Bioinformatique: Analyse d'article

Daniel Gautheret
ESIL, Université de la Méditerranée

V.2004.1 http://www.esil.univ-mrs.fr/~dgaut/Cours

Quizz

- ★ Après avoir lu l'article « *Computational and experimental analysis reveals a novel Src family kinase in the C.elegans genome* », répondez aux questions suivantes:
  1. Quelles sont les protéines Src connues dans le génome de *C. elegans* au début de l'étude?
  2. Que découvrent les auteurs à l'issue de leur premier Tblastn?
  3. Dans la partie « Materiel et Méthodes » le crible avec un fragment de 800bp du virus du sarcome de Rous sert à récupérer des cDNA proches de v-src. A quoi servent ces cDNA (un d'entre eux en fait) dans la suite de l'article?
  4. Quels arguments sont utilisés pour réannoter le gène de la protéine hypothétique Y92H12A.1?
  5. Que s'avère être Y92H12A.1 en fait?
  6. Pourquoi voudrait-on faire en plus une analyse phylogénétique ici?
  7. Y a-t-il des branchements incertains dans l'arbre?
  8. Les auteurs ne concluent pas sur l'origine des src de *C.elegans* et de Drosophile: aidez-les à proposer quelque chose à partir de cet arbre!

Les réponses seront données au dernier cours du ½ module

BIOINFORMATICS DISCOVERY NOTE

Vol. 19 no. 2 2003
Pages 169–172

Computational and experimental analysis reveals a novel Src family kinase in the C.elegans genome

Akhilesh Pandey[1,*], Suraj Peri[1], Colin Thacker[2], Chery A. Whipple[3], John J. Collins[4] and Matthias Mann[1]

[1]Center for Experimental Bioinformatics, University of Southern Denmark, Odense M, DK-5230, Denmark, [2]Biotechnology Laboratory, University of British Columbia, Vancouver, B.C. V6T 1Z4, Canada, [3]The Genetics Program and [4]Department of Biochemistry and Molecular Biology, University of New Hampshire, Durham, NH 03824, USA

Received on May 8, 2002; revised on August 26, 2002; accepted on August 28, 2002

ABSTRACT

Motivation: The complete genomes of a number of organisms have already been sequenced. However, the vast majority of annotated genes are derived by gene prediction methods. It is important to not only validate the predicted coding regions but also to identify genes that may have been missed by these programs.
Methods: We searched the entire *C.elegans* genomic sequence database maintained by the Sanger Center using human c-Src sequence in a TBLASN search. We have confirmed one of the predicted regions by isolation of a cDNA and carried out a phylogenetic analysis of Src kinase family members in the worm, fly and several vertebrate species.
Results: Our analysis identified a novel tyrosine kinase in the *C.elegans* genome that contains functional features typical of the Src family kinases that we have designated as Src-1. The open reading frame contains a conserved N-terminal myristoylation site and a tyrosine residue within the C-terminus that is crucial for regulating the activity of Src kinases. Our phylogenetic analysis of Src family members from *C.elegans*, *Drosophila* and other higher organisms revealed a relationship among Src kinases from *C.elegans* and *Drosophila*.
Contact: pandey@cebi.sdu.dk

INTRODUCTION

c-Src was the first cellular homolog to be discovered for any oncogene (Stehelin *et al.*, 1976). Over the years, a number of proteins have been discovered that share the domain structure with the prototypical member, Src. The Src family of kinases currently includes nine members, Src, Yes, Fgr, Lck, Fyn, Lyn, Hck, Blk and Yrk. Based on phylogenetic analyses, they have been classified into two groups: SrcA and SrcB, where SrcA includes Src, Fgr, Fyn, Yrk and Yes and SrcB includes Lyn, Hck, Blk and Lck (Hughes, 1996).

With the completion of sequencing of genomes of several organisms, it is often expected that most proteins with significant similarity to important molecules in other species have already been identified and annotated. However, annotation and other errors in genomes can occur due to a number of reasons (Galperin and Koonin, 1998). For instance, a distinct signal transducer and activation of transcription (STAT) gene was identified in *Caenorhabditis elegans* (*C.elegans*) when a detailed analysis was carried out to specifically look for STAT orthologs in the worm (Liu *et al.*, 1999).

