J Mol Evol (2002) 54:30–34
DOI: 10.1007/s00239-001-0014-0

JOURNAL OF MOLECULAR EVOLUTION

The Evolutionary History of Prosaposin: Two Successive Tandem-Duplication Events Gave Rise to the Four Saposin Domains in Vertebrates

Einat Hazkani-Covo,[1] Neta Altman,[2] Mia Horowitz,[2] Dan Graur[1]

[1] Department of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel
[2] Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel

Received: 8 February 2001 / Accepted: 29 June 2001

Abstract. Prosaposin is a multifunctional protein encoded by a single-copy gene. It contains four saposin domains (A, B, C, and D) occurring as tandem repeats connected by linker sequences. Because the saposin domains are similar to one another, it is deduced that they were created by sequential duplications of an ancestral domain. There are two types of evolutionary scenarios that may explain the creation of the four-domain gene: (1) two rounds of tandem internal gene duplication and (2) three rounds of duplications. An evolutionary and phylogenetic analysis of saposin DNA and amino acid sequences from human, mouse, rat, chicken, and zebrafish indicates that the first evolutionary scenario is the most likely. Accordingly, an ancestral saposin-unit duplication produced a two-domain gene, which, subsequently, underwent a second complete tandem duplication to give rise to the present four-domain structure of the prosaposin gene.

Key words: Prosaposin — Saposin domains — Vertebrates — Tandem duplication

Introduction

Prosaposin is a multifunctional protein encoded by a single-copy gene. It contains four saposin domains (A, B, C, and D) occurring as tandem repeats connected by linker sequences. The four functional saposins are generated by postranslational processing of the prosaposin precursor in the lysosome. Each saposin is relatively specific with respect to substrate, i.e., it activates a specific glycosphingolipid hydrolase in the lysosome, but some overlapping specificities are known (Sandhoff et al. 1995). In addition to serving as a precursor of four saposins, intact prosaposin has an *in vitro* nerve-regenerating function, whose active site is most probably located within the saposin C domain (Qi et al. 1999).

All four saposins possess several primary features in common, e.g., a length of about 80 amino acids, six cysteines as homologous positions, a glycosylation site, and a conserved proline. It is, therefore, reasonable to assume that the prosaposin gene was created by duplications of a single ancient saposin domain.

Several scenarios can explain the creation of prosaposin from a single ancestral saposin domain. One scenario invokes two rounds of duplication and six scenarios invoke three rounds of duplication (Fig. 1). According to the two-step scenario, an ancestral single-saposin gene was duplicated and gave rise to a gene containing two saposin domains in tandem. Subsequently, a duplication involving both domains occurred and the four-domain prosaposin was produced. According to this scenario, saposins A and C are phylogenetically more closely related to each other than either is to saposins B or D, and conversely, saposins B and D are phylogenetically closer to each other than either is to saposins A or C. In an unrooted phylogenetic tree we

