

**TargetFinder and Annotator: a Simple Approach for Finding Full-length Target
cDNAs and for Annotating EST Sequences**

X. J. Min¹, G. Butler^{1,2}, R. Storms^{1,3}, A. Tsang^{1,3}

¹Centre for Structural and Functional Genomics, ²Department of Computer Science, and

³Department of Biology, Concordia University, 1455 de Maisonneuve Blvd., W.,

Montreal, Quebec, Canada H3G 1M8

jack@gene.concordia.ca

Running Title: Finding Full-length cDNAs and Annotating EST Sequences

Key words: Full-length cDNA, Annotation, EST

Correspondence:

X. J. Min

Centre for Structural and Functional Genomics

Concordia University

1455 de Maisonneuve Blvd., W. H1223

Montreal, Quebec, Canada H3G 1M8

E-mail: jack@gene.concordia.ca

ABSTRACT

In a large scale EST (expressed sequence tag) or cDNA sequencing project, it is often desirable to know whether the ESTs identify genes of interest and whether the cloned cDNAs include intact coding regions (are of full-length). In this work, we present two Perl tools, TargetFinder and Annotator. TargetFinder automates the identification of full-length cDNAs from assembled EST sequences including singletons and contigs. Annotator is used to annotate ESTs and their assembled sequences by assigning a provisional function to each sequence and predicting whether or not they include intact coding regions. The programs use the output of BLASTX to predict the correct reading frame, search for a putative start codon, and predict whether a query sequence includes an intact ORF. In addition the programs also predict whether the sequence of the coding region of a cDNA clone or a contig is complete. Using our own *Aspergillus niger* EST data, TargetFinder rapidly and accurately found full-length target genes within a large set of assembled ESTs and Annotator functionally annotated the ESTs and their assembled sequences.

Introduction

Single-pass cDNA sequencing (or EST, expressed sequence tag) remains the primary tool for the identification of novel genes as well as for functional genomics analyses. The number of EST entries deposited at GenBank has reached 15,847,502 as of February 28, 2003 (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). The comprehensively analyzed and annotated gene indices in The Institute of Genomic Research (TIGR, <http://www.tigr.org/tdb/tgi/>) currently cover 59 species. Full-length cDNA databases have been built up in mouse and *Arabidopsis* [1, 2]. Critical questions often raised in these works include:

- Is a cDNA or EST sequence of full-length, i. e. does it contain the translation initiation site?
- How many full-length genes have been obtained for a species in such a project?
- When comparing to a database of protein or nucleotide sequences in GenBank, how many sequences have a match of high statistical significance?

BLAST and its variants [3, 4], as well as similar programs such as DPS [5] are widely used for searching protein or nucleotide databases to find sequences that are similar to a query sequence. BLASTX (a nucleotide query against a protein database) has been shown to be able to reliably identify protein coding regions of a DNA query by database similarity search [6]. Some programs, such as ESTScan [7], and DIANA-EST [8], were developed to predict coding regions of ESTs. To our knowledge, only two programs exist to predict whether or not a cDNA contains a translation start codon. The first one, ATGpr, uses a linear discriminant approach [9]. The second, ATGpr_sim, an improved version of the first one, showing increased sensitivity and specificity, uses an algorithm combining statistical and similarity information [10]. However, the algorithm of ATGpr_sim [10] uses the alignment position parameters of the high score pair in BLASTX to predict whether a query sequence is full-length. Their program, therefore, cannot identify full-length queries that are shorter in length than the database match

(subject) sequence in BLASTX. Moreover, the existence of a start codon within the frame is not confirmed either. The publicly available ATGpr_sim program can only process one query sequence at a time.

Using sequence similarity for functional annotation of ESTs and cDNAs is widely accepted by major genomics organizations [9, 10, 11]. However, at the time of this submission, we were unaware of any available program that combined similarity-based functional annotation and full-length prediction for ESTs or cDNAs. In this work, we present a new, practical and simple algorithm to find full-length target genes from a large set of assembled ESTs and to annotate ESTs and assembled sequences by combining functional annotation and full-length prediction. Furthermore, the program also predicts whether the protein coding region of a query sequence is completely determined. Using data generated from our *Aspergillus niger* EST project, we found that TargetFinder and Annotator were highly effective and accurate in identifying full-length target genes and in annotating EST sequences.

