Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Computational and molecular characterization of multiple isoforms of *lfe-2* gene in nematode *C. elegans*

Luv Kashyap[1], Mohammad Tabish[1]*, Subramaniam Ganesh[2], and Deepti Dubey[2]

[1]Department of Biochemistry, Faculty of Life Sciences, Aligarh Muslim University, Aligarh, India; [2]Department of Biological Sciences and Bioengineering, Indian Institute of Technology, Kanpur, India; Mohammad Tabish - Email: tabish.biochem@gmail.com; Phone: +91-571-270-0741; * Corresponding author

received May 01, 2007; accepted May 02, 2007; published online May 30, 2007

Abstract:
C. elegans C46H11.4 gene encodes a Let-23 fertility effector/regulator protein of the EGF-receptor class of the tyrosine kinase family. Alternative splicing is a major mechanism of generating protein diversity in higher eukaryotes. *C. elegans* genome sequencing consortium has reported three alternatively spliced transcripts of C46H11.4 gene which encodes for three hypothetical proteins namely, C46H11.4a, C46H11.4b and C46H11.4c. Using a combination of various bioinformatics tools like gene or exon finding programmes, blast searches, alignment tools etc followed by experimental validation, we report the presence of three more alternatively spliced transcripts which encode for novel hypothetical proteins C46H11.4d, C46H11.4e and C46H11.4f. These isoforms arise as a result of alternative splicing in the pre-mRNA encoded by gene C46H11.4. These novel un-reported spliced variants not only point towards the extent of alternative splicing in *C. elegans* genes but also hint towards the complex nature of alternative splicing.

Keywords: *lfe-2*; Let-23 fertility effector/regulator protein; exon prediction; alternative splicing; *C. elegans*

Background:
Protein tyrosine kinases (PTKs) are important molecules for intra- and inter-cellular communication as well as for survival in eukaryotes, playing a major role in signal transduction processes. **[1]** There are two main classes of PTKs: Receptor Tyrosine Kinases (RTKs) or cellular, and Non-Receptor Tyrosine Kinases (NRTKs) or non cellular. **[2]** RTKs are transmembrane proteins that are involved in the control and regulation of several key cellular processes. To date, two groups of RTKs have been described in *C. elegans*: orthologs of RTKs known in mammals (e.g. DAF-2, EGL-15, KIN-8, LET-23 AND VAB-1) and RTKs identified only in *C. elegans* (e.g. KIN-15 and KIN-16). **[3]**

C. elegans C46H11.4 gene encodes *lfe-2*, a Let-23 fertility effector/regulator protein of the gene class LET-23, a member of the EGF-receptor tyrosine kinases family. *lfe-2* activity is required for normal levels of hermaphrodite fertility and for tissue-specific negative regulation of the LET-23 signaling pathway. **[4]** LET-23 is seen to play a pivotal role in free living nematode *C. elegans*. It is required for vulval induction as it helps the VPCs (vulval precursor cells) to respond to the signals that induces vulval differentiation. **[5]** It also signals through the RAS/MAPK pathway to specific cell fates in the ventral cord, vulva, male tail and excretory system and through a PLC/IP3 pathway to regulate ovulation. **[6]** Large numbers of spliced transcripts have been reported from various groups of RTKs and their associated families in almost all higher eukaryotes like humans, mouse and several other organisms including *C. elegans.* Alternative splicing is a powerful means of regulating gene expression and enhancing protein diversity. In fact, the majority of metazoan genes encode pre-mRNAs that are alternatively spliced to produce anywhere from two to tens of thousands of mRNA isoforms. Alternatively spliced exons are typically identified by aligning EST clusters to reference mRNAs or genomic DNA. **[7]** This approach has several limitations; firstly they may underestimate alternative splicing because of their incomplete coverage and lack of information regarding combinations of exons that are utilized. Secondly, it is hard to detect highly specific as well as low copy number splice variants by EST-based approaches **[8]** or microarray technologies. **[9, 10]** Most of the other approaches used work on the genomic level by predicting splice sites and introns/exons using sequence signals and composition differences. **[11, 12]** These methods are more devoted to gene prediction and usually yield only one optimal gene structure. Thus, methods that facilitate the identification of alternative exons would be quite useful to assist in genome annotation. Although several approaches for the ab initio prediction of gene structure have been developed, the ab initio prediction of alternative splicing