Correspondence to: Dan Graur; *email*: graur@kimura.tau.ac.il

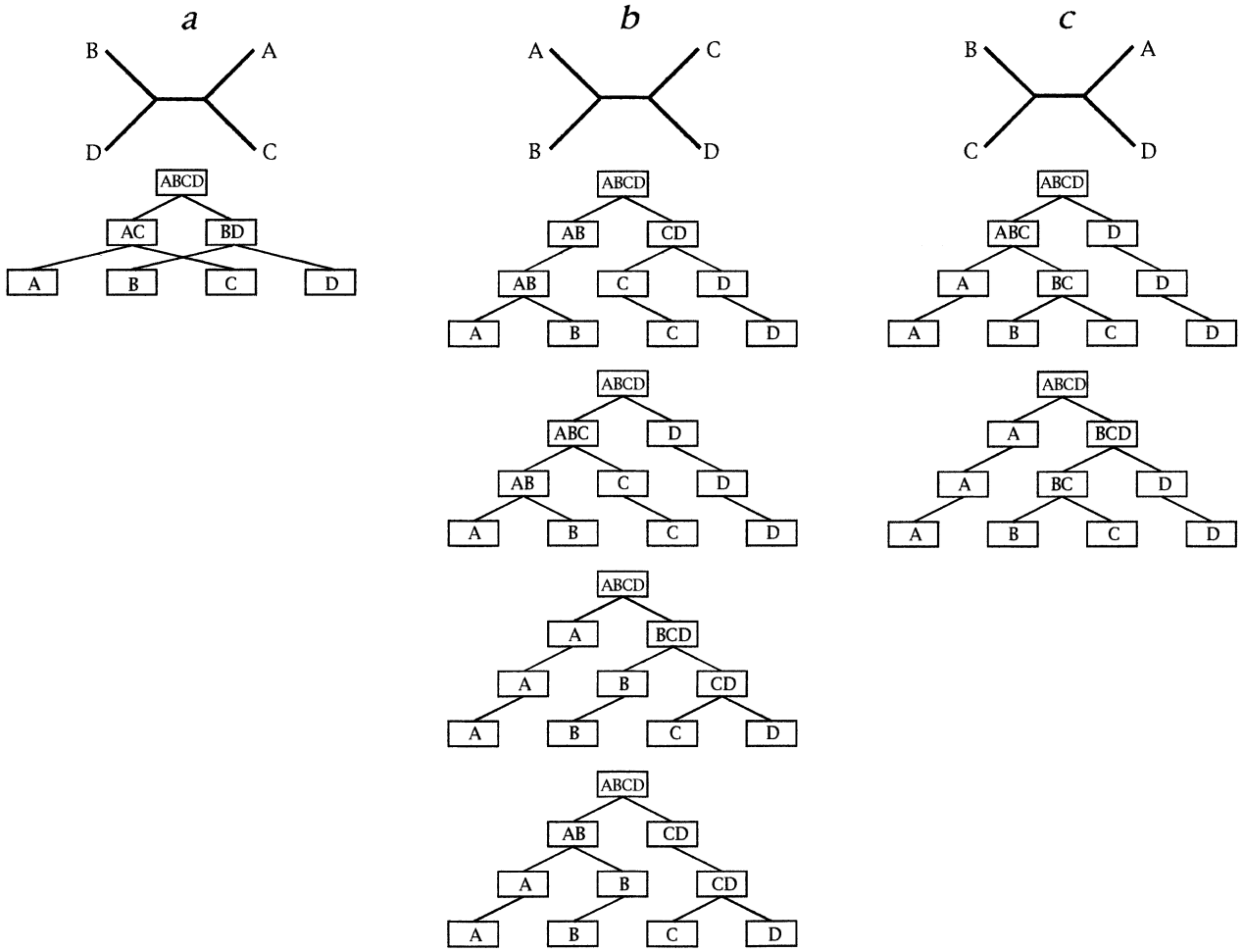

Fig. 1. Possible evolutionary scenarios for the creation of the four-domain prosaposin through sequential tandem duplications. *a* A two-step scenario. *b* and *c* Four three-step and two three-step scenarios, respectively. The unrooted topologies for saposins A, B, C, and D that are compatible with the evolutionary scenarios are shown at the top.

will, thus, expect saposins A and C to be neighbors and saposins B and D to be neighbors (Fig. 1a). According to the other six scenarios, the evolution of prosaposin from a single-domain gene required three internal duplications. The first duplication created a two-domain coding gene. The second duplication involved only one of the resulting domains and produced a three-domain coding gene. Finally, in the third round of duplication, one of the three saposin domains was duplicated to produce prosaposin. The six possible three-duplication scenarios may yield two unrooted phylogenetic topologies (Figs. 1b and c), and depending on the order in which the single domains were duplicated. We note that the tree in Fig. 1a cannot be obtained via three rounds of duplication and that the trees in Figs. 1b and c cannot be obtained via two rounds of duplication, unless one assumes the occurrence of additional processes, e.g., unequal crossing-over (Rorman et al. 1992). We further note that evolutionary scenarios other than the ones in Fig. 1 are possible, however, obtaining the A, B, C, and D subunits in linear sequential order would require additional processes, such as transposition and unequal recombination.