Methods

Definitions

Almost all mRNAs in eukaryotes are monocistronic, i.e. they encode a single polypeptide chain as indicated by their single start site for protein synthesis, although some polycistronic genes are found in plants [12] and mammals [13]. Hence a typical full-length cDNA in eukaryotes contains a 5' untranslated region (5' UTR), an amino acid coding region or an open reading frame (ORF), and a 3' untranslated region (3' UTR) (Figure 1A). An ORF is marked by the presence of a start codon (ATG) in the 5' region and a stop codon (TAA, TAG, or TGA) in the 3' region [14]. In the 5' UTR there are often one or more stop codons. Full-length in this work refers to a sequence (individual ESTs, singletons and contigs assembled from several ESTs) that possesses a putative translation initiation site. As our cDNAs were constructed using oligo-dT as primers in the first-strand synthesis, all the cDNA clones are expected to have intact 3' regions. To predict whether a sequence is full-length, we used the approach of searching for a start codon and for stop codons before a start codon in the putative 5' UTR (referred to here as 5' stop codons). To determine whether the coding region of a cDNA is completely sequenced, we searched for the stop codon at the 3' end of the sequence, which we call 3' stop codon, within the reading frame, and also considered the position of the 3' stop codon in the query sequence relative to the length of the subject sequence in BLASTX.

We used the translation frame predicted by BLASTX in the highest score pair to search for a putative start codon. The cDNA sequences were classified into the following categories: full-length, short full-length, possible full-length, ambiguous, and partial. The criteria for these categories are described below.

Full-length: A sequence is considered full-length when it satisfies either one of the following 2 criteria. (1) The sequence has a 5' stop codon followed by a start codon (Figure 1A). (2) It has a start codon only, and based on the alignment the predicted start codon in the query sequence matches to a position that is no further than 10 codons downstream of the start of the subject sequence (Figure 1C).

Short full-length: The sequence has an in-frame ATG codon but without a 5' stop codon (Figure 1C). Further, the putative start codon in the query sequence matches to a position

that is between 10 to 49 codons downstream of the start of the subject sequence. A number will be given by the programs to indicate this position.

Possible full-length: In some instances, the sequence quality is poor and an extensive portion of the 5' region would be trimmed by Lucy [15]. Using the cleansed sequence after trimming, the query sequence is not categorized as full-length or short full-length. If the length of the region removed by Lucy is added, the resulting sequence would encode a peptide that is roughly the length of a predicted full-length clone. This sequence will be categorized as "possible full-length". This cDNA clone will need to be sequenced again to confirm whether it can be classified as a full-length clone before it is used for further laboratory experiments.

Ambiguous: The sequence has a 5' stop codon but does not have a start codon (Figure 1D).

Partial: A sequence that does not belong to any other above described categories (Figure 1E).

Unknown: The programs described in this work are designed to analyze cDNAs that are sequenced from the 5' end, i.e. the sense strand. In the BLASTX report, these sequences should align with the subject sequences in the positive frame (+1, +2, or +3). Therefore, if the query sequence aligns to the subject sequence in a negative frame (-1, -2, or -3) and it does not belong to the full-length or short full-length categories, it is classified as "unknown." Since an antisense strand is sequenced from the 3' end, it is impossible to know whether it is a partial or full-length clone.

Algorithm and implementation

Input: BLASTX output, DNA sequence file in FASTA format, assembly (ace) file from Phrap [16] (Figure 2), a file containing information about the sequence length of a low quality region removed by the Lucy program (Lucy file) [15] (Figure 2). The ace file from Phrap provides the information regarding individual ESTs being assembled in a contig, and the Lucy file provides the length of a low quality region of an EST removed by Lucy.

Output: Query list with annotated features.

Variables used in the algorithm (Figure 1):

d1: the predicted peptide length from the start codon to the first amino acid of the query in the highest score pair (HSP) alignment in the BLASTX. It is calculated using the formula: $d1 = (\text{query beginning position in the HSP alignment} - \text{predicted start codon position} - \text{frame} + 1)/3$;

d2: the subject's beginning position in the HSP alignment in the BLASTX;

d3: the estimated length of the low quality portion of cDNA sequence removed by Lucy. The length of the low quality portion of a sequence removed by Lucy, obtained from the Lucy file, contains a portion of a vector, an adaptor, and a low quality region of cDNA insert. Examining the untrimmed EST sequence, we found that the vector and adaptor length is about 60 bp, therefore, $d3 = \text{total trimmed length} - 60$.

The output of BLASTX is parsed, and if there is a match to an entry in a database for a query sequence, the frame used for predicting the protein sequences of the query in the highest score pair is obtained. The query sequence is retrieved from the DNA FASTA file. If the frame is negative, the reverse complementary sequence is generated, and the alignment parameters need to be recalculated. The query sequence is divided into two subsequences:

(i) Subsequence 1 is from the frame to the end position of the first line of the query in the HSP alignment in BLASTX,

(ii) Subsequence 2 is the rest of the query sequence, i.e. the end position in the first line in the HSP alignment plus 1 to the end of the sequence.