using a combination of gene/exon finding programmes has not been considered. Genefinder prediction by the *C. elegans* sequencing consortium of genomic sequence of C46H11.4 has reported three spliced variants C46H11.4a, C46H11.4b and C46H11.4c. In spliced transcript C46H11.4a, a new exon originates from the large intronic region between exon 5 and exon 6 of C46H11.4c (the biggest isoform of C46H11.4 gene) and splices with the 6th exon of the parent transcript C46H11.4c to form a new spliced variant (Figure 1). Similarly, a new exon originates from the intronic gap between exon 4 and exon 5 of C46H11.4c and splices with the 5th exon of the parent transcript C46H11.4c to form its new spliced transcript C46H11.4b (Figure 1)**.** Here, we report the detailed analysis data of the large intronic and 5' and 3' untranslated regions (5'-UTR and 3'-UTR) of C46H11.4 gene using various gene/exon finding programmes and several other bioinformatics tools. Computational analysis predicted the existence of three new spliced variants C46H11.4d, C46H11.4e and C46H11.4f which were subsequently confirmed by the presence of different corresponding transcripts by RT-PCR using gene specific primers and RNA isolated from mixed population of *C. elegans*.

Methodology:

Animals

The Bristol N2 wild-type strain of *C. elegans*, grown on NGM plates and maintained on *Escherichia coli* (OP50) essentially as described in **[13]** were used in all experiments.

Reagents

RevertAid™ M-MuLV Reverse Transcriptase and oligo $(dT)_{18}$ primer were purchased from Fermentas, Hanover, (USA), *Taq* DNA Polymerase and PCR-Buffer were purchased from Bangalore Genie Pvt. Ltd., India., dNTP Mix (2.5 mM each) and 1kb DNA ladder was purchased from MBI Fermentas, USA. All other chemicals used in the experiments were of molecular biology grade.

Primers

The following oligonucleotides primers were custom synthesized from MWG Biotech, Pvt. Ltd., India. The numbers in the bracket indicate the position of the primer sequence in the cosmid C46H11 (Accession No.U88314) and the size of the primers respectively.

1. F1 (7457-7478; 22 mer):
   5' TGCTAGAGGCGATTTACGCCAA 3'
2. F2 (8919-8943; 25 mer):
   5' GAACTACGCTCATTCTGCTAGTGAG 3'
3. F3 (7981-8002; 22 mer):
   5' CCTAGTCAGGAGCCTGTTGTAC 3'
4. F4 (4138-4164; 27 mer):
   5'GTCTCATGTCAAATGTGGAATGCAAGG 3'
5. R1 (16984-17005; 22 mer):
   5' CGTCGGGTACGGGAAGCAGCTG 3'

The direction and relative position of the primers are indicated in figure 1.

Preparation of RNA

Total RNA was isolated from mixed-stage nematodes using the method described earlier. **[14]** Finally, total RNA was dissolved in diethyl pyrocarbonate-treated distilled water. Purity of RNA was checked using denaturing agarose gel electrophoresis.

