Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Comparative analysis of various gene finders specific to *Caenorhabditis elegans* genome

Luv Kashyap and Mohammad Tabish*

Department of Biochemistry, Faculty of Life Sciences, A. M. University, Aligarh, U.P. 202002, India;
Mohammad Tabish* - Email: tabish.biochem@gmail.com; Phone: +91 571 2700741 and +91 98 371 08550;
* Corresponding author

received July 26, 2006; revised September 12, 2006; accepted October 02, 2006; published online October 07, 2006

Abstract:

Computational gene prediction and identifying alternatively spliced isoforms have always been a challenging task. In this paper, we describe the performance of three gene/exon finding programmes namely Fex, Gen view2 and Gene builder capable of predicting open reading frames or exons for a given set of sequences from *C. elegans* genome. The predicted exons were compared with the 'sequencing consortium' identified exons and degree of consensus among them is discussed. We found that exon prediction by Fex was similar to the consortium prediction as compared to Gen view2 and Gene builder results. Interestingly, some exons (six exons in five genes) predicted positive only by Fex and not by the 'sequencing consortium' are found at the *C. elegans* EST database. This data is critical for further debate and discussion on gene finding in *C. elegans*.

Key words: gene prediction; multiple transcripts; exons; isoforms; *C. elegans*

Background:

Locating and finding coding regions from a given set of sequence is gene prediction and it is a challenging problem in post-genome era. **[1]** The task is to decipher the meaningful messages or information coded in these sequences. Several tools such as Gen view, GeneMark.HMM, Gene builder, Geneid and many others are available. **[2-5]** However, no single gene prediction algorithm has been able to predict all the genes with high accuracy.

A number of gene finders have been used to identify new transcripts encoded by a single gene. This arises as a result of alternative splicing of new exons transcribed either from untranslated sequences or from intronic regions to the internal exons of the transcripts. Alternative splicing has been reported from highly divergent organisms from yeast to human including *Caenorhabditis elegans.* **[6]** *C. elegans* is the first organism to have its genome completely sequenced and is about 97Mb with approximately 20,000 genes scattered in six chromosomes. **[7]** It is well suited for genetic studies and in recent years much work has been done on it taking it as a model system because of its gene homology with humans. Over the past few years several genes have been reported from *C. elegans* which give rise to multiple transcripts due to different splicing mechanisms either at the 3' and/or 5' region of the transcript. **[6]** Thus splicing gives rise to transcripts encoding either similar or different proteins having related or may be entirely different functions. **[8]** Everyday new genes are being identified and listed in the database. Several gene-finding programmes have been compared and evaluated earlier to elucidate their predictions. Although *C. elegans* sequencing consortium has annotated the whole genomic sequence of *C. elegans* but further analysis using different gene finders may possibly identify large number of protein coding sequences not available in the database. Related analyses have been performed earlier **[9-13]** for various organisms other than *C. elegans*. Here we describe the comparison of exon prediction by Fex, Gen view2, Gene builder and with that 'sequencing consortium' and discuss the consensus among them.

Methodology:

Sequence data set

We have downloaded the genomic sequences of 120 hypothetical genes from http://www.sanger.ac.uk/Projects/C_elegans/Genomic_Sequence.shtml. *C. elegans* Sequencing Consortium has predicted the whole genome of *C. elegans* using the genefinder for which functional parameters are not accessible to us. All the selected genes have exons ranging from one to six. These genes were selected from all six chromosomes of *C. elegans*, taking twenty genes from each chromosome. Thus, our data consisted of 120 genes containing 348 exons from the whole genome scattered in all six chromosomes.

Prediction programmes

We selected 3 gene finders namely FEX **[14, 15]**, Gene Builder **[4]** and GeneView2 **[2]** for the comparative study. We selected these programs satisfying the following criteria: [1] free availability; **[2]** easy interface; **[3]** exon prediction with protein sequence. All programmes were suitable to predict genes/exons according to the organisms selected and could be used for human, mouse, Drosophila and *C. elegans* sequences.