The reason we only use the end position of the first line in the HSP alignment as the end position of subsequence 1 is that there are 60 amino acids per line in the BLASTX output, and if there is no in-frame start codon in that part of the sequence, it will be a partial gene because we use 50 as a threshold. Subsequence 1 is used to search for a 5' end stop codon and a start codon, subsequence 2 is used to search for a stop codon at the 3' end.

All the cases of full-length cDNA can be summarized as follows:

Case 1: If a cDNA contains a 5' stop codon and a start codon within a frame (Figure 1A, 1B), it is full-length.

Case 2: If there is a start codon but without a 5' stop codon (Figure 1C), the length difference between the query sequence and the subject will be considered.

Subcase 2.1: If $(d2 - d1) < 10$, the cDNA is full-length;

Subcase 2.2: If $10 \leq (d2 - d1) < 50$, the cDNA is short full-length with a number given to show the length difference as compared to the subject. We have chosen the limit of 50, because aligning close related members of a protein family showed that the length differences are rarely more than 50 amino acids.

Subcase 2.3: The sequence length (containing vector, adaptor, and a low quality region of cDNA insert) trimmed by Lucy [15] (Figure 2) will be taken into consideration when judging whether a cDNA is full-length.

Subcase 2.3.1: If $(d1 + d3/3 - d2) \geq 0$, the query is possibly full-length;

Subcase 2.3.2: If $(d1 + d3/3 - d2) < 0$, the query is partial.

Case 3: If a 5' stop codon can be found without a start codon (Figure 1D), the sequence is categorized as ambiguous. This anomaly may be caused by a frame shift due to an insertion or a deletion within a sequence, a common phenomenon in ESTs because most ESTs are sequenced only once. Such a sequence needs to be manually examined and sequenced again if it is used for further experimentation.

Case 4: If both a 5' stop codon and a start codon are absent (Figure 1E), the length of the sequence trimmed by Lucy will be examined:

Subcase 4.1: If $(d1 + d3/3 - d2) \geq 0$, the query is possibly full-length;

Subcase 4.2: If $(d1 + d3/3 - d2) < 0$, the query is partial;

All the above cases (Case 1 to Case 4) are for a 5' to 3' (sense) sequence. For an antisense (3' to 5') strand, if it is not categorized as full-length (Case 1, Case 2: Subcase 2.1) or short full-length (Case 2: Subcase 2.2), it will be categorized as "unknown," since it may not be completely sequenced.

In addition to predicting whether a cDNA is of full-length, we also predicted whether the coding region of a cDNA clone at 3' end is completely sequenced for an EST clone, or whether a contig's coding region at 3' end is completely derived by assembling overlapped ESTs. Such a prediction is needed to determine whether an entire coding region of a query sequence is obtained, and this information may be helpful when further characterizing the gene in laboratory experiments. To predict whether the 3' coding region is completely obtained, we search for a 3' end stop codon in the subsequence-2

and also consider the relative lengths of a query and a subject. If there is a stop-codon at the 3' end and the subject length – subject beginning position in the alignment – (query length – query beginning position in the alignment)/3 \leq 0, then the 3' coding region of a query is completely obtained.

Both TargetFinder and Annotator were implemented with the same algorithm as described above but with a slight difference in output format. TargetFinder is intended to find full-length targets quickly from a large set of assembled EST sequences including singletons and contigs. BLASTX was performed against a specifically built target database containing full-length target protein sequences of our interest. We used TargetFinder to process the output of BLASTX. The output of TargetFinder includes the gene name, query name, prediction of full-length, sequence status of the 3' end coding region, and the HSP complete head information in BLASTX including the subject definition line, score, E-value, identities, and frame. In order to easily sort genes by gene names, the terms including 'probable', 'putative', 'possible', and 'similar to' in the subject definition were removed by TargetFinder. Each field in the output of TargetFinder is tab separated, thus, the output can be exported into a spreadsheet, and sorted easily to identify target genes.

The aim of Annotator is to automate the annotation of each EST sequence before Phrap assembling and that of each assembled sequence including contigs and singletons after Phrap (Figure 2). Therefore, all query sequences are searched against the NCBI non-redundant protein database with an E-value limited to 1E-5 to control the reliability of functional annotation. The output of Annotator includes the query sequence name, gene definition with or without a qualifier, prediction of full-length, and sequence status of the 3' end of the coding region. The subject information and related parameters in BLASTX are not included in the output of Annotator since the information about the subjects of the top five high score pairs in BLASTX will be processed by another program called “BlastParser” and will be integrated into our database. Annotator assigns a qualifier to a putative gene function to indicate the level of similarity between the query and the subject. For a high score pair with an identity \geq 90% in BLASTX, no qualifier is given and the subject's function is directly assigned to the query. If 70% \leq identity $<$ 90%, the qualifier is “Homologous to,” and for 50% \leq identity $<$ 70%, the qualifier is “Similar to,” The rest is assigned as “Weakly similar to.”