Reverse transcriptase (RT)-PCR

Computationally predicted spliced variants were validated using RT-PCR. *C. elegans* total RNA (3 micro gram) was primed with oligo $(dT)_{18}$, and single-stranded cDNA was synthesized using Revert Aid™ Reverse Transcriptase at 42°C for 1 h (total vol. 20 micro liter). Subsequently, 1 micro liter of the single-stranded cDNA preparation was added to PCR system buffer for PCR using appropriate sets of primers (total vol. 20 micro liter). PCR was performed for 30 cycles (denaturation, 94°C, 5 min, 1 cycle; denaturation, 94°C, 45 sec; annealing, 58°C, 1 min; extension, 72°C, 1 min; finally, 72°C for 7 min). RT-PCR product (5 micro liter) was subjected to electrophoresis on a 2% (w/v) agarose gel, stained with ethidium bromide and photographed on a UV-trans illuminator.

Bioinformatics tools

Genomic sequence of cosmid C46H11 (Accession No.U88314) was downloaded from the NCBI nucleotide database (http://www.ncbi.nlm.nih.gov/entrez). Various other databases used for searches were those at the *C. elegans* Sequencing Project Center at (http://sanger.ac.uk/Projects/C$_h$elegans/blast$_h$server.shtml) and Yuji Khoara's *C. elegans* expressed sequence tag (EST) database (http://ddbj.nig.ac.jp/htmls/c-elegans/html/CE_INDEX.html). Homology and similarity searches of the polypeptide and nucleotide sequences were performed using the BLASTN and BLASTP non-redundant database (http://www.ncbi.nlm.nih.gov/BLAST). Gene/Exon finding tools used were HMMgene (http://www.cbs.dtu.dk/services/HMMgene), Genescan (http:// genes.mit.edu/ GENSCAN.html), GeneSplicer (http://cbcb.umd.edu/ software/ GeneSplicer/), FGENESH (http://www.softberry.com/ berry.phtml). ORF findings tools used were located at (http://www.bioinformatics.vg/sms/orf_find.html) and (http://www.prl.msu.edu/clab/dnajava/orf_find.html). Alignment analysis was carried out using the Gene stream Align tool (http://www2.igh.cnrs.fr/bin/align-guess.cgi) and ClustalW tool available at (www.ebi.ac.uk/clustalw). Primers were designed manually or using the tool Primer3 (http://biotools.umassmed.edu/bioapps/primer3www.cgi). Various other tools and bioinformatics programmes used were same as described in our previous study/earlier. **[15]**

Results and discussion:

Computational analysis and prediction of new alternatively spliced transcripts of *lfe-2* encoding gene C46H11.4

Computational analysis of genomic DNA sequences using gene finders to predict spliced transcripts and subsequent confirmation of the predicted spliced variants have been successfully demonstrated in *C. elegans.* **[14]** Several studies aimed at detecting new spliced variants arising due to alternative splicing have been done in various RTKs and related families in Humans **[16, 17, 18]**, mouse **[19, 20]** etc. In light of these previous studies, we started with the detailed analysis of the 5' and 3' untranslated (UTR) and large intronic regions of C46H11.4 gene. Analysis comprised of a series of bioinformatics tools like gene/exon finding programmes, ORF finders and several other bioinformatics tools (as mentioned in materials and methods). These tools predicted several exons possibly capable of replacing the existing exon(s) and thus creating alternatively spliced variant of the gene. From the several exons predicted above, we selected only the "common exons" capable of replacing the existing exon(s) and thus creating spliced transcript of the gene without affecting the reading frame of the protein. Further, the possibility of occurrence of that spliced exon/transcript was analyzed by a comparative analysis between the original protein and the new protein formed by addition/exclusion of alternative exons using various alignment tools. Lastly, several other parameters like percent-amino acid replacement, codon usage, sense nature i.e. whether from positive or negative strand, the probability score of occurrence of that exon etc. were also checked to ensure accuracy of predicted spliced transcript of the gene. Thus, Gene/Exon finder analysis of about 1.6 kb of intronic gap region between exon 1 and exon 2 of existing C46H11.4c and about 4 kb of 5' untranslated region (between the predicted stop codon of the upstream gene C46H11.6 and the initiation codon of C46H11.4c) predicted three new exons. These new exons 1d, 1e and 1f encoded 16, 22 and 24 amino acids residues and have methionine as the initiation codon (Table 1 in supplementary material). These newly identified exons were potentially capable of splicing with exon 2 or exon 3 of the parent gene C46H11.4c (Figure 1, Table 1 in supplementary material) without affecting the reading frame of the protein.

