A Web Tool to Discover Full-Length Sequences: Full-Lengther

Antonio J Lara[1], Guillermo Pérez-Trabado[2], David P Villalobos[1], Sara Díaz-Moreno[1], Francisco R Cantón[1], and M Gonzalo Claros[3]

[1] Biología Molecular y Bioquímica, Universidad de Málaga, Campus Universitario de Teatinos, E-29071 Málaga, Spain,
[2] Arquitectura de Computadores, E.T.S.I. Informática, Campus de Teatinos, E-29071 Málaga, Spain,
[3] Departamento de Biología Molecular y Bioquímica
Facultad de Ciencias Universidad de Málaga
29071 Málaga (Spain)
Tel: +34 95 213 72 84
Fax: +34 95 213 20 41
E-mail: claros@uma.es

Summary. Many Expressed Sequence Tags (EST) sequencing projects produce thousands of sequences that must be cleaned and annotated. Here it is presented Full-Lengther, an algorithm that can find out full-length cDNA sequences from EST data. To accomplish this task, Full-Lenther is based on a BLAST report using a protein database such as UniProt. Blast alignments will guide to locate protein coding regions, mainly the start codon. Full-Lengther contains an ORF prediction algorithm for those cases that do not deploy any alignment in the BLAST output. The algorithm is implemented as a web tool to simplify its use and portability. This can be worldwide accessible via `http://castanea.ac.uma.es/genuma/full-lengther/`.

1 Introduction

New biological technology produces a large amount of sequences in form of ESTs (Expressed Sequence Tags). These sequences have to be thoroughly annotated to uncover, for example, its funtion. Currently, the task of annotating EST sequences does not keep pace with the rate at which they are generated [1] since:

1. EST sequence annotation is computationally intensive and often returns no results;
2. EST data suffers from inconsistency problems (error rate, contaminant sequences, low complexity regions, etc.);
3. gene identification programs perform inconsistently as they are sensitive to errors.

One of the most important and difficult tasks is to discover whether the EST sequence was derived from a full-length cDNA. Basically, a sequence is considered full-length when it contains the gene 5' end. In other words, it a EST is flagged as full-length EST, the plasmid where the EST comes from contains the complete cDNA. The final aim is to flag the "full-length" EST sequences among a long EST list, which will facilitate further study of contigs containing a complete gene for its functionality.

Due to the enormous number of ESTs, such a process is not feasible one by one by researchers, so we need to process the ESTs by computers in an automatic form [2]. There are several algorithms that have been developed to predict ORFs in ESTs for which no known orthologues are available, such as NetStart [3], ESTScan [4], ATGpr [5] or OrfPredictor [6], which use neural networks, Markov models and a linear discriminant approach. These algorithms can predict coding regions but they have to be trained with organism-specific sequences [7]. Recently, webservers for identification of full-length cDNAs using BLAST have been described [8, 9] that can receive various input formats and outputs a tab-delimited spreadsheet.

To solve the problem of training to be as generic as possible as well as using the algorithm in a workflow of sequence analysis and/or annotation, we have developed Full-Lengther. Full-Lengther classifies ESTs as full-length, putative full-length or non full-length based on matches (similarities) found by executing BLAST against a protein database [2, 9] (i.e. UniProt). BLAST alignments will guide detection of protein coding regions, mainly the start codon. The algorithm also cares about ESTs that do not show any known alignment in the BLAST output, and it offers an alternative method to detect whether or not it is full-length. We also have implemented a web interface for the algorithm to improve ist accesibility to internet users.

2 Algorithm and Implementation Details

Any standard mRNA has two edges, a start 5'-UTR and an end 3'-UTR. Protein coding region is inside this range, and such region has a well defined start (ATG) and end codons (TAA, TAG or TGA). Clones are produced by a method that guarantees the presence of the 3' edge, but not the 5' end. Since 5' edge can be absent, the presence of this edge in the sequence enables to flag a sequence as complete and we call it "full-length". Our mission is to identify the presence of the 5' edge.

The algorithm will flag sequences with one of the following tags:

Full-Length: It will appear when the algorithm has unambiguously detected that the 5' edge exists.

Non Full-Length: It will appear when the algorithm has unambiguously determined that the 5' edge is not present.

Putative Full-Length: Sometimes it is not clear whether the sequence is full-length. Due to problems like short sequence

length, bad quality, lack of reliability... In these cases the result does not give too much information except that "the sequence maybe is full-length, but it is not sure".