FEX

It is an exon prediction programme with a web interface available at http://www.softberry.com/berry.phtml. The FEX (Find EXon) **[14, 15]** programme predicts internal exons by linear discriminant function, evaluating open reading frames flanked by GT and AG base pairs (the 5' and 3' ends of typical introns). Potential 5'- and 3'- exons are predicted by corresponding discriminant functions on the left side of the first internal exon and on the right side from last internal exon, respectively.

GENE BUILDER

It is also a gene finding programme with web-based interface at http://l25.itba.mi.cnr.it/~webgene/genebuilder.html. **[4]** Gene Builder is based on prediction of functional signals

and coding regions by different approaches in combination with similarity searches in proteins and EST databases.

GEN VIEW2

It is available at http://l25.itba.mi.cnr.it/~webgene/wwwgene.html. GenView2 system is based on prediction of splice signals by classification approach and coding regions by dicodon statistic. **[2]** Potential gene structure is constructed using dynamic programming approach.

Other bioinformatics tools:

DNA Tools

In order to format the genomic sequences compatible for each programme, we used a web interface dna tool, available freely at http://biology.semo.edu/cgi-bin/dnatools.pl.

***C. elegans* EST**

To verify the existence of new exon(s) in the transcripts predicted by gene finders, *C. elegans* EST database was searched at http://www.sanger.ac.uk/cgi-bin/blast/submitblast/c_elegans.

TBLASTN

Newly predicted exon sequences were used to search EST database by BLASTN or TBLASTN **[16]** to identify any corresponding cDNA sequence present at http://www.sanger.ac.uk/cgi-bin/blast/submitblast/c_elegans. The cDNA sequences identified from the database with high percentage of identity with new exons were further analysed to rule out for any genomic DNA contaminations.

The evaluation methodology

To compare the prediction of gene/exons by Fex, Gene Builder and Gen View2, evaluation of these programmes was made on the basis of their ability not only to correctly predict the number of exons but also accurately predict the annotated proteins when matched to the individual amino acid level. Thus the accuracy was measured at the level of exon and amino acid.

Our evaluation methodology is based on the terms defined earlier **[9, 13]**, briefly these are explained below:

Actual exons: Total number of exons predicted for the respective genes by *C. elegans* Sequencing Consortium and submitted to the database.

Predicted exons: Total number of exons predicted in this study by the exon/gene finder when genomic DNA sequence was analysed

Positives: Total number of exons predicted by the exon/gene finder having exactly the same amino acid sequence as that of the exon predicted and submitted to the database by *C. elegans* Sequencing Consortium.

Negatives: Total number of exons predicted by the exon/gene finder having different or missing amino acid sequence from that of the exons predicted and submitted to the database by *C. elegans* Sequencing Consortium.

New exons: Total number of exons which were new prediction in this study and were neither identified nor predicted earlier by *C. elegans* Sequencing Consortium.

Degree of consensus: Percentage similarity between amino acid sequences encoded by the exons predicted by exon/gene finding programmes with that already predicted actual exons by *C. elegans* Sequence Consortium.

Results and discussion:

Relative accuracy in exon predictions

With the development of genome sequence for many organisms, more and more raw sequences need to be annotated correctly. To predict the protein coding exons and genes, a number of programmes have been developed. **[17, 18]** Selecting the best gene and/or exon predicting programme claiming to be highly specific and accurate for the organism of interest could be difficult. Secondly, the programmes used in exon identification are generally tuned for specific organism. Although lot of work has been done in this direction to identify the best exon predicting programme for different organisms but yet incomplete information is available for most of the organisms including nematode *C. elegans*. Our approach was to identify the best exon prediction tool available among the several possible programmes which may be the best suited for gene/exon predictions for *C. elegans*. We selected the programs satisfying the following criteria: **[1]** free availability; **[2]** easy interface; **[3]** exon prediction with protein sequence and **[4]** well suited for several organisms including *C. elegans*. The three programmes we selected which fulfilled the above criteria were Fex, Gene builder and Gen view2 from among several others.