To facilitate target finding and annotation, we also implemented two other programs, including:

- “Translation,” which translates all query sequences into protein sequences using the predicted frame in BLASTX;
- “BlastParser,” which extracts BLASTX and BLASTN results up to the top five high score pairs, including the subject definition, E-value, score, the beginning and end positions in the aligned query and subject, and the coverage of the alignment as complete or partial.

Sequence quality control and application of TargetFinder and Annotator

We have applied TargetFinder to find full-length targets from singletons and contigs assembled from our own generated ESTs, and Annotator to annotate all ESTs and all assembled sequences. In our project, the majority of the ESTs were sequenced from the 5' end only, a small portion of them were sequenced from both ends, and a small portion of them were sequenced from the 3' end only.

The general procedure for EST sequence quality control, assembling, target

finding and EST annotation is shown in Figure 2. In brief, sequence chromatograms were traced and DNA quality values were assigned by Phred [17, 18], then low quality regions of DNA were trimmed, vector sequences at both ends were spliced, vector contaminations as well as sequences with an insert less than 100 bp in length were removed by Lucy [11, 15], and individual ESTs were assembled into singletons or contigs by Phrap [16], then BLASTX was performed and the output of BLASTX was processed by TargetFinder or Annotator.

To test the accuracy of the predictions by TargetFinder and Annotator, we downloaded 449 protein sequences of *Aspergillus niger* from the NCBI GenBank. The same species was used to generate our ESTs. There are 403 full-length protein sequences among the downloaded protein sequences. We only used full-length protein sequences to build up a database for performing BLASTX to test the programs, since if the subject in a high score pair is not of full-length, it is difficult to judge the prediction accuracy of the programs. We used 3866 non-redundant singletons and contigs assembled from 8176 ESTs to perform BLASTX against the *A. niger* database. As both TargetFinder and Annotator are implemented using the same algorithm, we first made sure that the outputs of the prediction were identical for both programs. To make a manual comparison more reliable, we only selected query sequences that showed an identity $\geq 90\%$ with a subject sequence. These selected sequences were translated into protein sequences by the Translation program we implemented using the frame as determined by BLASTX. To examine the accuracy of the prediction as to whether a query is full-length and whether its 3' coding region is completely obtained, we aligned the translated query protein sequence and the subject sequence using ClustaX [19].

To find genes of interest, TargetFinder was applied to process the output of the BLASTX of assembled sequences against our own full-length target protein database that was built with protein sequences interesting for us and retrieved from the NCBI. Annotator was used to annotate all individual ESTs, singletons and contigs. The annotated sequences with BLASTX results are stored in a database. All individual ESTs, singletons and contigs were used to perform the BLASTX search against the NCBI non-redundant protein database for annotation. Some sequences may not have a protein coding region, i. e. only have a 5' UTR or a 3' UTR, and others may be ribosome RNAs. For this reason, all sequences without a database match shown as "No hits found" with an E-value limited to $1E-5$ in the output of the BLASTX were then retrieved and used to perform the BLASTN search against the NCBI non-redundant nucleotide database with an E-value limited to $1E-5$. Thereafter, Annotator was applied to process the BLASTN output and a provisional function was assigned to queries having a database match. The values of full-length prediction and sequence status of the 3' coding region were assigned as "unchecked," since no protein translation frame was available in BLASTN.

Results

The results of a manual examination regarding the accuracy of TargetFinder and Annotator are summarized in Table 1. Altogether 90 query sequences showed identities $\geq 90\%$ with subjects in the BLASTX against the downloaded *A. niger* full-length protein sequences. 49 sequences were predicted to be full-length. Among them, 48 sequences were correctly predicted. There was only 1 sequence predicted as full-length, which was one amino acid shorter than the full-length subject. There are two ATG sites at the 5' end of the subject ORF, with the second ATG separated by only 3 codons from the first. Since the query sequence contained only the second ATG, it was predicted to be full-

length. Thus the percentage of sequences correctly predicted as full-length was 98%. The prediction was correct for all the partial sequences. Only one short full-length sequence is predicted, and it is a partial gene, 16 codons shorter than the subject, but the start codon in the query sequence is 49 codons shorter than the start codon in the subject sequence. The one ambiguous sequence is a partial sequence, but there is a stop codon in the open reading frame. This occurred because of a frame shift that is caused by an insertion. There are 5 partial sequences having negative frames, since they were sequenced from 3' end and are not categorized as full-length or short full-length, so they are correctly classified as "unknown." All the predictions for the sequence status of the 3' end coding region of a query were correct. In summary, manual comparison of the outputs of TargetFinder and Annotator to published protein sequences confirmed that the predictions of both full-length and the 3' coding sequence status by TargetFinder and Annotator were highly accurate.