In this study, we attempted to reconstruct the order of internal duplications that gave rise to the four saposins by using phylogenetic tools. In our analysis, we assumed (1) that all internal duplications resulted in tandemly repeated sequences and (2) that the unalignable intersaposin linker sequences can be safely ignored in the phylogenetic analyses.

Materials and Methods

Prosaposin Data

Five prosaposin DNA and amino acid sequences were collected from human (*Homo sapiens*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), chicken (*Gallus gallus*), and zebrafish (*Danio rerio*). The accession numbers for the DNA and amino acid sequences are M32221 and AAA60303 for human, U57999 and AAB02695 for mouse, M19936 and AAA42136 for rat, AF108656 and AAF05899 for chicken, and AF108655 and AAG32919 for zebrafish. The prosaposin sequences were processed into single saposins according to O'Brien and Kishimoto (1991). Only complete prosaposin sequences were used; partial sequences even if they included complete saposin domains were omit-

ted from the analysis. The 20 saposins were aligned using ClustalW (Higgins et al. 1996). Manual modifications of the multiple alignments were performed using SEAVIEW (Galtier et al. 1996). The alignments are available at http://kimura.tau.ac.il/. Subsequent analyses were performed at both the protein and the DNA levels.

Identification of Nonvertebrate Prosaposin-Like Sequences

We used FindPatterns (Womble 2000) and PSI-BLAST (Altschul et al. 1997) to search for nonvertebrate "prosaposin like proteins," i.e., proteins consisting of two or more saposin-like domains. In our searches, we looked for at least two repetitions (separated by a linker sequence of any length) of the following motif C–X(2)–C–X(27,28)–C–X(10,11)–C–X(23,25)–C–X(5)–C, where C denotes cysteine, X denotes any amino acid, and the numbers in parentheses denote either the number or the range of consecutive residues.

Phylogenic Analyses

A maximum-likelihood species tree for the prosaposin gene sequences was reconstructed with the DNAML program (Felsenstein 1993). SEQBOOT and CONSENSE from the PHYLIP package were used for computing bootstrap values. Similarly the maximum-likelihood method was used to ascertain the monophyly of each paralogous saposin-domain group. In the final analysis we computed the likelihood values of three user trees representing all possible topological relationships among the four saposins. Within each saposin group the species phylogenetic topology was fixed. In this part of the study, we used DNAML as well as PROTML (Adachi and Hasgawa 1996) with the JTT amino-acid replacement model (Jones et al. 1992).

Phylogenetic analyses of saposins were also performed by using inferred ancestral protein sequences. That is, for each of the four groups of saposins, the ancestral amino acid sequence was inferred using FastML (Pupko et al. 2000), and the three possible unrooted phylogenies were tested by PROTML.

Results

A maximum-likelihood phylogenetic reconstruction for the prosaposins from the five taxa is shown in Fig. 2. The congruence between this tree and the classical species tree rules out most processes, such as independent taxon-specific evolution, that may confuse the phylogenetic picture. A phylogenic reconstruction of the 20 saposins (Fig. 3) indicates that each orthologous group is monophyletic. The maximum-likelihood reconstruction differs from the one shown in Fig. 3 in the placement of chicken and zebrafish within the saposin C and D groups. However, the bootstrap values on these branches were very low, so that these misplacements may be the result of long-branch attraction (Felsenstein 1978). Thus, within each saposin group, we assumed that the species topology in Fig. 2 is the true tree.

All analyses indicate that the tree in Fig. 1a is the most likely, followed in descending order by the trees in Figs. 1b and c. However, as shown in Table 1 the likelihoods were not statistically different from one another. All other methods of phylogenetic reconstruction (e.g., neighbor joining, maximum parsimony) yielded essentially the same results and identical levels of branch resolution (not shown).