In order to perform the analysis, we selected 120 genes scattered in all six chromosomes of *C. elegans* from the *C. elegans* database. *C. elegans* dataset itself is predicted by Sequencing Consortium using certain genefinder for which details are not available. All genes selected for this study had number of exons ranging from one to six (Fig.1). For gene/exon predictions, the genomic sequence containing upstream untranslated region and complete gene including exons and intron as defined by *C. elegans* Sequencing Consortium was selected for each gene. A comparative performance of the exon/gene predicting programmes including various parameters have been clearly demonstrated in table 1. As apparent from table 1, total number of exons predicted by Fex was 384 against 348 exons predicted by *C. elegans* Sequencing Consortium. Gene Builder and Gen View2 predicted only 149 and 72 exons respectively. Fex prediction has outperformed the other selected exon prediction programmes with the highest number of positives 283 followed by Gene builder 77 and lastly Gen view2 55 out of 348 exons. In comparison with Gene Builder and Gen View2 which predicted large number of negative exons 162 and 230 respectively, Fex had only 63 negative exons with respect to the total number of exons established by *C. elegans* Sequencing Consortium.

To understand the accuracy of these gene finders, a term “degree of consensus” was defined, which is a measure of degree of similarity between amino acid sequence encoded by the exon predicted by exon/gene finding programmes and that already predicted and reported in the *C. elegans* sequence dataset by *C. elegans* Sequencing Consortium. It was very remarkable that, exons predicted by Fex were found to have

very high degree of consensus with *C. elegans* Sequencing Consortium predicted dataset as compared to the exons predicted by Gene Builder and Gen View2 (Table 1). Gen Builder successfully predicted exons with a considerable degree of consensus whereas the performance of Gen View2 was relatively poor. Gen View2 was found to have the least degree of consensus with *C. elegans* Sequencing Consortium predicted dataset among the three prediction programmes.

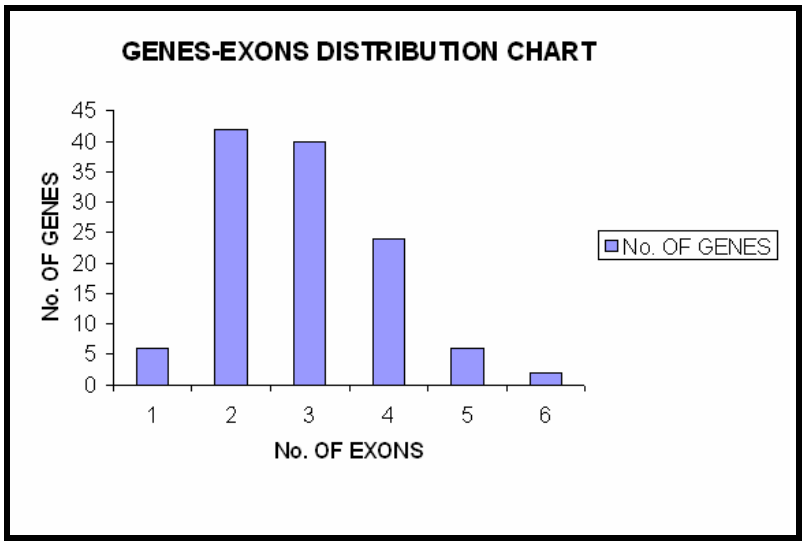

Figure 1: A plot between number of exons and number of genes. A chart showing number of genes having varying number of exons ranging from one to six. Genes containing one (6), two (42), three (40), four (24), five (6) and six (2) exons are shown.