We have regularly applied TargetFinder to find full-length target genes from our own ESTs generated from *A. niger* and other fungal species, along with the progress of sequencing in fungal EST projects. Some of the key targets found by TargetFinder were completely sequenced and are being further characterized in laboratory experiments.

Annotator was applied after the sequencing of ESTs in *Aspergillus niger* was completed in our project. Annotated sequences will soon be publicly accessible. After running the sequence quality control and assembling pipeline, we obtained 4845 non-redundant singletons and contigs from 11544 ESTs (Table 2). Among the assembled sequences, 2846 have at least one database match in BLASTX with an E-value $\leq 1E-5$, and hence each of them was assigned a provisional function. Annotation results showed that, among these assembled sequences with database matches in BLASTX, there were 1330 full-length and, among them 342 contain an entire coding region, 98 short full-length, 6 possible full-length, 1317 partial, 19 ambiguous, and 76 "unknown". The "unknown" sequences were sequenced from the 3' end, as they may not be completely sequenced, it could not be predicted whether they were full-length.

The complete set of *A. niger* ESTs was also annotated with Annotator. Among 11544 EST sequences, 7123 ESTs have a hit with an E-value $\leq 1E-5$ in BLASTX. Among them, there are 3205 full-length, 324 short full-length, 46 possible full-length, 3352 partial, 45 ambiguous, and 151 unknown. Sequences without a hit in BLASTX were then retrieved and used to perform BLASTN against the NCBI non-redundant nucleotide sequences to find a database match with E-value $\leq 1E-5$. There are 27 assembled sequences and 126 ESTs having a database match in BLASTN. They were then annotated with Annotator. Since the translation frames of these sequences were unknown, the value "unchecked" was assigned for both full-length prediction and the status prediction of the 3' end coding region.

Discussion

Using sequence similarity to find homologues or closely related genes is a widely adopted practice by researchers. Using the same approach to assigning a gene function to ESTs or genes predicted from genomic sequences is also widely accepted by major genomic institutes such as TIGR [20, 21] and RIKEN [1]. We used a similar approach for functional annotation of our ESTs and for finding target genes from our ESTs and assembled sequences. However, in this work, we also predict whether a query sequence is full-length and whether its 3' coding region is complete by searching for a putative start codon, and a 3' stop codon and simultaneously considering the relative lengths of

the query and the subject. Though there are programs available for cDNA full-length prediction [9, 10], to our knowledge we are the first to combine functional annotation and full-length prediction for ESTs or cDNAs. Because the algorithm used in TargetFinder and Annotator allows direct searching for a start codon and stop codons within a translation frame, the programs do not require that all entries in the database used for BLASTX search are full-length. However, since not all the entries in the NCBI non-redundant protein database are full-length, some inaccuracy may occur for Annotator when using the lengths of a query and a subject to predict whether a query is full-length, when a subject itself is not full-length. To improve the accuracy of the TargetFinder output, we removed all the incomplete protein sequences from our own target database. Since the available ATGpr_sim program processes only one query sequence at a time [10], we are not able to compare our results directly. However, performing manual comparison to known full-length protein sequences in the same species, we found that both TargetFinder and Annotator are highly accurate (Table 1). We believe that both programs are more accurate than ATGpr_sim since the algorithm in our program uses the predicted frame in BLASTX to search for a start codon directly and also take the relative sequence length into consideration, while ATGpr_sim only considers the relative length of the query sequence as compared to the subject [10]. In addition to predicting whether cDNA sequence is full-length, our programs also predict accurately whether the 3' end of the coding region is complete. Predictions of full-length and the sequence status of 3' end of the coding region are helpful to identify full-length target genes with an entire coding region obtained. That will facilitate the characterization of target genes in laboratory experiments.

The algorithm used to implement TargetFinder and Annotator is simple and robust therefore it can be implemented in other programming languages and with more user options for the parameters used in the algorithm. Both TargetFinder and Annotator could also be used to process data without Lucy file from Lucy and Ace file from Phrap. However, the prediction of the category of "possible full-length" may be less accurate without those two files as the length of a low quality region of a cDNA at the 5' end, trimmed by a program such as Lucy, will be considered as zero.