The sequencing of the 97-megabase genome of *C.elegans* was completed almost three years ago by the *C.elegans* sequencing consortium which revealed the presence of 19 099 genes (The *C.elegans* sequencing consortium, 1998). A previous study indicated that only a single representative of the Src family was present in *C.elegans* (Plowman *et al.*, 1999) although two Src-related molecules have been described in *Drosophila* (Dsrc64 and Dsrc41) (Simon *et al.*, 1985; Takahashi *et al.*, 1996). Since this protein was more related to Fyn than to Src, it was referred to as a Fyn-like kinase. We therefore decided to take another look at the *C.elegans* genome to determine the final number of Src family members.

METHODS

RNA isolation

Six *Caenorhabditis elegans* worms (Bristol strain N2) were picked in 5 μl of RNase-free distilled water to

[*]To whom correspondence should be addressed.

which 50 μl of GITC buffer (4 M guanidinium thiocyanate, 50 mM Tris-HC1 pH 7.4, 50 mM EDTA, 1% sarkosyl) was added. The samples were incubated in a dry ice/ethanol bath for 15 min. After thawing for 1 min at 65 °C, the samples were lyophilized. The dried pellet was resuspended in distilled water and extracted with phenol and chloroform.

cDNA isolation

An amplified cDNA library was made from poly A RNA extracted from a pool of *C.elegans* worms (Bristol strain N2) at different developmental stages. The library was screened with an 800 bp fragment of the Rous sarcoma virus, Schmidt-Ruppin strain. The fragment was kindly provided by J. Michael Bishop. Positive plaques were identified by hybridization performed at low stringency and sequenced.

Sequence analysis

The *C.elegans* genomic sequence database maintained by the Sanger Center (http://www.sanger.ac.uk) was used to search for regions encoding genes homologous to c-Src sequences using TBLASTN algorithm (Wu-BLAST 2.0). This algorithm compares a query protein sequence with nucleotide database translated in all six frames. The BLAST output was analyzed to identify orthologs based on *p*-values. The sequence identities between Src and Fyn members were determined using the BLAST2 algorithm (Tatusova and Madden, 1996).

The exons of the predicted hypothetical protein (Y92H12A.1; GenBank accession # AAK29735) were corrected using homology-based alignments based on publicly available EST sequences from *C.elegans*. These exons were further confirmed by a cloned cDNA and also supported by several ESTs corresponding to this cDNA.

Domain and phylogenetic analysis

We used SMART (Schultz *et al.*, 1998) and Pfam (Sonnhammer *et al.*, 1998) domain analysis tools for prediction of domain structures. Multiple sequence alignment was performed using ClustalW version 1.74 (Thompson *et al.*, 1994). A phylogenetic tree was constructed using the neighbor-joining algorithm as described (Saitou and Nei, 1987). The statistical significance of the branches was estimated using 500 bootstrap analyses. Trees produced by phylogenetic analyses were viewed with TreeView program (Page, 1996).

RESULTS

We searched the entire *C.elegans* genomic sequence database maintained by the Sanger Center using TBLASTN search algorithm. When human c-Src sequence was used to search this database, we found that the entry with the lowest E-value ($7.0e^{-119}$) was annotated as encoding a hypothetical protein (Y92H12A.1; GenBank accession # AAK29735) from cosmid clone Y92H12A. The entry with the next lowest score (E-value of $2.7e^{-84}$) was annotated as Fyn-like kinase. There were also several other entries with significant scores—these included tyrosine kinases Abl-1, vab-1, let-23 (*C.elegans* EGF receptor), DAF-2 (*C.elegans* insulin receptor homolog), Csk and Fak as well as several members of Fes/Fer family of SH2-containing tyrosine kinases. We examined the genomic region encoding the hypothetical protein labeled as Y92H12A.1 and performed comparisons with ESTs as well as homology-based alignments. Finally, we isolated a cDNA clone (accession # AF419171) to confirm our intron exon assignments based on homology. Figure 1a shows a comparison of the predicted gene with that of the experimentally obtained cDNA clone. The predicted exon IV and part of exon VII as indicated by the blue colored boxes in the figure were not found in the cDNA. These regions are contained within the coding region and do not encode sequences homologous to the Src-family kinases and therefore represent false positive predictions. Conversely, we found five exons or regions within exons in the cDNA that were absent in the predicted gene (false negatives) (indicated in red in Figure 1a). The availability of the cDNA and the knowledge of the domain structure of this well-studied gene family allowed us to refine the computer prediction. Similar efforts can easily be carried out to annotate other gene families as well.