2.1 Selection of Significant Alignments

First of all, it is necessary to determine if there are similar protein sequences in the database for every EST sequence. This is quite easy using BLASTX against a protein database, by now UniProt. At this point we have a list of alignments, each one associated to an expect value (quality value), which is a real number, and the smaller it is, the better. The list of alignments has to be then filtered by a cut off value to remove least significant alignments. The cut off value is set by default to 10^{-6}, although users can modify it.

The SwissProt part of UniProt is the current database used for the BLAST analysis since it is not too large and provides high quality annotations. However, when no hits were found in SwissProt, the Full-Lengther algorithm uses the poorly annotated TrEMBL though it contains a lot of incomplete sequences.

The set of alignments are filtered once again taking into account the best expect value. A new cut off value is generated dynamically to make this new filter. The expect value is a positive real number always less or equal than 1.0. Values closer to zero are better, and usually useful values are very close to zero. The selection criteria is based on the exponent value of the expect. It only passes alignments with expect values that have a similar order of magnitude than the best one. The cut off is chosen using the expression $10^{e-\frac{e}{10}}$, where e is the best expect exponent.

At this point a filtered list of the best alignments is obtained, but this list may be empty if an EST sequence does not produce any significant match in the database. The biological reason for this absence of matches is beyond the scope of this section. To deal with this cases, Full-Lengther provides two ways of analysis, one using the alignments and another taking into account just the structure of the sequence.

2.2 Analysis of Alignments

We describe here the analysis that Full-Lengther performs on the significant alignments found for an EST. Each alignment among EST and subject sequences is processed and classified in one of the three categories (full-length, putative full-length or non full-length) described above. Matches with the most repeated expect value are used to classify the sequence. When it is ambiguous, the average expect value is calculated for every selected category and the best one is then chosen. If even then, there is more than one category with the same expect value, the best expect value of each category is used to determine which one is chosen.

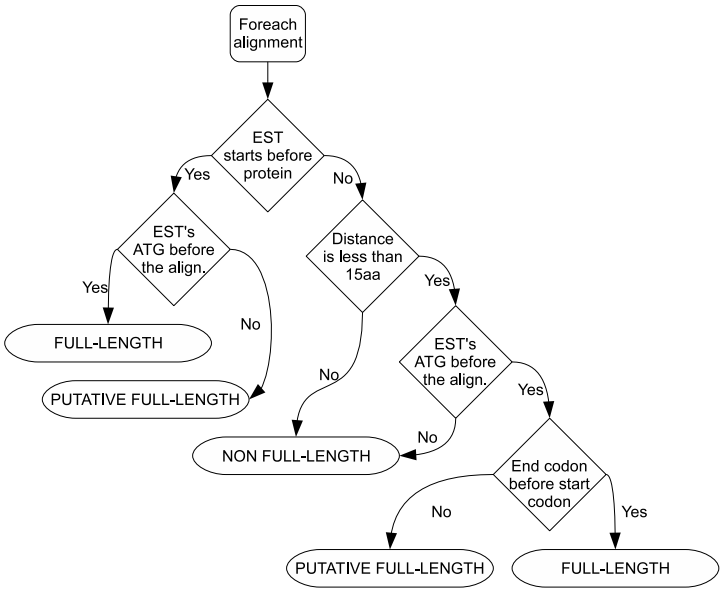

Fig. 1. Processing steps for each alignment

To process an alignment (Fig. 1), first, we have to distinguish between three cases:

1. The EST sequence starts before the subject protein. In this case, the alignment will be full-length if there is a start codon in the sequence before the alignment (Fig. 2(a)), or will be putative full-length if there is no in-frame ATG codon or the first is within the aligned sequence.
2. The EST sequence starts after the protein which has the alignment (Fig. 2(b)), but within a short distance from the beginning. This distance is set to 15 amino acid by default, although users can modify it. Then if there is an in-frame ATG codon, the EST will be flagged as putative full-length, but if there is also an in-frame end codon, EST will be full-length.
3. Any other case will result in non full-length flag.

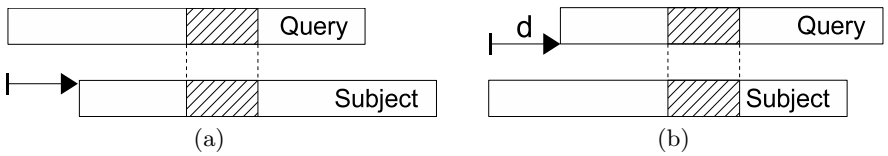

Fig. 2. The query (EST sequence) starts **before** 2(a) or **after** 2(b) the subject (aligned protein), with distance d. Gray regions correspond to the aligned sequence

2.3 Analysis without Alignments

When BLAST renders no similarities for an EST, the analysis must consider only the features of that sequence. The EST is supposed to contain a high quality sequence since all automatic sequencers provide a quality value to qualify removal of any ambiguous fragment [10, 11, 12, 13]. Hence, Full-Lengther will look for the longest possible open reading frame (ORF). All possible ORFs are determined by a simple analysis locating start and stop codons, and then the longest one is selected. The rationale for ORF classification is explained in Fig. 3. ORFs longer than 66 nucleotides are full-length when start (ATG) and end codons are present or when the ATG is at no longer than 150 nucleotides from the 5' end without an in-frame end condon.