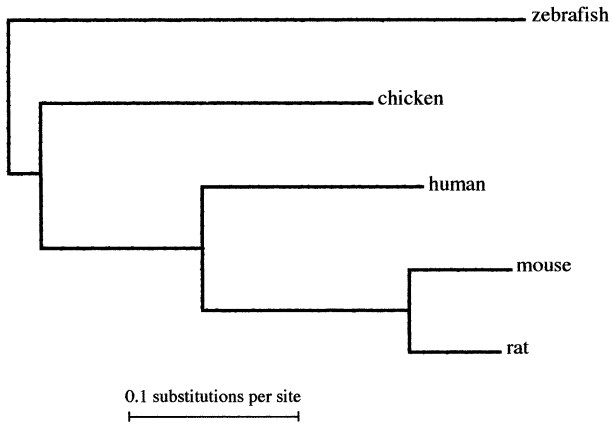

Fig. 2. Maximum-likelihood phylogenetic tree of prosaposin DNA sequences from five vertebrate taxa. All internal branches are supported by 100% bootstrap replicates.

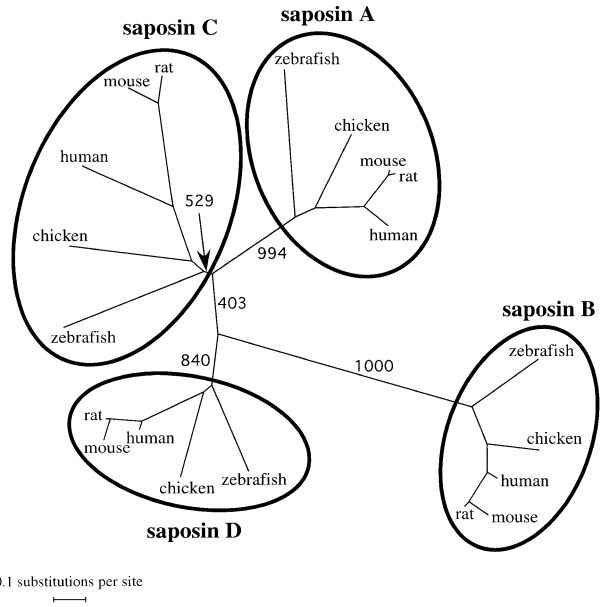

Fig. 3. A scaled phylogenetic reconstruction of 20 saposin DNA sequences indicates that each orthologous saposin group is monophyletic (ellipses). The maximum-likelihood reconstruction differs from the one shown here in the placement of chicken and zebrafish within the saposin C and D groups (see text). Numbers on the branches leading to the four saposin clades indicate bootstrap values based on protein maximum-parsimony reconstructions with 1000 pseudo-replicates.

In the entire nonredundant NCBI database, we identified five nonvertebrate prosaposin-like proteins (Table 2). Several preliminary phylogenetic analyses (not shown) indicate that the duplications giving rise to the nonvertebrate prosaposin-like genes have occurred independently in each of the major lineages (insects, nematodes, slime molds, and sarcodine amoebae) and independently of the duplications in the vertebrates.

Table 1. Comparison of likelihoods (L) among user-specified trees[a]

	DNA ML		Protein ML		Protein ML of ancestral sequences	
Tree topology	ln*L*	Δln*L* ± SE	ln*L*	Δln*L* ± SE	ln*L*	Δln*L* ± SE
((SapA, SapC), SapB, SapD)	−4324.5	ML	−2418.8	ML	−797.0	ML
((SapA, SapB), SapC, SapD)	−4327.7	−3.19 ± 3.8	−2422.2	−3.3 ± 3.9	−798.9	−1.9 ± 3.3
((SapA, SapD), SapB, SapC)	−4328.5	−3.95 ± 3.2	−2422.2	−3.4 ± 3.7	−799.1	−2.1 ± 3.3

[a]Each terminal node in DNA ML and protein ML represents a species tree for a saposin, e.g., SapA represents the tree (((mouse, rat), human), chicken, zebrafish) for saposin A. Maximum-likelihood trees are marked ML.

