dbEST — database for "expressed sequence tags"

Sir—The accumulation and analysis of partial, "single-pass" cDNA sequences ("expressed sequence tags", ESTs) has become an important component of genome research and has been the subject of numerous articles[1–6] and commentaries[7–9] in *Nature Genetics*. We would like to call your attention to a new database that has been created specifically to meet the unique informatics challenges posed by EST data.

ESTs are qualitatively and quantitatively different from traditional database entries in terms of accuracy, completeness and rate of acquisition. Automated "single pass" sequencing results in an ~3% error or base ambiguity rate and contamination with vector and other spurious sequences has been a problem. The average length of an EST is a function of current automated sequencing technology and is about 300–400 nucleotides. It is not generally known in advance if these fragments come from coding or non-coding regions of an mRNA, thus sequence characterization and annotation are minimal. Furthermore, up to 70% of ESTs have no recognizable homologues at the time of submission. Thus there is a continual need for periodic re-annotation by database similarity searching. Lastly, ESTs now represent the most rapidly expanding source of new human sequences: more than 14,000 human EST sequences (Table 1) have been released in the past two years compared with about 19,000 human sequences that have appeared in GenBank over the last decade or more. Below we describe the design and operation of a special database for ESTs as well as the specialized processing and annotation performed

Table 1 Current contents of dbEST

Organism	No. of sequences
Homo sapiens[a]	14,556
Caenorhabditis elegans	4,699
Arabidopsis thaliana	1,764
Oryza sativa (rice)	1,023
Plasmodium falciparum	339
Mus musculus, domesticus	119
Macropus eugenii	36
Total	22,537

[a]The majority of human sequences are from brain (63%), a lymphoblastoid cell line (16%) and liver (9%) with the remainder coming from adrenals, bone, kidney, placenta, testis, retina, skeletal muscle, heart muscle and rhabdomyosarcoma.

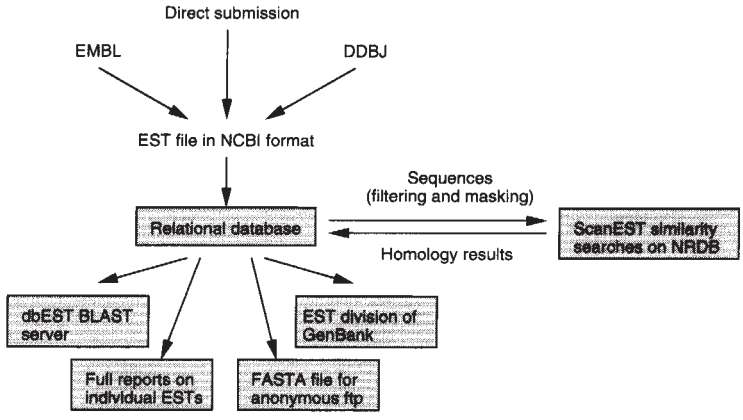

Fig. 1 Schematic of EST database and information processing. Special analyses and annotation are performed as follows: we use a software system, constructed in part from the BLAST function library (W. Gish, personal communication), to carry out homology searches, when the data are submitted, that supplement any homology information supplied by the author(s). Thereafter, we periodically update our homology annotation by searching the entire contents of dbEST against sequence database updates. Prior to searching, the ESTs are preprocessed to identify and mask problematic subsequences. The first "filtering" step involves comparing each EST against a small nucleotide database consisting of prototypic repetitive sequences[11] (such as Alu and L1) and common cloning vectors (such as pBluescript®). A slightly modified version of the BLASTP algorithm that works with nucleotide sequences is used for greater sensitivity than obtained with BLASTN. For each match with a Karlin-Altschul[12] score ≥ 160, the corresponding EST segment is masked (replaced by X's) and this fact is retained for the dbEST record. At this point, a nonredundant nucleotide sequence database (see below) is searched with the masked sequence using the program ScanEST that employs a more sensitive version of BLASTN as well as an expanded scoring matrix that allows proper weighting of ambiguous nucleotides common among EST sequences. For each EST, Karlin-Altschul scores and Poisson probability values for up to 25 of the most similar database sequences are saved and incorporated into the relational database. To compare the ESTs against the protein database, each sequence is conceptually translated in all six reading frames and then filtered using the Seg and XNU algorithms[13,14]. These programs detect regions of locally biased amino acid composition ("low complexity" subsequences such as polyQ) that can lead to hundreds of spurious matches in a database search and have forced some groups to use overly conservative scoring thresholds to avoid this problem. The resulting, filtered open reading frames (ORFs) are then compared to the protein database using a variant of the BLASTX algorithm[15]. Matches between EST ORFs and database proteins are then stored in dbEST. In addition to scanning the protein database, each filtered EST ORF is also compared to a special representative set of protein families that have been conserved across phyla for about 500 million years (T.M.J.L. *et al.*, manuscript in preparation). This information is intended to provide some degree of cross-referencing among different EST collections and may prove to be a useful measure of expression complexity as EST collections grow. The nonredundant databases of nucleotide and amino acid sequences are derived from GenBank, EMBL, DDBJ, PIR International, SWISS-PROT and the NCBI backbone databases by a method that will be described elsewhere (W. Gish, manuscript in preparation).