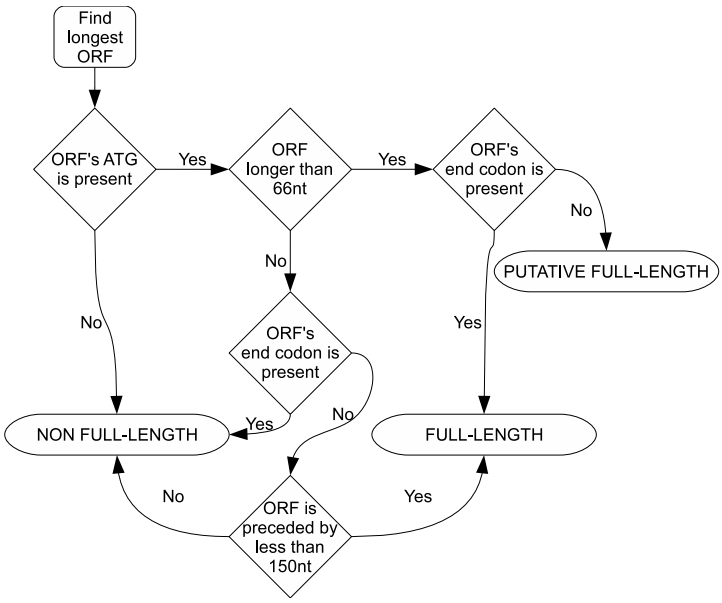

Fig. 3. Processing steps when there are no alignments

The minimum length to consider an acceptable ORF is 66 nucleotides because the probability to obtain an end codon is 1/66, so that shorter ORFs can not be considered significant.

3 Interface

Full-Lengther has a web interface that can process a set of sequences at once. It is actually very easy to use since users only have to fill out a web form (Fig. 4). The sequences to analyze can be provided by pasting text or chosing

a file to upload. The valid input format for sequences is FASTA files containing the set of sequences to be processed. The default parameters for the algorithm are suitable for most purposes and hence, it is not compulsory to change them. The meaning of parameters are as follows:

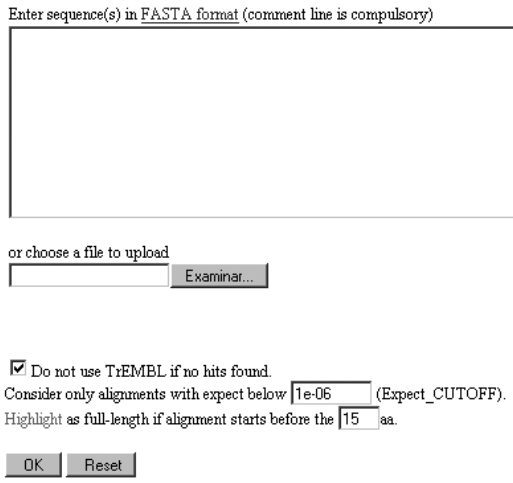

Fig. 4. Full-Lengther form

- A check-box to mark whether we want to use TrEMBL when there are no hits found using Swiss-Prot. By default the use of TrEMBL is disabled.
- The expect cutoff is a BLAST output filter that removes any subject sequence in the alignment that have an expect value greater than specified (10^{-6}).
- The last parameter is the number of amino acid that will be considered as maximum distance from the beginning of the sequence to the protein which is aligned to, to be considered full-length. By default this value is 15 amino acid.

Once the form is submitted and the process is finished we will have access to the results under the form of a list of sequences (Fig. 5), where the sequences are highlighted according to its category, including information about length and position of start codon. A detailed view of each EST sequence can be presented, showing several sections (Fig. 6):

1. The first section shows the ORFs result, where the best and the longest ORF are highlighted.
2. The second section shows the Testcode result for the best ORF.
3. The last section is the graphical list of all the EST alignments. Different colours are used to identify information.