Thus computational analysis of the large intronic and 5' and 3' untranslated (UTR) region predicted three new spliced transcripts of the gene C46H11.4, which we have named C46H11.4d, C46H11.4e and C46H11.4f. In all the three new spliced transcripts a new N-terminal exon in each splices with the already existing exon of the parent isoform in the manner as depicted (Figure 1). In C46H11.4d, a new potential exon originates from the 5'UTR region of C46H11.4c and splices with the 3rd exon of the parent transcript C46H11.4c. (Figure 1). Whereas, in spliced variant C46H11.4e, a new potential exon originates from the large intronic gap between exon 1 and exon 2 of C46H11.4c and splices with the 2nd exon of C46H11.4c (Figure 1). Similarly, a new potential exon originates from the large intronic gap between exon 1 and exon 2 of C46H11.4c and splices with the 3rd exon of C46H11.4c to form the new transcript C46H11.4f. The three new spliced transcripts C46H11.4d, C46H11.4e and C46H11.4f have a 16, 22 and 24 amino acids long first exon and have their new 5' start 3322 bp upstream, 473 bp downstream and 1438 bp downstream respectively with respect to the parent gene C46H11.4c (Table 1 in supplementary material). In all spliced transcripts, the last exons remain the same only the N-terminal (5') encoding exon varies depending upon the splicing pattern. Following the computational predictions of new spliced isoforms, Yuji Kohara's *C. elegans* EST database was searched for putative EST/cDNA support for possible occurrence of these new exons/transcripts. A search of Yuji Kohara's *C. elegans* EST database didn't yield any EST match for these new exons, the most probable reason for this being that the current available EST database for *C. elegans* is not adequately represented. Secondly, *C. elegans* sequencing consortium has also made prediction C46H11.4a which has no conclusive EST/cDNA support for its unique 1st exon present at the 5'end of the transcript. This also supports our hypothesis that only EST data is not good enough for detection of all possible alternatively spliced variants of a gene. NCBI BLAST search was accomplished for finding out homology of these new spliced variants; however, no significant information was available about the prospective similarity with other polypeptides. Due to non-availability of required information for supporting EST/cDNA matches for the newly predicted exons, presence of transcripts were validated by performing RT-PCR using total RNA isolated from the mixed population of *C. elegans*.

Experimental confirmation of computationally identified new spliced transcripts encoded by *lfe-2* gene C46H11.4

In order to confirm the possible existence of new transcripts of Let-23 fertility effector/regulator protein encoded by *lfe-2* gene as predicted by bioinformatics tools, RT-PCR amplification was performed. We selected the transcript encoding C46H11.4c as a control in our experiments. Total RNA, prepared from mixed-stage *C. elegans*, was reverse transcribed and the resulting single-stranded cDNA was PCR amplified (as detailed in methodology). Forward and Reverse gene specific primers corresponding to each of the newly predicted exons were designed using a combination of bioinformatics tools and manual methods. These Forward and Reverse primers, specific for a particular exon were successfully able to confirm the occurrence of the spliced transcript by giving a band of anticipated size (Figure 2) when the primer pair specific for that exon was PCR amplified.