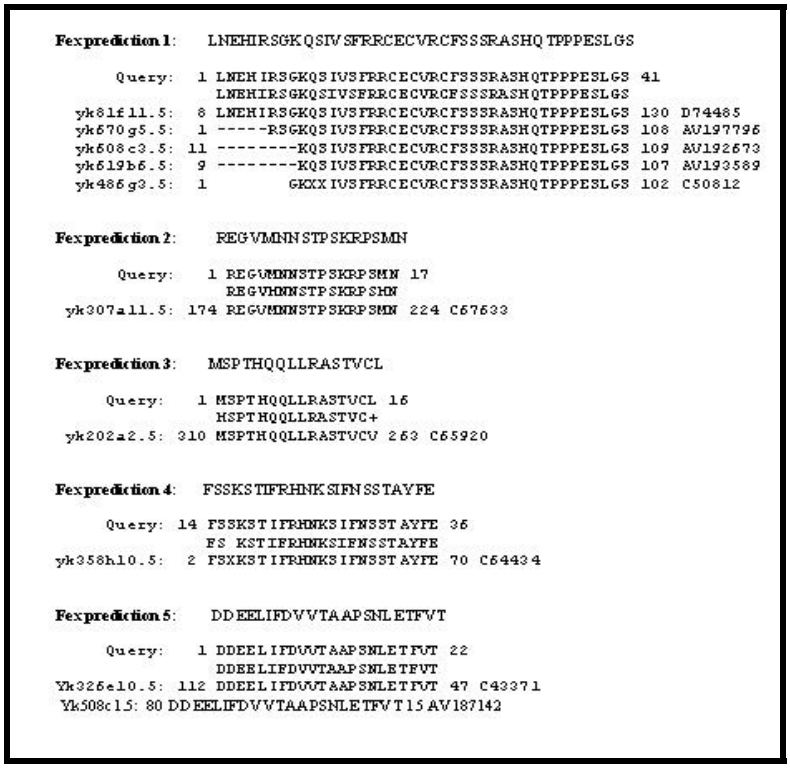

Figure 2: Homology of the Fex predicted exons with cDNA sequences of *C. elegans* EST database. Sequences were compared by TBLASTN computer-based sequence analysis. **[16]** Hypothetical gene name designated by *C. elegans* sequencing consortium is given in bold letters and amino acid sequence predicted by Fex is given next to the hypothetical gene name in bracket. Newly predicted exons were compared with cDNA sequences separately. Numbers on the left and right side of amino acid sequences indicate the position of these amino acid residues in the exons predicted by Fex (query) and cDNA hits (yk series) during TBLASTN search. Names of each cDNA clone are mentioned on the left side and their EMBL accession numbers are given on the right side of the aligned sequences.

Parameters	FEX	GEN VIEW2	GENE BUILDER
Actual exons	348	348	348
Predicted exons	384	72	149
Positives	283	55	77
Negatives	63	230	162
New exons	101	17	72
Degree of consensus	0.73	0.76	0.51

Table 1: Comparison of various parameters for three exon predicting programmes: A comparative view of essential parameters including positives, negatives and degree of consensus for three gene/exon predicting programmes to ensure the accuracy in their predictions

We have also found a large number of new exons prediction in this study, maximum by Fex (101) followed by Gene Builder (72) and lastly Gen View2 (17), which did not match the already established exons. New exons prediction was of special interest because these could be very important and may possibly be involved in generating new transcripts not identified or predicted by *C. elegans* Sequencing Consortium. Since the *C. elegans* genome data is itself predicted, it is likely possible that the actual number of genes present in the genome may be underestimated. Negative prediction was least for Fex and was highest for Gen View2, indicated that Fex could predict most of the exons predicted by *C. elegans* Sequencing Consortium to very high accuracy. Thus, with these analyses, we have found that the exon prediction by Fex was better than Gen Builder and Gen View2 for *C. elegans* genome with a very high degree of consensus with that of the sequences predicted by *C. elegans* Sequencing Consortium.

Identification of ESTs corresponding to new exons predicted by Fex

The completion of *C. elegans* genome sequencing project and the rapid increase in the size of Expressed Sequence Tag (EST) databases has lead to the discovery of multiple transcripts including alternatively spliced transcripts. Recently, many genes having multiple transcripts have been shown to exist in the nematode *C. elegans*. While analyzing genome sequence using any gene finder, prediction of multiple transcripts is particularly problematic. However, it was demonstrated clearly that the transcripts predicted for any gene using bioinformatics tools could be correctly identified and characterized. **[6]**

