Molecular Evolution of Serine/Arginine Splicing Factors Family (SR) by Positive Selection

Andrés J. Gutiérrez Escobar, Aylan Farid Arenas and Jorge Enrique Gómez-Marin*

*Grupo de Parasitología Molecular (GEPAMOL), Centro de Investigaciones Biomédicas, Facultad de Ciencias de la Salud, Universidad del Quindío, Armenia, Colombia

Edited by E. Wingender; received 17 April 2006; revised and accepted 11 June 2006; published 22 June 2006

ABSTRACT: The serine-rich (SR) protein family is involved in the pre-mRNA splicing process and the DNA sequences of the corresponding genes are highly conserved in the metazoan organisms. The mammalian SR proteins consist of one or two characteristic RNA binding domains (RBD), containing the signature sequences RDAEDA and SWQDLKD and a RS (arginine/serine-rich) domain. We used the amino acid and nucleotide sequences deposited in GenBank and Swiss-Prot databases to perform a phylogenetic analysis using bioinformatics tools. The results of the phylogenetic trees suggest that this family has evolved by several gene duplication events as a result of a positive selection mechanism.

KEYWORDS: SR protein family, splicing, Toxoplasma, molecular evolution

INTRODUCTION

Members of the serine/arginine (SR)-rich protein family are essential components of the basic metazoan splicing machinery [1–3]. SR proteins contain one or two N-terminal RNA binding domains (RRM) [4, 5] and a variable-length C-terminal arginine/serine (RS)-rich domain [2,3]. Previous studies have shown that SR proteins interact with the other components of the basic splicing machinery, such as the U1 small nuclear ribonucleoprotein (snRNP)-specific 70 kDa protein (U1-70 K) and the 35 kDa subunit of the U2 auxiliary factor (U2AF35) [6,7]. SR proteins function in two distinct aspects of the splicing reaction. First, SR proteins are required for the splicing of all metazoan introns and therefore play a central role in the basic splicing reaction promoting the binding of U1 snRNP to 5’-splice sites [7–10]. Secondly, a large number of splicing regulatory elements, termed splicing enhancers, have been identified and are normally located downstream of the regulated intron. SR proteins have been shown to bind to many of these elements [2,3]. Alternative splicing has an important role to maintain the proteomic diversity in natural populations [11]. Here, we describe the evolutionary history for the SR gene family using the sequences deposited in the databases.

*Corresponding author. Tel./Fax: +57 67 460168; E-mail: gepamol2@uniquindio.edu.co.

Electronic publication can be found in In Silico Biol. 6, 0032 <http://www.bioinfo.de/isb/2006/06/0032/>, 22 June 2006.

1386-6338/06/$17.00 © 2006 – IOS Press and Bioinformation Systems e.V. and the authors. All rights reserved
METHODS

Data Sets

We searched the GenBank database for sequences that have the structural characteristics of SR proteins [1]. To find all available sequences that belong to the SR family, gapped BLAST and PSI-BLAST searches were performed in the Swiss-Prot database using the *Rattus norvegicus* (gi|57164147) SR family member as a query sequence. After an exhaustive search, partial and redundant sequences were removed from further analysis. The final data set includes 24 complete SR sequences of protist to mammalian origin.

Alignments and phylogenetic analysis

A total of 24 SR gene sequences from the following species were obtained from GenBank (gene numbers are shown in brackets): 2 protists: *Toxoplasma gondii* (1), *Plasmodium yoelii yoelii* (2); 2 plants: *Arabidopsis* (2), *Oryza* (1); 7 animals: *Xenopus* (3), *Gallus* (3), human (4), mouse (4), rat (2) and *Danio* (2). The mRNA sequences were aligned by using Clustal X [12] using the multiple alignment parameters: Gap Opening: 7.55, Gap Extension: 3.33, Delay Divergent Sequences: 24%, DNA Transition Weight: 0.5 and Use Negative Matrix: ON [12] and checked visually for any possible errors afterwards. The proportions of synonymous and nonsynonymous differences per site were computed by the modified Nei-Gojobori method using the overall average tool [13]. Phylogenetic trees were constructed by the neighbor-joining method [14]. All analyses were conducted by using the computer program MEGA, Version 3.1 [15].

RESULTS

Looking for SR genes in protozoa, we found a new gene in *Toxoplasma gondii* (contig TGG 993202) using the Blast program offered by ToxoDB. This sequence contains the principal domain (RRM) and the signature sequence RDAEDA that define the SR family protein. To further understand the driving force for sequence divergence of SR genes during evolution, we tested whether positive selection has occurred by estimating the number of synonymous substitutions per synonymous site ($d_S$) and the number of nonsynonymous substitutions per nonsynonymous site ($d_N$) using the modified method of Nei and Gojobori [13]. We computed these values for 24 members of the SR gene family and found that the mean $d_N$ was significantly greater than mean $d_S$ in the SR genes ($d_S$: 0.3722 ± 0.3266, $d_N$: 1.0000 ± 0.3363). These results support the hypothesis that positive selection has acted to diversify the SR gene family.

The nucleotide phylogenetic tree indicates that the evolution of the SR gene family occurred by gene duplication following the birth and death model by positive selection (Fig. 1). The phylogenetic protein tree clearly shows a good congruence with the nucleotide sequences suggesting that this family evolves to improved new types of proteins with different functions.

DISCUSSION

Alternative splicing is a major mechanism for modulating the expression of cellular and viral genes and enables a single gene to increase its coding capacity, allowing the synthesis of several structurally and
Fig. 1. (a) The phylogenetic tree of the SR gene family. The neighbor-joining algorithm was used to infer the topology based on the multiple sequence alignment with $p$-distance. All bootstrap scores are presented. The black arrows represent the time positions of three rounds of gene duplications, the red arrow represents the ancestral node in the origin of the SR gene family that produce the clade A and B. The *Toxoplasma gondii* SR new putative gene (Black Circle). Hs: Homo sapiens, Cf: Canis familiaris, Mm: Mus musculus, Xt: Xenopus tropicalis, Dr: Danio rerio, Rn: Rattus norvegicus, Tg: Toxoplasma gondii, Os: Oryza sativa, At: Arabidopsis thaliana, Py: Plasmodium yoelii yoelii, Pf: Plasmodium falciparum, Gg: Gallus gallus; the number alongside is the gi code of GenBank database. (b) The phylogenetic tree of the SR protein sequences.

functionally distinct protein isoforms [11]. Four successive gene duplications played an essential role in the case of serine/arginine (SR)-rich splicing factors evolution (Fig. 1). The SR genes have high rates of nonsynonymous substitutions due to positive selection favoring the gain of new functions. This selection model drives rapid change in the protein sequences and, therefore, will probably generate diversity that it is beneficial for the fitness within and between the species.

Gene duplication followed by positive selection indeed has been observed in several gene families involved in immune responses as those from the major histocompatibility complex and those from the immunoglobulin proteins [16,17]. Divergence of these immune genes often leads to either an additional layer of functional redundancy or acquisition of functional novelties. Therefore, as different species live in quite different ecological niches, the production of species-specific SR proteins would presumably allow them to better respond to the specific challenges that they face.

On the other hand, we report here the identification of a new SR gene for *Toxoplasma gondii* with a good level of structural similarity with other SR family members. It is important to note that this protozoa parasite exhibits a clonal population structure [18] and that the role of the alternative splicing process would be an important factor for gaining proteomic diversity in *Toxoplasma gondii*.
REFERENCES