**SEARCHING FOR THE IMMUNOGLOBULIN SUPERFAMILY GENES IN C.ELEGANS.**

TARANIN A.V.+, ROGOZIN I.B.

Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, 10 Lavrentiev Ave., Novosibirsk, 630090, Russia;
e-mail: taranin@bionet.nsc.ru

+Corresponding author

Keywords: genes, immunoglobulin superfamily, exon/intron structure prediction

The immunoglobulin superfamily (IgSF) is the largest known group of related proteins. It includes hundreds of members believed to have arisen from a single primordial Ig-like domain during evolution of multicellular metazoans (Williams and Barklay, 1988; Doolittle, 1995). The upcoming fruition of the*C.elegans*genome sequencing project (expected to be completed in 1998) offers opportunities to reveal structural and functional relationships of vertebrate and invertebrate IgSF members and to reconstruct molecular events which have led to the diversification of IgSF. Of particular importance is the opportunity to trace the evolution of complex functional systems built of IgSF proteins, such as, for example, the vertebrate immune system.

Establishement of evolutionary relationships among worm and vertebrate IgSF members is not an easy task because of the abundance of Ig-like domains in metazoan proteins, similar homology level among such domains from different proteins and mosaic structure of many IgSF proteins. Since a bulk of information available for*C.elegans*is presented as genomic sequences and expression sequence tags (EST), accuracy of gene structure predictions is of critical importance for correct identification and assignment of IgSF members in this species. The*C.elegans*Genome Sequencing Consortium currently uses the Genefinder program for gene structure predictions. The program is based only on analysis of potential functional signals and statistical properties of protein coding regions. Programs of such kind were shown to be not completely reliable in the case of large genomic sequences with complex exon/intron structure (Burset and Guigo, 1996; Milanesi and Rogozin, 1998).

The aims of this study were to estimate the amounts of IgSF encoding genes in*C.elegans*genome and to compare applicability of different gene prediction approaches for identification of worm IgSF members.

TBLASTN search of 6-frame translated*C.elegans*genomic sequences (URL: www.sanger.ac.uk/Projects/C\_elegans/blast\_server.shtml) against a set of structurally different vertebrate IgSF proteins showed that nearly 70 cosmids contain genes encoding proteins with Ig-like domains. At the time of the search, 38 of these cosmids have been completely sequenced by the Consortium, and corresponding Genefinder predicted proteins have been deposited in the WormPep database (URL:[www.Sanger.ac.uk/ Projects/C\_elegans/wormpep/](http://www.sanger.ac.uk/%20Projects/C_elegans/wormpep/)).

Six of 38 genes have been previously experimentally characterized and assigned to distinct vertebrate homologs. The predicted Ig-domain containing proteins for the remaining 32 cosmids were taken from the WormPep database, and databases of amino acid sequences, as well as our collection of 465 IgSF members, were searched for their closest matches. The results showed that the majority of predicted worm IgSF proteins could not be reliably indentified with respect to vertebrate or insect IgSF members. We explained this to be mainly due to the errors in protein predictions resulting from incorrect exon/intron structure determinations by Genefinder program.

The particular cases of cosmids C18f3 and C33f10 were studied in detail. According to the Genefinder predictions each cosmid include two linked genes for the Ig-domain containing proteins: C18f3.2/C18f3.3 and C33f10.5/C33f10.6. We found that C18f3.2 and C18f3.3 predicted proteins are closest to the different parts of the L1-like cell adhesion molecules. Similarly, C33f10.5 and C33f10.6 match predominantly with the different parts of the F11-like cell adhesion molecules. These data suggested that C18f3 and C33f10 may contain genes homologous to L1 and F11, respectively.

In vertebrate species, five groups of cell adhesion molecules composed of Ig and FnIII-domains (Ig/FnIII family) have been described: L1 (6 Ig-domains + 5 FnIII-domains), F11 (6 + 4), NCAM (5 + 2), DCC (4 + 6) and Robo (5 + 3). Each subfamily includes 2-5 members distinguished on the basis of their function and primary structure. All these molecules play important role in neural tissue differentiation (Walsh and Doherty, 1997). In invertebrate species unique representatives of each group (except F11) were found, thereby suggesting that duplications within subfamilies occurred mainly during deuterostomes evolution. In*C.elegans*, only DCC and Robo homologs have been described.

The nucleotide sequences of C18f3 and C33f10 cosmids were further analysed using the GeneBuilder program for gene prediction (URL:www.itba.mi.cnr.it/webgene (Milanesi, D’Angelo, Rogozin, manuscript in preparation)). The program maps exons on the basis of both statistical analysis and searches for homologous sequences in protein databases and dbEST. Such approach, first suggested by Gish and States (1993) has been shown to significantly improve accuracy of gene structure prediction (Rogozin et al., 1996; Gelfand et al., 1996). Fruit fly neuroglian (L1-subfamily) and rat NB3 (F11-subfamily) were used as key proteins for the C18f3 and C33f10 gene predictions, respectively. As a result, several variants of prediction were obtained in each case. The GeneBuilder program was able to find the Ig-domain encoding exons in the cosmid regions predicted by Genefinder as intergenic regions. The proteins predicted in the first round of analysis were aligned with key proteins, partially corrected on the basis of TBLASTN search results and used as new keys in the second round of analysis by GeneBuilder.