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

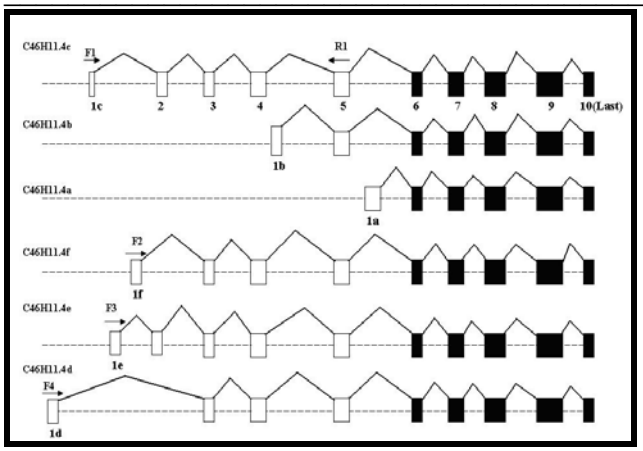

Figure 1: Organization of the C46H11.4 gene of *C. elegans* along with the predicted spliced variants: The Exon/Intron organization of the C46H11.4 gene, along with its existing spliced variants C46H11.4a, C46H11.4b and C46H11.4c and the newly predicted alternatively spliced variants C46H11.4d, C46H11.4e, C46H11.4f. Exons are indicated by rectangular boxes, dotted lines indicate the intronic and the untranslated regions, while solid joining lines show the splicing pattern of each spliced variant. Arrows (F1, F2, F3, F4 and R1) indicate the Primer designed specific for each predicted exon

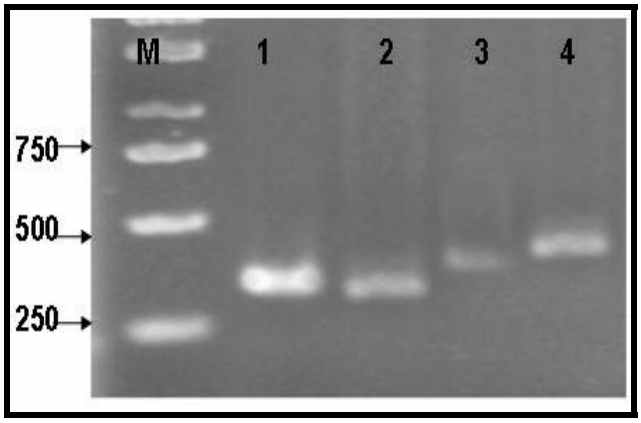

Figure 2: RT-PCR analysis of predicted spliced variant of the *lfe-2* gene C46H11.4: RT-PCR amplification was used to determine the presence of transcripts, containing predicted exon, in total RNA prepared from mixed-stage *C. elegans* as described in the Materials and methods section. The migration of a series of size markers (M) is indicated on the left. RT-PCR products were obtained using a common reverse primer 5R1 (from exon five) and exon specific forward primers representing each spliced variants. Lanes 1-4 represent the product size (bp) of 326, 316, 340 and 350 obtained using a common reverse primer in combination with forward primers F1, F2, F3 and F4 respectively. Presence of anticipated DNA band size in lanes 1 - 4 represent the spliced transcripts amplified which encode for C46H11.4c (control), and the computationally predicted C46H11.4f, 46H11.4e and C46H11.4d transcripts respectively

Our goal here was to use a novel bioinformatics method capable of delineating all possible spliced transcripts of a gene. Our results therefore, extend the findings of the previous studies that the conventional methods of detection of spliced variants of a gene are not good enough to detect all possible spliced isoforms. So for efficient delineation of all possible putative spliced variants of a gene and to minimize junk predictions (biological noise or false predictions), we propose to combine the ab initio prediction of alternative splicing using a combination of gene/exon finding programmes with experimental validation. These new coding sequences, not annotated or identified earlier will not only add to the available splice data base of *C. elegans* but will also enhance our knowledge about understanding of the genome structure and evolution of higher eukaryotes specially in context to humans. Moreover as now its well established that *C. elegans* are well characterized organisms that have a lot of biological functions that are similar to almost all higher organisms including humans, so based on our work, similar studies can be conducted in several other organisms specially humans with whom *C. elegans* share a close gene homology. Lastly, due to limited domain of our work, further studies using more

advanced techniques like the RNA interference (RNAi) could be taken up which would enhance our knowledge about the biological and functional significance of these spliced transcripts and their possible role in *C. elegans* gene working and regulation.