DOMAIN AND SEQUENCE ANALYSIS

Domain analysis of this corrected protein showed that it contained an SH3 and an SH2 domain followed by a tyrosine kinase domain—an arrangement that is typical of Src family members (Pawson and Scott, 1997). Our analysis reveals that there are two Src-like kinases in *C.elegans* that are clearly distinct from each other. We therefore propose that the novel member of Src family that we have identified be designated as Src-1 and that the other known Fyn-like kinase be termed as Src-2.

The kinase activity as well as transforming potential of Src proteins is activated by dephosphorylation of a crucial tyrosine residue, Y527, which is conserved in the C-terminus of all Src family members (Superti-Furga and Courtneidge, 1995). A multiple alignment of Src-1 with *Drosophila* and human Src and Fyn sequences shows conservation of this tyrosine residue (Figure 1b). A second distinguishing feature of Src family kinases is their membrane localization that is dictated by the presence of a myristoylation site as well as charged residues within the N-terminal region (Resh, 1994). Membrane localization is an important event that brings these proteins in close

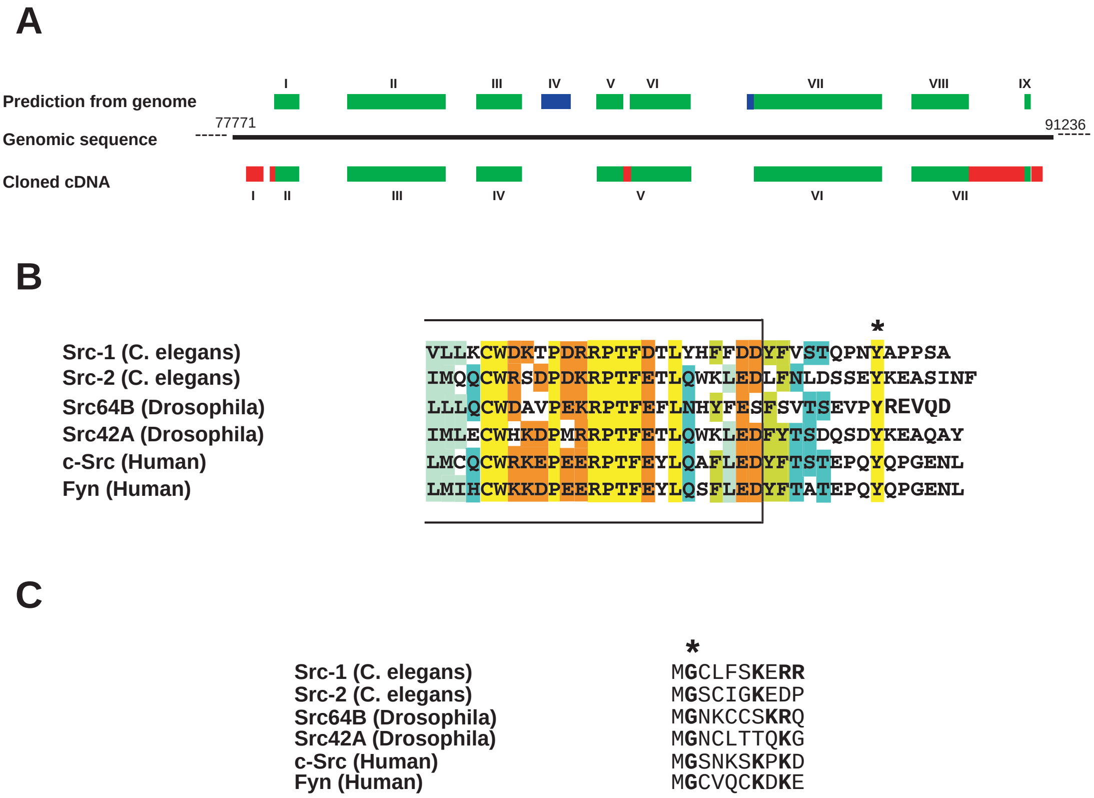

Fig. 1. (a) Genomic structure and sequence analysis of a novel Src family kinase in *C.elegans*, Src-1. Comparison of predicted exons from genomic DNA with a cloned cDNA. Full length cDNA sequence compared to the predicted transcript from a cosmid clone, Y92H12A (GenBank accession #AAK29735), shows discrepancies. Regions marked in red color indicate exons that were not predicted using the prediction program (false negatives) and exons colored in blue indicate exons that are not found in the cDNA (false positives). Exons marked in green color show authentic exons (i.e. common to both predicted and cDNA). (b) Conservation of a C-terminal tyrosine residue and a putative myristoylation site in Src-1. Multiple alignment of the C-terminal region of the novel Src-like kinase (Src-1) with Fyn-like kinase