Abstract

Motivation: Expressed Sequence Tags (ESTs) are short single-pass DNA sequences obtained from either ends of cDNA clones. To exploit these sequences efficiently, a dynamic Web-tool has been developed which uses these data to perform fast virtual cloning of cDNAs.

Results: Starting with a query sequence, the user is able to identify related ESTs and extend the sequence of interest step by step, possibly to a full-length transcript. Graphical views of the clustering are used to monitor the progress of a particular ‘cloning’ project. Potential open reading frames are detected by positional base preference, and hyperlinks to other Worldwide Web sites allow the user to retrieve information relevant to each EST in a cluster (e.g. sequence traces, clone size, plate position). Apart from cDNA cloning, this tool also provides a mechanism for collating gene families and polymorphism sites.

Availability: This tool is available at the following URL: http://www3.ebi.ac.uk/~gill/ESTblast

Contact: E-mail: rwg23944@ggr.co.uk; psl4446@ggr.co.uk

Introduction

With the advent of automated DNA sequencers, it became possible in the early 1990s to perform high-throughput end sequencing of cDNA clones. These relatively short cDNA sequences are called Expressed Sequence Tags (ESTs). This approach was first described in 1991 (Wilcox et al., 1991) and developed on a large-scale format by Venter’s group at the National Institutes of Health (NIH) (Adams et al., 1992). In 1994, Merck & Co. and Cambridge University started a collaboration to deposit EST sequences in a public database through GenBank (Benson et al., 1994; Williamson et al., 1995). The EST database, dbEST (Boguski et al., 1993), contains sequences from the Merck–WashU project, but also from smaller contributors, and includes not only human ESTs, but sequences from a variety of organisms. As of 16 December 1996, dbEST contained 412,992 human sequences from the Merck–WashU initiative out of a total of 775,636 sequences (539,706 human sequences). Twenty-seven normalized oligo-dT-primed cDNA libraries prepared from various tissues (Soares et al., 1994) have been used for the Merck–WashU project with both ends of the cDNA clones sequenced in the majority of cases. In addition, the length of many of the cDNA clones have been determined. Since cDNA clones belonging to same gene could be sequenced more than once and sequence information can start at different positions in the transcript (especially at the 5’ end), it is possible to rebuild and extend a cDNA clone by sequence assembly. Since a computer is used, the terms ‘in silico’ (Kalogeropoulos, 1993) or ‘virtual’ cDNA cloning have been used to describe this approach.

With the continuous growth of dbEST, in silico cloning is becoming one of the methods of choice to obtain cDNAs of interest when only limited experimental data are available. In fact, one server at the NCBI, called Unigene, is taking advantage of all the sequence information to present clusters of all human GenBank entries that are part of the same gene, including EST sequences (Boguski and Schuler, 1995). As of 6 January 1997, Unigene contained over 50,000 sets or clusters corresponding probably to a similar number of distinct human genes (http://www.ncbi.nlm.nih.gov/Schuler/Unigene/). Whilst chromosomal localization or peptide similarity information is available for some of the clusters in Unigene, the main information remains a list of accession numbers. The consensus sequence, the sequence length and the cluster organization are not accessible. The user has to rebuild the consensus sequence of the gene of interest independently. Unigene currently contains human sequences only and no clusters have been built for ESTs from other species.

While this tool was developed, the Tentative Human Consensus (THC) sequences created at TIGR (The Institute for Genomic Research) have become available at the NCBI Web server (http://www3.ncbi.nlm.nih.gov/cgi-bin/THCBlast/nph-thcblast). As of March 1997, the NCBI’s Blast server contained 38,000 THC sequences created by assembling ESTs into virtual transcripts. THCs are a good alternative to Unigene, but the graphical alignment of the clusters is lacking information when both ends of a cDNA clone have been sequenced or hyperlinks to trace files for example.

Because of the static format of Unigene and the drawback in the display of the clusters for the THCs, we have developed a new Web-based tool to perform fast and dynamic in silico
cloning that allows the user to create a graphical display of EST clusters with hyperlinks, multiple alignments and consensus sequences directly from the information in dbEST.