Acknowledgements
We thank Warren Gish for invaluable advice and assistance.

at NCBI.

EST data are usually submitted as a set of many sequences and we have designed a simple, tagged "flat file" format (available on request) to streamline the direct submission process. EST sequences and associated information may be sent electronically (via e-mail or Internet file transfer protocol) to NCBI; GenBank accession numbers are then issued. We also scan the daily updates from general submissions to GenBank and also from EMBL and the DNA Database of Japan (DDBJ) for EST data submitted to these locations (Fig. 1).

Once new data has been accessioned into dbEST, it is immediately made available for public access unless submitters specify that it be kept confidential until publication. Access and distribution are accomplished by four different mechanisms:

(i) Network and e-mail BLAST searches. EST sequences, preceded by header lines containing the NCBI-dbEST identification number, the cloning library name, the organism name and the best protein or nucleotide match (if any), are posted to the BLAST network and e-mail servers as a stand alone database for homology searching. (For information on using these services, send e-mail to blast@ncbi.nlm.nih.gov with the word 'help' in the body of the message.)

(ii) Full reports. Complete descriptions of EST sequences are available from est_report@ncbi.nlm.nih.gov (type 'help' in the body, leave subject line blank). The report contains complete annotation for the sequence(s), including publication (if any), contributor name and address, cloning library information, organism and tissue description, mapping information, putative homology assigned by the contributor, as well as information on significant similarities with other protein and nucleotide sequences generated by NCBI (see below). For human ESTs that have been deposited with the American Type Culture Collection (ATCC), the ATCC identification numbers are provided along with ordering information for the physical DNA clones. (An example of a full report is included in the e-mail 'help' document.) For physical DNA clones from other sources (for example, Généthon, the *C. elegans* sequencing consortium[6]) contact information is provided. We are also working with the Genome Data Base to coordinate our efforts in making available genetic map locations.

(iii) FASTA format. The sequences, with descriptive header lines as described in (i), are placed in the NCBI Data Repository for down-loading by anonymous ftp from ncbi.nlm.nih.gov.

(iv) New EST Division of GenBank. The dbEST sequences are available in the major releases and daily updates of the EST division of GenBank on CD-ROM and by anonymous ftp.

dbEST is not simply a static repository but attempts to keep the information about ESTs current by periodically performing homology searches against new data in GenBank and the protein sequence databases. Searches are carried out after "filtering" the queries to increase the sensitivity, to minimize non-specific or uninformative database matches and to avoid interference from potential contaminants (see Fig. 1). DNA and protein similarities are stored in dbEST and are available in the full reports along with the date of the most recent search.

EST data have many possible uses including the identification of previously unknown gene products for genetic mapping purposes and the study of the mechanisms of tissue differentiation and ontogeny by developing profiles of sequences that are differentially expressed in particular cell types or at specific developmental stages. Comparative sequence analysis has also provided provocative insights into protein evolution[10]. An immediate practical value of interest to a broad range of biomedical researchers is the accelerated cloning of human genes for which homologues in other organisms have already been functionally characterized. We believe that dbEST is a useful resource that will facilitate all of the these activities.

Mark S. Boguski
Todd M.J. Lowe
Carolyn M. Tolstoshev
National Center for Biotechnology Information,
National Library of Medicine,
National Institutes of Health,
Building 38A, Room 8N-805
8600 Rockville Pike,
Bethesda, Maryland 20894, USA

References

1. Adams, M.D., Kerlavage, A.R., Fields, C. & Venter, J.C. *Nature Genet.* **4**, 256–267 (1993).
2. Adams, M.D. *et al. Nature Genet.* **4**, 381–386 (1993).
3. Kahn, A.S. *et al. Nature Genet.* **2**, 180–185 (1992).
4. McCombie, W.R. *et al. Nature Genet.* **1**, 124–131 (1992).
5. Okubo, K. *et al. Nature Genet.* **2**, 173–179 (1992).
6. Waterston, R. *et al. Nature Genet.* **1**, 114–123 (1992).
7. Ballabio, A. *Nature Genet.* **3**, 277–279 (1993).
8. Editorial *Nature Genet.* **2**, 167–168 (1992).
9. Sikela, J.M. & Auffray, C. *Nature Genet.* **3**, 189–191 (1993).
10. Green, P. *et al. Science* **259**, 1711–1716 (1993).
11. Jurka, J., Walichiewicz, J. & Milosavljevic, A. *J. molec. Evol.* **35**, 286–291 (1992).
12. Karlin, S. & Altschul, S.F. *Proc. natn. Acad. Sci. U.S.A.* **87**, (1990).
13. Wootton, J.C. & Federhen, S. *Comp. Chem.* (in the press).
14. Claverie, J.-M. & States, D.J. *Comp. Chem.* (in the press).
15. Gish, W. & States, D.J. *Nature Genet.* **3**, 266–272 (1993).