The final prediction assumes C18f3 cosmid to contain a gene, encoding L1-like protein (28-30% overall homology). Like other members of the L1-subfamily the predicted protein is composed of 6 Ig-like domains, 5 FnIII-like domains, a transmembrane region and an intracellular region containing characteristic ankyrin-binding motif. Similarly, cosmid C33f10 was predicted to contain a gene for F11-like protein (26-28% of homology) with 6 Ig-like domains, 4 FnIII-like domains and a stretch of the C-terminal hydrophobic residues suggesting its membrane attachment via GPI anchor. The assignment of the predicted proteins to the L1 and F11 cell adhesion molecules was further confirmed by pairwise comparisons of the distinct Ig- and FnIII-domains. Combined with the previously obtained data, these results show that all the members of Ig/FnIII family of cell adhesion molecules diverged at an early stage of metazoan evolution.

A striking feature of the vertebrate Ig/FnIII adhesion molecules is conserved exon/intron arrangment of their genes. Both Ig and FnIII-like domains of these molecules are encoded by two exons. Moreover, the interdomain introns are exclusively in phase 1, whereas the intradomain introns may have different intron phases in the genes for all these molecules.

The predicted exon/intron arrangement of the*C.elegans*L1- and F11-like genes is different from each other and from that of vertebrate genes. Some of the Ig-like domains of both L1 and F11 homologs are encoded by three exons and vice versa, some exons code for three domains (Fig.1). In contrast to 22 and 20 exons for extracellular domains in vertebrate L1 and F11 genes, their*C.elegans*counterparts have only 13 and 11 exons, respectively. However, all the interdomain introns are in phase 1 in both worm genes, like in vertebrates.

The predicted*C.elegans*proteins were aligned with their vertebrate homologs, and location of intron positions was compared in the extracellular regions. Of 19 intron positions common to the vertebrate L1 and F11 genes, 9 are completely conserved, 1 is separated by a codon and 2 are separated by two codons. Of 10 positions common to the vertebrate and worm L1 genes, 4 are identical and 2 are separated by a codon. Of 9 positions common to the vertebrate and worm F11 genes 4 are identical. Of the latter, 2 intron positions are F11-specific. The*C.elegans*L1 and F11 genes are very different in structure having only one exon/intron boundary at identical position.

The results of comparisons further support the accuracy of gene structure predictions generated by the GeneBuilder system and demonstrate that L1 and F11 genes have arisen from a common ancestor. Secondly, it may be concluded that the primordial L1/F11 gene had an exon/intron arrangement similar to that found in vertebrate genes (2 exons/1 domain). The conservation of this organization pattern for so long period suggest its strong positive selection in vertebrate lineage. Loss of almost a half of introns in the*C. elegans*Ig/FnIII-like genes, as well as in fruit fly genes of this family (not shown), demonstrate that such selection did not operate in the case of invertebrate species.

Finally, we conclude that there are at least several dozens of the IgSF genes in*C.elegans*genom. However, gene predictions generated by the Sequencing Consortium may be of limited value for tracing the relationships among*C.elegans*and vertebrate IgSF genes. In the absence of experimental data, this goal may be achieved by an individual analysis of the Ig-domain encoding regions using similarity searches for predictions of their exon/intron structure.



Figure 1. Schematic representation of the predicted exon/intron structures of the C.elegans (upper) L1- and F11-like genes in comparison with the human (lower) L1 and axonin-1 genes. Exons are depicted as filled boxes. Broken lines show borders of the Ig- and FnIII-like domains. Numbers indicate intron phase.

**References**

1. Burset M. and R. Guigo, (1996) Genomics 34: 353.
2. Doolittle R.F., (1995) Annu. Rev. Biochem. 64: 287
3. Gelfand M.S., A.A. Mironov and P.A. Pevzner, (1996) Proc. Natl. Acad. Sci. USA 93: 9061.
4. Gish W. and States D.J., (1993) Nat.Genet. 3: 266
5. Milanesi L. and I.B. Rogozin, (1998) In: "Guide to Human Genome Computing (2nd ed.)" (Ed. M.J.Bishop), Academic Press, Cambridge, 1998.
6. Rogozin I.B., L. Milanesi and N.A. Kolchanov, (1996).Comput. Applic. Biosci. 12: 161.
7. Walsh F.S. and P. Doherty, (1997) Annu. Rev. Cell. Dev. Bio. 13: 425
8. Williams A.F. and A.N.Barklay, (1988) Annu. Rev. Immunol. 6: 381