from *C.elegans* (Src-2), two Src family kinases from *Drosophila melanogaster*, and Src and Fyn from *Homo sapiens*. The amino acid residues are colored on the basis of conservation of similar residues. The conserved tyrosine found in the C terminus of all Src family members corresponding to Y527 of mouse Src is indicated by an asterisk. The box shown in the figure marks the end of the kinase domain. (c) Conservation of a myristoylation site in Src-1. The figure shows the presence of a conserved myristoylation site (indicated by the asterisk) in the N-terminal part of Src-1 as well as other Src and Fyn like kinases as described in panel B. Charged residues required for membrane attachment are shown in bold.

proximity to receptors residing in the plasma membrane. An alignment of Src-1 and Src-2 with prototypical Src-family kinases from other species shows the presence of these two key features (Figure 1c). Therefore the novel Src-like kinase, Src-1, in *C.elegans* shares two key functional features besides sequence similarity to the Src family members.

PHYLOGENETIC ANALYSIS

To understand the phylogenetic relationship among these Src-family kinases, we aligned the known Src kinases derived from *C.elegans* to *Homo sapiens*. The resulting alignment was used to construct a phylogenetic tree using the neighbor-joining tree algorithm. The statistical significance of various branches was estimated using the bootstrap method (Felsenstein, 1985). The phylogenetic tree of Src family members shows three clades whose branches are supported by high bootstrap values (Figure 2). Two of these clades represent SrcA and SrcB members. Interestingly, the third clade contains only the Src related kinases from *C.elegans* and *Drosophila*. The phylogenetic tree shows that Src-1 and Src-2 in *C.elegans* are related to Dsrc64 and Dsrc41, respectively, in *Drosophila melanogaster*. More detailed analysis will be necessary to study the observed relationship and to trace back the point of divergence among these Src-family kinases.