In addition to applying TargetFinder and Annotator to find full-length target genes from assembled ESTs and to annotate ESTs and assemble sequences, we have also tested TargetFinder to find target genes from a set of predicted genes from genomic sequences as well as Annotator to annotate the same set of predicted genes. The set of predicted genes are generated with a gene prediction tool against genomic DNA sequences. We found that TargetFinder was very effective at finding target genes from a set of predicted genes and Annotator was able to functionally annotate them.

In conclusion, we present a new and simple algorithm to find full-length target genes from assembled ESTs and to annotate ESTs and assembled sequences. This algorithm combines the similarity based functional annotation with the prediction of full-length and the sequence status of the 3' end of the coding region. We implemented TargetFinder and Annotator in Perl using the above algorithm and have applied TargetFinder to identify full-length target genes and Annotator to annotate ESTs in our projects. We found that the predictions of full-length characteristic and sequence status of the 3' coding region are both highly accurate. We believe that TargetFinder could be well applied to identify full-length genes of interest from EST data in GenBank and Annotator could be used to functionally annotate those sets of ESTs. There are also potential applications for TargetFinder to find selected target genes within the set of

genes predicted from genomic sequences and for Annotator to functionally annotate these predicted genes.

Acknowledgments

The project is supported by Genome Quebec and Genome Canada. We thank Dr. M. Foldeaki for her suggestions and careful editing of the manuscripts.

References

- [1] H. Bono, T. Kasukawa, M. Furuno, Y. Hayashizaki, and Y. Okazaki, "FANTOM DB: database for functional annotation of RIKEN mouse cDNA clones", *Nucleic Acids Res.*, Oxford University Press, 2002, 30(1), pp. 116-118.
- [2] M. Seki, M. Narusaka, S. Kamiya, J. Ishida, M. Satou, T. Sakurai, M. Nakajima, A. Enju, K. Akiyama, Y. Oono, M. Muramatsu, Y. Hayashizaki, J. Kawai, P. Carninci, M. Itoh, Y. Ishii, T. Arakawa, K. Shibata, A. Shinagawa, K. Shinozaki, "Functional annotation of a full-length Arabidopsis cDNA collection", *Science*, American Association for the Advancement of Science, 2002, 296, pp. 141-145.
- [3] S.F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool", *J. Mol. Biol.*, Academic Press, 1990, 215, pp. 403-410.
- [4] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res.*, 25, pp. 3389-3402.
- [5] X. Huang, "Fast comparison of a DNA sequence with a protein sequence database", *Microbial and Comparative Genomics*, 1996, 1 (4), pp. 281-291.
- [6] W. Gish, and D. J. State, "Identification of protein coding regions by database similarity search", *Nature Genet.*, 1993, 3, pp. 266-272.
- [7] C. Iseli, C. V. Jongeneel, P. Bucher, "ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences", *Proceeding of the Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, CA, 1999.
- [8] A. G. Hatzigeorgiou, P. Fizev, and M. Reczko, "DIANA-EST: a statistical analysis", *Bioinformatics*, Oxford University Press, 2001, 17(10), pp. 913-919.
- [9] A. Salamov, T. Nishikawa, and M. B. Swindells, "Assessing protein coding region integrity in cDNA sequencing projects", *Bioinformatics*, Oxford University Press, 1998, 14(5), pp. 384-390.
- [10] T. Nishikawa, T. Ota, and T. Isogai, "Prediction whether a human cDNA sequence contains initiation codon by combining statistical information and similarity with protein sequences", *Bioinformatics*, Oxford University Press, 2000, 16 (11), pp. 960-967.
- [11] F. Liang, I Holt, G. Pertea, S. Karamycheva, S. Salzberg, and J. Quackenbush, "An optimized protocol for analysis of EST sequences", *Nucleic Acids Res.*, Oxford University Press, 2000, 28, pp. 3657-3665.
- [12] D. J. Leader, G.P. Clark, J. Watters, A. F. Beven, P. J. Shaw, and J. W. Brown, "Clusters of multiple different small nucleolar RNA genes in plants are expressed as and processed from polycistronic pre-snoRNA", *EMBO J.*, Oxford University Press, 1997, 16, pp. 5742-5751.
- [13] T. Blumenthal, "Gene clusters and polycistronic transcription in eukaryotes", *Bioassays*, Wiley, Cambridge, UK, 1998, 20, pp. 480-487.
- [14] F. Mignone, C. Gissi, S. Liuni, and G. Pesole, "Untranslated regions of mRNAs", *Genome Biol.*, BioMed Central Ltd, London, UK, 2002, 3, reviews0004.1 – 00001.10.

