones en ámbitos menos teóricos.

Una finalidad al respecto sería poder diferenciar los genes más importantes de los que tienen funciones secundarias.

Referencias

- [1] Taylor J y Raes, J. Small-Scale Gene Duplications. En: Ryan Gregory, T. (Ed.) *The Evolution of the Genome*, pp. 289-327. Academic Press, Cambridge, 2005.
- [2] Hoffmann F y otros. Evolution of the Globin Gene Family in Deuterostomes: Lineage-Specific Patterns of Diversification and Attrition. *Mol. Biol. Evol.*, 29: 1735-1745, 2012.
- [3] Carroll SB y otros. *From DNA to diversity: molecular genetics and the evolution of animal design*. Blackwell Science, Hoboken, Nueva Jersey, 2001.
- [4] Sun B y Chi H. Cathepsin S of *Sciaenops ocellatus*: Identification, transcriptional expression and enzymatic activity. *Int. J. Biol. Macromol.*, 82: 76-82, 2016.
- [5] Wang Y y otros. Identification and activity of a paralog of cathepsin S from yellow catfish (*Pelteobagrus fulvidraco*) involved in immune response. *Fish Shellfish Immunol.*, 61: 16-23, 2017.
- [6] Zhou J y otros. Evolutionary History of Cathepsin L (L-like) Family Genes in Vertebrates. *Int. J. Biol. Sci.*, 11: 1016-1025, 2015.
- [7] Sol-Church y otros. Evolution of placentally expressed cathepsins. *Biochem. Biophys. Res. Commun.*, 293: 23-29, 2002.
- [8] Gallwitz M y otros. Expansion of the mast cell chymase locus over the past 200 million years of mammalian evolution. *Immunogenetics*, 58: 655-669, 2006.
- [9] Mason R. Emerging functions of placental cathepsins. *Placenta*, 29: 385-390, 2008.
- [10] Raymond W y otros. How immune peptidases change specificity: cathepsin G gained tryptic function but lost efficiency during primate evolution. *J. Immunol.*, 185: 5360-5368, 2010.
- [11] Oikawa D y Iwawaki T. Positive contribution of IRE1 α -XBP1 pathway to the expression of placental cathepsins. *Biochem. Biophys. Res. Commun.*, 433: 426-431, 2013.
- [12] Kirkegaard T y Jäättelä M. Lysosomal involvement in cell death and cancer. *Biochim. Biophys. Acta*, 1793: 746-754, 2009.