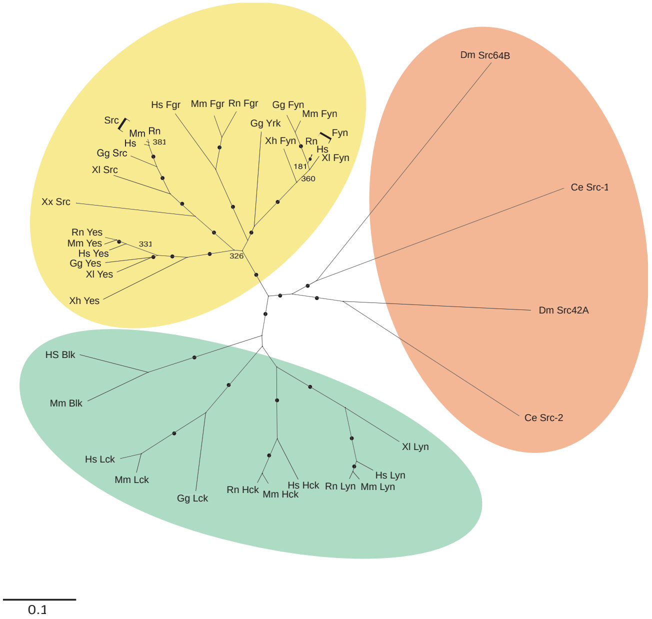

Fig. 2. Phylogenetic relationship among Src family tyrosine kinases. Sequences of Src family kinase members from *C.elegans* (Ce), *Drosophila melanogaster* (Dm), *Xiphophorus helleri* (Xh), *Xiphophorus xiphidium* (Xx), *Xenopus laevis* (Xl), *Gallus gallus* (Gg)(chicken), *Mus musculus* (Mm), *Rattus norwegicus* (Rn) and *Homo sapiens* (Hs) were used for multiple alignment. ClustalW program was used for multiple alignment and a phylogenetic tree construction using the neighbor-joining tree algorithm. The branch lengths are to scale and the scale units are a measure divergence and there is no correction for multiple substitutions. The statistical significance of various branches was estimated using the bootstrap method. The dots in the figure represent bootstrap values greater than 85 percent for 500 trials with the actual values shown if the value was found to be less than 85 percent. The three clades described in the text are enclosed in yellow, green or salmon colored circles.

DISCUSSION

We have identified a novel member of the Src family in *C.elegans* almost three years after the complete sequencing of the worm genome which is about 40 times smaller in its overall size as compared to the human genome. Approximately 32 000 to 39 000 genes have already been predicted as a result of the complete sequencing of the human genome (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001). Our study suggests that automated predictions coupled with manual curation of genes (especially gene families) is a good approach to annotate all protein coding regions from the human genome and will require a collective and concerted action from computational as well as experimental biologists.

REFERENCES

Felsenstein,J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

Galperin,M.Y. and Koonin,E.V. (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol.*, **1**, 55–67.

Hughes,A.L. (1996) Evolution of the src-related protein tyrosine kinases. *J. Mol. Evol.*, **42**, 247–256.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Liu,X. *et al.* (1999) STAT Genes Found in *C.elegans*. *Science*, 285–167a.

Page,R.D. (1996) Tree View: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.*, **12**, 12357–358.

Pawson,R. and Scott,J.D. (1997) Signaling through scaffold, anchoring, and adaptor proteins. *Science*, **278**, 2075–2080.

Plowman,G.D. *et al.* (1999) The protein kinases of *Caenorhabditis elegans*: a model for signal transduction in multicellular organisms. *Proc. Natl Acad. Sci. USA*, **96**, 13603–13610.

Resh,M.D. (1994) Myristylation and palmitylation of Src family members: the fats of the matter. *Cell*, **76**, 411–413.

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

Schultz,J. *et al.* (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA*, **95**, 5857–5864.

Simon,M.A. *et al.* (1985) The nucleotide sequence and the tissue-specific expression of *Drosophila* c-src. *Cell*, **42**, 831–840.

Sonnhammer,E.L. *et al.* (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.*, **26**, 320–322.

Stehelin,D. *et al.* (1976) DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature*, **260**, 170–173.

Superti-Furga,G. and Courtneidge,S.A. (1995) Structure-function relationships in Src family and related protein tyrosine Kinases. *Bioessays*, **17**, 321–330.

Takahashi,F. *et al.* (1996) Regulation of cell-cell contacts in developing *Drosophila* eyes by Dsrc41, a new, close relative of vertebrate c-src. *Genes Dev.*, **10**, 1645–1656.

Tatusova,T.A. and Madden,T.L. (1996) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.*, **174**, 247–250.

The *C.elegans* sequencing consortium (1998) Genome sequence of the nematode *C.elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.

Thompson,J.D. *et al.* (1994) ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Venter,J.C. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.