- [15] H. Chou, and M. H. Holmes, "DNA sequence quality trimming and vector removal", *Bioinformatics*, Oxford University Press, 2001, 17, pp. 1093-1104.
- [16] Phil Green, "Documentation for Phrap and Cross-match (Version 0.990319)", <http://www.phrap.org/phrap.docs/phrap.html>, 1999.
- [17] B. Ewing, L. Hillier, M. Wendl, P. Green, "Base-calling of automated sequencer traces using Phred. I. Accuracy assessment", *Genome Res.*, Cold Spring Harbor Press, 1998, 8, pp. 175-185.
- [18] B. Ewing, and P. Green, "Base-calling of automated sequencer traces using Phred. II. Error probabilities", *Genome Res.*, Cold Spring Harbor Press, 1998, 8, pp. 186-194.
- [19] A. Aiyar, "The use of CLUSTAL W and CLUSTAL X for multiple sequence alignment", *Bioinformatics Methods and Protocols, Methods in Molecular Biology*, Volume 32, Edited by S. Misener and S. A. Krawetz, Humana Press, Totowa, New Jersey, pp. 221-241.
- [20] J. Quackenbush, F. Liang, I. Holt, G. Pertea, and J. Upton, "The TIGR gene indices: reconstruction and representation of expressed gene sequences", *Nucleic Acids Res.*, Oxford University Press, 2000, 28 (1), pp. 141-145.
- [21] J. Quackenbush, J. Cho, D. Lee, F. Liang, I. Holt, S. Karamycheva, B. Parviz, G. Pertea, R. Sultana, and J. White, "The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species", *Nucleic Acids Res.*, Oxford University Press, 2001, 29(1), pp. 159-164.

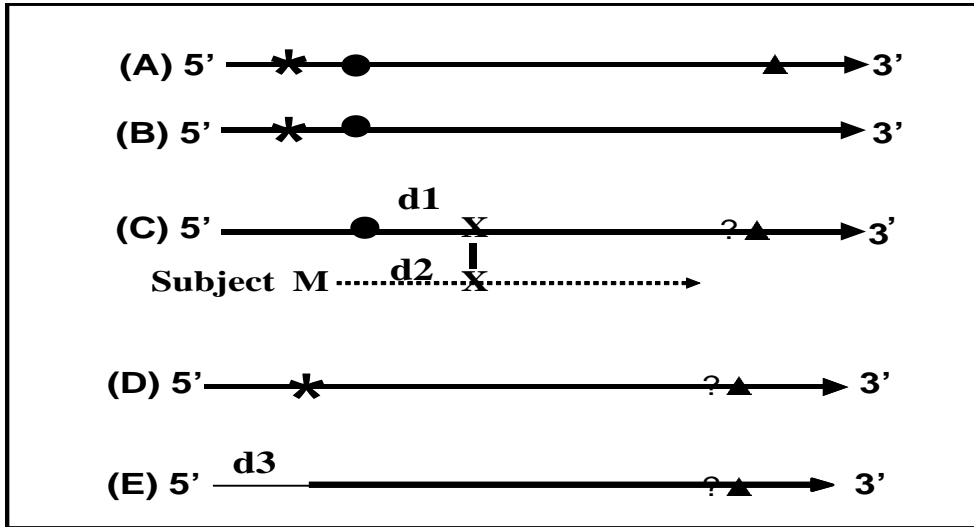


Figure 1. Categories of cDNA sequences. (*) Stop codon before start codon (5' end stop codon); (●) predicted start codon within an open reading frame; (▲) stop codon after start codon (3' end stop codon); (?) indicates checking if a stop codon after start codon exists; (X) the first amino acid in the alignment of the highest score pair in BLASTX; (M) methionine; (d1) the length of predicted amino acids from a predicted starting codon to X; (d2) the length of M to X in the subject sequence of the highest score pair in BLASTX; (d3) length of cDNA sequence trimmed by Lucy including a portion of a vector, an adaptor and a low quality region of a cDNA insert; thick solid line: cDNA sequences after processing by Lucy; thin solid line: a low quality region of a cDNA sequence removed by Lucy at 5' end; dashed line: amino acid sequence of the subject in BLASTX. (A) A typical full-length cDNA including one or more stop codon(s) (5' end stop codon) before a predicted start codon, a start codon, and a 3' stop codon within an open reading frame (ORF). (B) A full-length cDNA without a stop codon at the 3' end shows it is not completely sequenced. (C) A cDNA with a start codon but without a 5' stop codon, whether this cDNA is full-length needs to be judged from the length difference of the predicted protein of the cDNA and the subject in BLASTX. (D) A cDNA having a 5' end stop codon but lacking a start codon after a stop codon. This is an ambiguous sequence and needs to be checked manually. (E) A cDNA sequence lacking a stop codon and a start codon at the 5' end. The length of the low quality region of cDNA removed by Lucy will be taken into consideration when predicting whether it is full-length. In all cases, the eventual existence of a stop codon at the 3' end and the relative lengths of a query and a subject will be used to predict whether the coding region of a cDNA is completely sequenced or whether its coding sequence is completely derived in case of a contig.