Algorithm

A flowchart (Figure 1) describes the principle of the tool. A query sequence is used as the starting point and pasted in the query box. The user can choose the type of Blast search required (Altschul et al., 1990) and section of the database to search: human ESTs only (default choice), mouse ESTs only or the complete dbEST. These database sections are updated on a daily basis by site-dependent protocols that do not impinge on this work. The split between the human and mouse ESTs is made as part of the daily updating, and is a mechanism for reducing the database search times. ESTs overlapping the query sequence can be selected by either clicking on each individual High-scoring Segment Pair (HSP) from the Blast output (marked up in html) or by selecting the number of HSPs to be kept from the output.

The Web server then returns a graphical display (Figure 2) of the HSPs for all the ESTs selected and their read-pairs—the sequences determined for the other end of the cDNA clones. The position of these sequences is determined from the cDNA clone length information present in dbEST. There is also an estimation of total contig length, when the data are available, giving a virtual map of the EST coverage overlapping, and with the query sequence. The graphical display is created by interrogation of flat files containing all EST information indexed on clone-id and accession number. The searching of these files is facilitated by use of the binary search algorithm (Knuth, 1972). These data plus positional data derived from the Blast search are then scaled and represented using hypertext-linked ASCII characters.

Three options are available at this stage: to use the hyperlinks to find out more about the clones, to search dbEST again, or to assemble the sequences of the set selected giving a consensus for the transcript. Usually, Blast is used again with a selected EST in an attempt to extend the contiguous region, and cycles of searches and contig display are repeated until no further ESTs are found. It is possible that some read-pair sequences have not been incorporated into the contiguous

Automated

User selected

EST Blast

Query sequence

Search Non redundant EST database

Select blast hits of interest

Retrieve information on read pairs of EST clone and estimates overall clone length. Calculates position of clone relative to HSP

Create Graphical overview

EST data viewer

Reblast any EST to extend coverage

Incomplete Create Consensus Complete Finished Consensus

Reblast

Clustering (UNIGENE)

Trace Data (WASH U)

Fig. 1. A flowchart depicting the typical procedure employed in EST Blast (or in silico cloner).
A new dynamic tool to assemble ESTs

Graphical view of EST blast output

(-) = unsequenced data
(+)= Blast search area
(=) = Known Sequence and position
() = Known Sequence unknown position (Clone size not available!)

---

**Fig. 2.** Graphical display of a contig after three cycles of Blast searching. The original query sequence was EST W07581 (dbEST accession number), shown in the centre of the alignment on the left side, which separates previous (upper) from the new and (lower) EST overlaps. In this cloning exercise, W07581 identified only N27805 in the first round in three contiguous HSPs, which accounts for this EST being shown three times. When dbEST was subsequently searched with the consensus sequence derived from W07581 and N27805, two more ESTs were identified: AA053610 (which was not picked up in the first cycle because its overlap with W07581 was too short) and H39620. The third cycle identified W31256 and AA053145 (which is the read-pair of AA053610). Underlined text marks hyperlinks: ‘==’ denotes regions where sequence data are available which are separated by ‘- -’ if no sequence information is available but the size of the cDNA clone is known, or by ‘...’ if the sequence is known but the size of the cDNA clone is not in the database. ‘No data’, as on the right-hand side of the contig in this example, denotes that no information (sequence, accession number) is available for this end of the cDNA. Hyperlinks are only applied to EST names or alignment positions once they have been confidently identified in a Blast search. For example the read-pair (N80811) of the initial query sequence (W07581) has still not been identified by Blast, indicating that the clone is longer than originally estimated and must constitute a substantial part of the parent mRNA, and that extra cycles are required. The template above the ESTs shows the estimated size of the ‘virtual’ transcript, with the region of sequence contig depicted by ‘++’. The ‘EB’ hyperlinks relaunch EST Blast with nearest left-hand or right-hand EST sequences. The checkboxes are used to mark which sequences are to be used for a sequence assembly and generation of a consensus (one checkbox is sufficient for each EST) before pressing the ‘create’ button to obtain a contig and a consensus sequence of the selected ESTs.

From the graphical view, ESTs can be selected and subjected to an automated sequence assembly by scripts using Staden’s Genome Assembly Package. Gel readings are added to the assembly ‘project’ beginning with the query sequence and then in the order that they appeared in the various Blast outputs. This ensures that the most similar sequences are aligned first, which increases the probability of producing a full-length consensus for the transcript. The alignments and consensus sequences (in html text areas where they may be edited if so desired) are returned to the screen, and the consensus sequence may be used either as the query of a Blast search or for detecting potential coding regions by viewing plots of positional base preference (Staden, 1990). The latter can also be helpful in identifying sequencing errors because the base preference has shifted from one frame to another.