Seq.Name	Min. Start	# Hits found	# Families found	Seq. Length (nt)
GEM-05-1D07-T3-A1.ab1	28	4	1	698
GEM.48-D07-T3-96-F1.ab1	59	6	6	640
GEM-04-1C08-T3-A1.ab1	25	2	1	689
GEMINI-81-D07-T3-96-F1.ab1	174	3	1	562
GEM.48-D10-T3-96-F1.ab1	19	5	4	570
GEM-05-1E01-T3-A1.ab1	261	1	4	693
GEMINI-82-G02-T3-96-F1.ab1	87	3	2	647
GEM.49-F06-T3-96-L1.ab1	158	2	4	710
GEM-05-1E02-T3-A1.ab1	379	4	1	658
GEM-05-1G08-T3-A1.ab1	428	2	3	721
GEM.48-H05-T3-96-F1.ab1	49	20	4	685
GEM-05-1C09-T3-A1.ab1	-	0	0	629

# FULL-LENGTH	2
# Putative full-length	2
# Non full-length	8
# Total seqs	12

Fig. 5. Summary of sequences

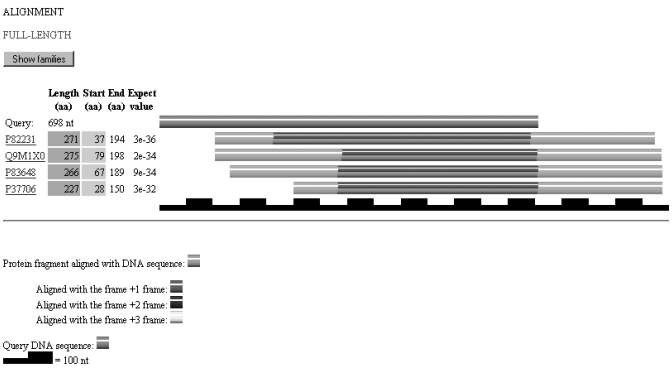

Fig. 6. EST sequence alignment result. Aligned hits are shown in green, unpaired sequences are in gray and the putative ORF appears in blue.

4 Discussion

Full-Lengther has been evaluated with real data to verify the correctness and accuracy of its results. In contrast to similar programs such as TargetIdentifier, Full-lengther shows a simple and intuitive output, making easy to form an accurate idea of the results. Moreover, it is designed as a library so that it can be easily integrated in other algorithms to build workflows for EST annotation.

Since Full-Lengther emerged from a collaboration between biologists and computer scientists, it is continuously improved and new features are included.

For example, it is planned to use databases of completely sequenced genomes in order to avoid partial sequences that can lead to erroneous interpretations.

5 Acknowledgements

This work is supported bye the Spanish MEC grants AGL-2006-07360/FOR and BIO2006-06216 as well as the Junta de Andalucía grant AGR-663 and foundings to the research groups CVI-114, TIC-160, and TIC-113.

References

1. Chen Y, Carlis J, Shoop E, Riedl J (2001) International Joint Conference on Artificial Intelligence 2001, Workshop on Inconsistency in Data and Knowledge, Seatle, WA
2. Terol J, Conesa A, Colmenero JM, Cercos M, Tadeo F, Agusti J, Alos E, Andres F, Soler G, Brumos J, Iglesias DJ, Gotz S, Legaz F, Argout X, Courtois B, Ollitrault P, Dossat C, Wincker P, Morillon R, Talon M (2007) Analysis of 13000 unique Citrus clusters associated with fruit quality, production and salinity tolerance. BMC Genomics 8, 31
3. Pedersen AG, Nielsen H (1997) Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. Proc Int Conf Intell Syst Mol Biol 226–233
4. Iseli C, Jongeneel CV, Bucher P (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol 138–148
5. Salamov A, Nishikawa T, Swindells MB (1998) Assessing protein coding region integrity in cDNA sequencing projects. Bioinformatics 14:384–390
6. Min XJ, Butler G, Storms R, Tsang A (2005) OrfPredictor: predicting protein-coding regions in EST-derived sequences. Nucleic Acids Res 33:677–680
7. Nadershani A, Fahrenkrug SC, Ellis LBM (2004) Comparison of computational methods for identifying translation initiation sites in EST data. BMC Bioinformatics 5, 14
8. Nishikawa T, Ota T, Isogai T (2000) Prediction whether a human cDNA sequence contains initiation codon by combining statistical information and similarity with protein sequences. Bioinformatics 16:960–967
9. Min XJ, Butler G, Storms R, Tsang A (2005) TargetIdentifier: a webserver for identifying full-length cDNAs from EST sequences. Nucleic Acids Research 33:669–672
10. Walther D, Bartha G, Morris M (2001) Basecalling with LifeTrace. Genome Res 11:875-888
11. Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. Genome Res 8:195-202
12. Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome research 9:868-877
13. Falgueras J, Lara A, Cantón FR, Pérez-Trabado G, Claros MG (2007) SeqTrim: a validation and trimming tool for all purpose sequence reads. This issue