Table 2. Nonvertebrate prosaposin-like proteins

Organism	Accession No.	Number of saposin-like subunits	Length of protein (aa)
Drosophila melanogaster	AAD38622	8	953
Bombyx mori	BAA23126	7	965
Caenorhabditis elegans	T15674	2	314
Dictyostelium discoideum	AAB06759	4	456
Naegleria fowleri	AAK21658	2	307

Discussion

In our analyses, the most likely tree was the one in which saposins A and C are neighbors, therefore, supporting the two-step evolutionary scenario. Admittedly, the differences among the log likelihoods (Table 1) are not statistically significant, however, there are additional pieces of evidence that strengthen our confidence in the correctness of the maximum-likelihood phylogenetic tree. First, there is a high similarity between the neurotrophic domain of saposin C and the homologous counterpart in saposin A, which does not exhibit any neurotrophic activity. Interestingly, it has been shown that the neurotrophic activity of saposin A can be restored by the replacement of a single amino acid (Qi et al. 1999).

Second, the lysosomal proteolysis of prosaposin indicates that regions immediately following the saposin A and C domains may be functionally homologous. According to Hiraiwa et al. (1997), the first stage in the proteolysis of prosaposin yields two trisaposins: trisaposin A, containing domains A, B, and C, and trisaposin B, containing domains B, C, and D. These two trisaposins are produced following the hydrolysis by cathepsin D of sites in the linker regions that are located next to the saposin A and C domains. These cathepsin D recognition sites suggest a possible structural similarity between saposin A and saposin C. We note, however, that it is not clear whether cathepsin D recognizes domains structure or linker regions.

Finally, both saposins activate β-galactocerebrosidase and glucocerebrosidase. Saposin A activates β-galactocerebrosidase in cells (Harzer et al. 1997), and glucocerebrosidase *in vitro* (Morimoto et al. 1989), while saposin C activates glucocerebrosidase *in vivo* (Sandhoff et al. 1995; Ho and O'Brien 1971) and β-galactocerebrosidase in cells (Harzer et al. 1997).

The evolutionary evidence provided by intron positions within the paralogous saposin domains is somewhat equivocal (Rorman et al. 1992). In Fig. 4, we see that one intron is positionally homologous between saposin A and saposin C, and one between saposin B and saposin D, thus supporting the tree in Fig. 1a. One additional intron has positional homology between saposin C and saposin D, supporting the tree in Fig. 1b. Allowing for intron sliding, three intron positions support the tree in Fig. 1b, and two positions support the tree in Fig. 1a. None of the positions was common to saposins A and D (or B and C), and one intron was unique to saposin D. These results can only be used to reject the two evolutionary scenarios represented by the unrooted tree in Fig. 1c. We note, however, that the information concerning intron positions indicates that phylogeny-obscuring events, such as crossing-over, may have occurred during saposin evolution.

In conclusion, it seems that the two-step evolutionary scenario is the most likely. Accordingly, an ancestral saposin-unit duplication produced a two-domain gene, which, subsequently, underwent a second duplication involving both tandemly repeated units to give rise to the present four-domain structure of the prosaposin gene. Interestingly, the most parsimonious scenario in terms of number of steps proved to be the most likely, despite the fact that parsimony considerations were not taken into account in the analysis.