Numerous hyperlinks are available at the different steps during an assembly project. If the EST sequence is derived from a Merck-WashU cDNA, it is easily possible to obtain, by clicking on the accession number, clone size, cDNA library information, the other end of the clone. Moreover, a hyperlink to Unigene can be used as an internal control for your own in silico project. The consensus sequence can be edited in the text box underneath the multiple sequence alignment. In order to facilitate the editing, hyperlinks to sequence-trace graphics (where available) are also encoded. Intermediate steps can be saved using a jot saving into a text editor.
Systems and methods

The 'in silico cloner' comprises a set of Perl 5.0 and Bourne shell scripts which call executables of Blastn (Altschul et al., 1990), Genome Assembly Package 96.0 (GAP 96.0) (Bonfield et al., 1995), Nucleotide Investigation Package (Staden, 1990) and Readseq (http://www.med.virginia.edu/achs/molbio/phy/lip/readseq.html). The default parameters for Blastn were used for searches. Sequence assembly was carried out using an initial match of 15 bases, accepting a maximum of 25 insertions in the contig consensus or new gel-reading sequences. GAP 96.0 provided the multiple alignment and preliminary consensus sequences which were filtered to exclude base calling errors by deleting consensus positions labelled with an asterisk. Positional base preference plots (Staden, 1990) were generated by a script using the default parameters of the Nucleotide Investigation Package, returning postscript as the output.

The software is accessed through the common gateway interface (cgi) of a Web server using a hypertext transfer protocol (http 1.0) resident initially on a Silicon Graphics Indy (running Irix 5.2), but later a Silicon Graphics Challenge Server. The user requires a networked computer on which any Web viewer (e.g. Netscape or Mosaic) may be used, although a MIME type should be set for interpreting postscript files.

Implementation

The initial development was carried out in the Glaxo-Welcome intranet. However, it is currently being ported to a Web site for public use (http://www3.ebi.ac.uk/~gill/EST-blast). Training information and examples are also available at the same web site. No client data are retained after processing, although user feedback is welcome.

Discussion and conclusion

We describe here a new tool to perform in silico cloning using the public information in dbEST. The main advantages of this dynamic tool are its speed and simplicity of use. The graphical view of the cluster greatly helps the user to monitor the progress of a virtual contiging project by aligning EST clusters from successive Blast searches. The manual selection of the ESTs at different stages allows a great flexibility of the tool. Blast is not species or gene dependent; hence, the in silico cloner also provides a mechanism for collating members of a gene family or finding specific gene polymorphisms.

In silico cloning is a very powerful way to utilize rapidly new sequence information from the public databases, but this approach has limitations. The user relies completely on the quality of the information in the databases. It is well known that chimaeric cDNA clones have been used to produce ESTs and that annotations for some entries could be wrong (Hillier et al., 1996). The normalization process of the cDNA libraries is always incomplete and therefore it is likely that rare transcripts are under-represented in dbEST. Moreover, from our personal experience, we found that intronic sequences and misprimed cDNAs were also present in dbEST. Also, the automated sequence assembly may give misleading results for several reasons: the Blast searches may have identified sequences from splice variants or other gene family members through conserved regions, or simply because the sequence base calling was poor. To ensure that the consensus sequences generated are correct, the user should go back to the bench to confirm their existence. This verification can be done rapidly using PCR-based techniques.

Future improvements will focus on the interactivity of the tool with a particular emphasis on the editing capabilities, such as correction of the frameshift errors, and providing additional hyperlinks, for example to EST mapping information and to BlastX searching to assist functional characterization of the transcript. We are also developing software additions which remove the requirement for specialized MIME types. Another possibility is the complete automation of the process by setting predefined rules for EST selection. If the user does not want to go through the different cycles, an option will be available where he can submit a query sequence and the tool would produce a graphical view and a consensus sequence where all the overlapping ESTs have been selected.

Acknowledgements

The authors are most grateful to Dr Steve Foord and Mr Simon Tate for helpful comments.

References

A new dynamic tool to assemble ESTs


Received on February 14, 1997; revised and accepted on April 15, 1997