Acknowledgement:
The authors are thankful to the CSIR, New Delhi, India for providing financial support. Authors are also grateful to UGC and DST, New Delhi, India for providing special grants to the Department in the form of DRS and FIST for developing infrastructure facilities.

References:

[01] S. K. Hanks, *et al., Science*, 241:4861(1988) [PMID: 3291115]
[02] K. Neet and T. Hunter, *Genes Cells*, 1:2(1996) [PMID: 9140060]
[03] C. Popovici, *et al., Genome Res.*, 9:11(1999) [PMID: 10568743]
[04] E. J. Hubbard and D. Greenstein, *Developmental Dynamics*, 218:1(2000) [PMID: 10822256]
[05] R. V. Aroian & P.W. Sternberg, *Genetics*, 128:2(1991) [PMID: 2071015]
[06] T. R. Clandinin, *et al., Cell*, 92:4 (1998) [PMID: 9491893]
[07] S. Gupta *et al., BMC Genomics*, 5:1(2004) [PMID: 15453915]
[08] B. Modrek and C. Lee, *Nat. Genet.*, 30:1(2002) [PMID: 11753382]
[09] G. K. Hu, *et al., Genome Res.*, 11:7(2001) [PMID: 11435406]
[10] J. M. Yeakley, *et al., Nat. Biotechnol.*, 20:4(2002) [PMID: 11923840]
[11] C. Burge and S. Karlin, *J. Mol. Biol.*, 268:1(1997) [PMID: 9149143]
[12] M. G. Reese, *et al., Genome Res.*, 10:4(2000) [PMID: 10779493]
[13] S. Brenner, *Genetics*, 77:1(1974) [PMID: 4366476]
[14] M. Tabish, *et al., Biochem. J*, 339:1(1999) [PMID: 10085246]
[15] L. Kashyap and M. Tabish, *Bioinformation*, 1:203 (2006)
[16] C. C. Lee and K. M. Yamada, *J. Biol. Chem.*, 269:30(1994) [PMID: 7518457]
[17] M. H. Wang, *et al., Carcinogenesis*, 21:8(2000) [PMID: 10910951]
[18] E. Preger *et al., Proc. Natl. Acad. Sci.*, 101:5(2004) [PMID: 14742870]
[19] M. Lindahl, *et al., Mol. Cell Neurosci.*, 15:6(2000) [PMID: 10860579]
[20] K. Y. Lee, *et al., Biochim. Biophys. Acta*, 1627:1(2003) [PMID: 12759189]

Edited by P. Kangueane
Citation: Kashyap *et al.*, Bioinformation 2(1): 17-21 (2007)

Supplementary material

Exon	Amino acid sequence	N-terminal exon size	Splicing pattern
1c	MLEAIYAKIMGVGTSR	8	1c-2-3-4----9-Last(CN)
1d	MCLMSNVECKEVVNCNGSSNSE	16	1d-3-4------9- Last (PR-1)
1e	MSLSNRVRAHLVRSLLYIYTFPIMGVGTSR	22	1e-2-3-4---9- Last (PR-2)
1f	MNYAHSASEKKSENKTLKKKGHIAGSSNSE	24	1f-3-4------9- Last (PR-3)

Table 1: Amino acid sequence encoded by the predicted alternatively spliced N-terminal exons
Deduced amino acid sequences encoded by alternative first exons (shown in regular font) and the first part of the amino acid sequence encoded by exon 2 or exon 3 (underlined) depending on the splicing pattern; number of amino acids encoded by the N-terminal exon is shown under N-terminal exon size. 1c, 1d, 1e and 1f are the first exons of C46H11.4c (control), C46H11.4d (predicted), C46H11.4e (predicted) and C46H11.4f (predicted) respectively