Because four-domain prosaposins were found in all Euteleostomi (bony vertebrates) studied to date, we conclude that the duplications leading to the creation of the four-domain protein occurred before the divergence of Actinopterygii (ray-finned fishes) from Tetrapoda (i.e., 300–450 million years ago). The subsequent evolution of the saposin domains during such a long period of time has a negative effect on our chances to resolve unambiguously the phylogenetic relationships among the four saposins. Many saposin-like proteins (SAPLIPs) are known in the literature (Munford et al. 1995), and a few of them are known to include more than a single saposin-like domain. However, the multidomain structure of the nonvertebrate prosaposin-like proteins that have been identified in this study has arisen independently of the

Fig. 4. Introns position within human and mouse saposins. *Capital letters* represent saposin regions; *lowercase letters* represent linker regions. *Vertical lines* represent intron positions. The alignment shown here is somewhat different from that of Rorman et al. (1992).

four-saposin domain in vertebrates and, thus, cannot be used to elucidate the order of internal gene duplications in the vertebrate lineage. In addition, the nonvertebrate sequences are almost unalignable with the vertebrate ones and cannot be reliably used to root the saposin tree. Thus, a further characterization of either the time of duplication or the primary sequence of the ancestral saposin subunit is unattainable.

Acknowledgments. We wish to thank Tsadok Cohen and Tal Pupko for their advice. This study was supported in part by the Magnet Da'at Consortium of the Israel Ministry of Industry and Trade.

References

Adachi J, Hasegawa M (1996) MOLPHY. Version 2.3: Programs for molecular phylogenetics based on maximum likelihood. Comput Sci Monogr 28:1–150

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27:401–410

Felsenstein J (1993) PHYLIP: Phylogeny inference package and manual. Version 3.5. Department of Genetics, University of Washington, Seattle

Galtier N, Gouy M, Gautier C (1996) SEAVIEW and PHYLO_WIN: Two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci 12:543–548

Harzer K, Paton BC, Christomanou H, Chatelut M, Levade T, Hiraiwa M, O'Brien JS (1997) Saposins (*sap*) A and C activate the degradation of galactosylceramide in living cells. FEBS Lett 417:270–274

Higgins DG, Thompson JD, Gibson TJ (1996) Using CLUSTAL for multiple sequence alignments Methods Enzymol 266:383–402

Hiraiwa M, Martin BM, Kishimoto Y, Conner GE, Tsuji S, O'Brien JS (1997) Lysosomal proteolysis of prosaposin, the precursor of saposins (sphingolipid activator proteins): Its mechanism and inhibition by ganglioside. Arch Biochem Biophys 341:17–24

Ho MW, O'Brien JS (1971) Gaucher's disease: Deficiency of "acid"-glucosidase and reconstitution of enzyme activity *in vitro*. Proc Natl Acad Sci USA 68:2810–2813

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282

Morimoto S, Martin BM, Yamamoto Y, Kretz KA, O'Brien JS, Kishimoto Y (1989) Saposin A: Second cerebrosidase activator protein. Proc Natl Acad Sci USA 86:3389–3393

Munford RS, Sheppard PO, O'Hara PJ (1995) Saposin-like proteins (SAPLIP) carry out diverse functions on a common backbone structure. J Lipid Res 36:1653–1663

O'Brien JS, Kishimoto Y (1991) Saposin proteins: Structure, function, and the role in human lysosomal storage disorders. FASEB J 5: 301–308

Pupko T, Pe'er I, Shamir R, Graur D (2000) A fast algorithm for joint reconstruction of ancestral amino acid sequences. Mol Biol Evol 17:890–896

Qi X, Kondoh K, Krusling D, Kelso GJ, Leonova T, Grabowski GA (1999) Conformational and amino acid residue requirements for the saposin C neuritrogenic effect. Biochemistry 38:6284–6291

Rorman EG, Scheinker V, Grabowski GA (1992) Structure and evolution of the human prosaposin chromosomal gene. Genomics 13: 312–318

Sandhoff K, Harzer K, Furst W (1995) Sphingolipid activator proteins. In: Scriver CR, Beaudet AL, Sly WS, Valle D (eds) The metabolic and molecular bases of inherited disease. 7th ed. McGraw–Hill, New York, Vol II, pp 2427–2441

Womble DD (2000) GCG: The Wisconsin package of sequence analysis programs. Methods Mol Biol 132:3–22