In order to verify the existence of new exons predicted by Fex, Gen Builder and Gen View2 and not by *C. elegans* Sequencing Consortium, we searched *C. elegans* EST database at http://www.sanger.ac.uk/cgi-bin/blast/submitblast/c_elegans. This search was performed at the level of nucleotide by BLASTN or at the level of amino acid sequences after translating the database sequences by TBLASTN. *C. elegans* EST database represents only about 1% of the total expressed genes. The new exons predicted by Gen View2 and Gene Builder failed to identify any corresponding cDNA sequence in EST database suggested that either the exon prediction was not correct or due to incomplete EST dataset. However, we have identified several ESTs namely yk81f11.5, yk670g5.5, yk608c3.5, yk619b6.5, yk486g3.5, yk307a11.5, yk202a2.5, yk358h10.5, yk326e10.5 and yk508c1.5 from *C. elegans* EST database corresponding to the new exons predicted/identified only FEX (Figure 2). The EST sequences identified (using new exons predicted by Fex) have very high identity at the levels of nucleotide (data not shown) and conceptual translated amino acids sequences (Fig 2). These new exons were neither identified nor annotated/predicted earlier in any transcript. Presence of new exons containing cDNA in EST database identified the presence of new transcripts encoded by *C. elegans* genome. Thus, identification of new transcripts using combined approaches i.e. Fex prediction and EST dataset searches will certainly increase the number of expressed genes in *C. elegans* dataset.

Conclusion:

In the present study, we have found that Fex was capable of predicting/identifying new exons not identified or detected earlier. These new exons were confirmed by identifying cDNA clones (from EST database) having such new exons, confirming new transcripts. Therefore, in addition to the genefinder used by *C. elegans* Sequencing Consortium, Fex prediction should also be considered for identifying the transcripts that arise from any gene. So, further studies using these programmes including Fex will help us in identifying novel and/or rarely represented transcripts not identified or characterized by other gene finders.

Acknowledgement:

The authors are thankful to the Council for Scientific and Industrial Research, New Delhi, India for providing financial support. Authors are also grateful to University Grants Commission and Department of Science and Technology, New Delhi, India for providing special grants to the Department in the form of DRS and FIST for developing infrastructure facilities. Authors are thankful to the anonymous referee for improving this manuscript.

References:

[01] R. J. Mural, *Methods Enzymol.*, 303:77 (1999) [PMID: 10349639]

[02] M. S. Gelfand & M. A. Roytberg, *BioSystems,* 30:1 (1993). [PMID: 8374074]

[03] A. V. Lukashin & M. Borodovsky, *Nucleic Acids Res.,* 26:4 (1998) [PMID:9461475]

[04] L. D. Milanesi, *et al., Bioinformatics,* 15:7 (1999) [PMID: 10487869]

[05] S. Parra, *et al., Genome Res.,* 10:4 (2000) [PMID: 10779490]

[06] M. Tabish, *et al., Biochem J.,* 339:3 (1999) [PMID: 10085246]

[07] *C. elegans* Sequencing Consortium, *Science,* 282: 5396 (1998) [PMID: 9851916]

[08] *A.* Alfonso, *et al., J. Mol. Biol.,* 241:4 (1994) [PMID: 8057385]
[09] M. Burset & A. R. Guig, *Genomics.,* 34:3 (1996) [PMID: 8786136]
[10] C. Burge & S. Karlin, *Curr. Opin. Struct. Biol.,* 8:3 (1998) [PMID: 9666331]
[11] R. Guigo, *et al., Genome Res.,* 10:10 (2000) [PMID: 11042160]
[12] E. Kraemer, *et al., Bioinformatics,* 17:10 (2001) [PMID: 11673234]
[13] S. Rogic, *et al., Genome Res.,* 11:5 (2001) [PMID: 11337477]
[14] V. V. Solovyev, *et al., Nucleic Acids Res.,* 22:24 (1994) [PMID: 7816600]
[15] C. M. Johnston, *et al., J. Immunol.,* 176:7 (2006) [PMID: 16547259]
[16] S. F. Altschul, *et al., J. Mol. Biol.,* 215:3 (1990) [PMID:2231712]
[17] I. Korf, *BMC Bioinformatics,* 5:59 (2004) [PMID: 15144565]
[18] M. Q. Zhang, *Nat. Rev. Genet.,* 3:9 (2002) [PMID: 12209144]

Edited by P. Kangueaane
Citation: Kashyap & Tabish, Bioinformation 1(6): 203-207 (2006)