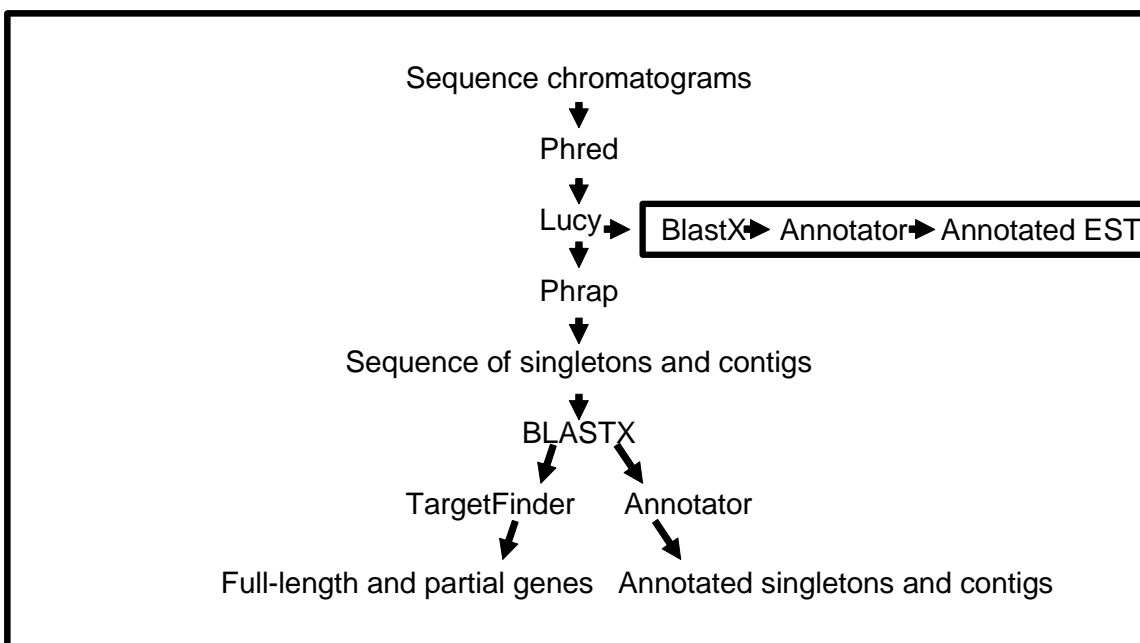


Figure 2. Procedure for analysis of EST sequences. Sequence chromatograms are traced by Phred, vector and low quality regions are removed by Lucy, and ESTs are assembled by Phrap. After BLASTX searching, full-length target genes are identified by TargetFinder from assembled EST sequences including singletons and contigs, and all individual ESTs, singletons and contigs are annotated by Annotator.

Table 1. Results of comparison of TargetFinder prediction of assembled EST sequences of *Aspergillus niger* with deposited full-length protein sequences in GenBank from the same species.

		Predicted number	Correctly predicted number	%
Categories of full-length	Full-length	49	48	98
	Short full-length	1	1	100
	Possible full-length	0	0	-
	Ambiguous	1	1	100
	Partial	33	33	100
	Unknown	6	6	100
Strand and 3' coding region status	Sense complete	24	24	100
	Sense partial	66	66	100
	Antisense complete	6	6	100
	Antisense partial	0	0	-

Table 2. A summary of annotation of ESTs and their assembled sequences of *Aspergillus niger* by Annotator. An individual EST is a good quality cleansed EST resulting from Lucy processing. ESTs are assembled by Phrap to generate a non-redundant set of singletons and contigs. The E-values in both BLASTX and BLASTN are limited to $\leq 1E-5$. BLASTX was first performed, and then query sequences without a match in the database in BLASTX were searched by BLASTN.

	Individual ESTs	Singletons/Contigs
Total	11544	4845
Sequences with no hit in BLASTX	4421	1999
Sequences with a hit in BLASTX	7123	2846
Full-length sequences	3205	1330
Short full-length sequences	324	98
Possible full-length sequences	46	6
Partial sequences	3352	1317
Ambiguous sequences	45	19
Unknown	151	76
Full-length and coding region complete	463	342
Sequences with a hit in BLASTN	126	27
Sequences without a hit in BLASTX and BLASTN	4295	1972