EST analysis online: WWW tools for detection of SNPs and alternative splice forms

David Brett
dbrett@mdc-berlin.de

Gerrit Lehmann
glehmann@mdc-berlin.de

Jens Hanke
hanke@mdc-berlin.de

Stefan Gross
sgross@mdc-berlin.de

Jens Reich
reich@mdc-berlin.de

Peer Bork*
bork@embl-heidelberg.de

Max Delbrueck Center (MDC) for Molecular Medicine, Robert-Rössle-Strasse 10, Berlin-Buch, 13125, Germany.
*EMBL, Meyerhofstr 1, 69012 Heidelberg, Germany.

Is my mRNA alternatively spliced? Does it contain coding single nucleotide polymorphisms (SNPs)? Is it preferentially expressed in a disease tissue? Is it mapped to a chromosome? Does it have a close family member or a homologue in another species. The answer to all of these questions could come from expressed sequence tags (ESTs).

ESTs were originally used for gene identification[1] and were subsequently found to be extremely useful as a mapping tool[2]. Furthermore, tissue distributions of genes have been derived from EST library comparisons[3] and more recently, ESTs have been used to study both the frequency of human SNPs and alternative splicing[4–7]. This article introduces some World Wide Web (WWW) tools that allow a single gene to be analysed with respect to information contained in ESTs.

SNP analysis

The WWW SNP finder tool

To the typical user working on a single gene or gene family, finding alternative splices or novel SNPs was a slow and fragmented process until recently. Clustering overlapping ESTs, filtering possible pseudogenes or paralogues, and checking trace files were performed by hand. To speed-up the process, and allow an end user to review a gene quickly for the possibility of an alternative splice form or cSNP, we have developed a package of EST tools with simple cut-and-paste WWW interfaces. This tool can be found at http://mahe.bioinf.mdc-berlin.de/home.html

Accuracy of candidate SNP predictions

Locating a candidate SNP within ESTs is always a question of accuracy. The first step is to align the ESTs to a query gene sequence (cDNA or EST itself). We align ESTs using BLASTN (Ref. 8) with an 'expect value' of 1e-30. The aligned ESTs are then filtered such that only those with over 95% identity over 100 bp are retained. This helps to avoid processed pseudogenes and close homologues. Although it is possible that these values will allow the inclusion of a small fraction of false positives, too strict a filtering procedure would remove many correct ESTs and with them, many valid SNPs. In practice, later evaluation of the sequence quality of the EST trace files is a strong enough filter system to exclude most false positives.

If a change in base-pairs (polymorphism) occurs when a number of ESTs are aligned, the polymorphic site must be validated in some way. The quality of many ESTs, particularly at the beginning and end of the sequence, is poor. The initial filtering thresholds during the alignment process help to eliminate such poor-quality sequences. The SNP finder tool automatically retrieves existing EST trace files. Individual base-pair calling at polymorphic sites are validated with the sequence trace file program Phred (Ref. 9). Using a Fourier transformation method, Phred can derive the quality of a base in a sequence from a trace file. Phred gives a value for each base, ranging from 1 to 60. This value should be seen on a logarithmic scale. So the Phred threshold of n = 30 (set by default) means a error probability of 1/1000. Our group and others have employed such strategies to study SNP frequency in the human genome[4,5]. A further step taken by one of these groups was to employ an 'at least two different ESTs' rule[5]: in other words, where several ESTs match the sequence, at least two of them must have the same polymorphic change that is different to the majority of ESTs that match the sequence at that point. In a study carried out by our group, 74% of SNPs with two or more different ESTs were confirmed by resequencing[4].

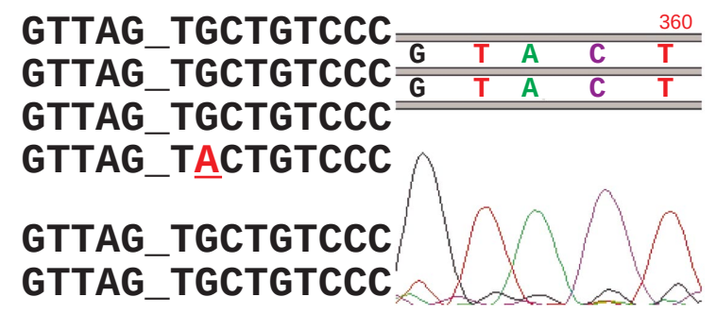
The WWW tool displays all aligned ESTs so that the number of polymorphic ESTs can be easily identified; the SNP finder is used to display the EST alignment file, with the location of the SNP highlighted in red. The corresponding trace file with Phred output scores is also displayed (Fig. 1). A general help file gives the background information and takes you through the procedure step by step.

Alternative splicing

The WWW AS finder tool

Alternative splicing (AS) is an important mechanism allowing tissue-specific or temporal expression of a novel gene form. AS allows one pre-mRNA to be processed into many different mature forms within a cell, each of which can have distinct functions. AS has also been shown to be specifically associated with disease phenotypes[10,11]. To identify candidate AS forms in ESTs, we created a simple-to-use WWW tool (http://mahe.bioinf.mdc-berlin.de/home.html). A protein or mRNA sequence is cut and pasted in a FASTA format and the AS finder tool runs a BLAST search using the dbEST subsection using only human ESTs. The resulting file is searched for matching ESTs with possible AS forms. We recommend using mRNA as query sequence rather than protein when given a choice; in our own studies, mRNAs gave more information than did protein sequence.

FIGURE 1. The SNP finder tool: an example



The picture on the left was taken from the results files of the SNP finder tool. It shows a candidate SNP, a G residue to an A (highlighted in red, EST W86556), detected from aligning ESTs to the PACT kinase mRNA (gi 3290197). The picture on the right shows the corresponding trace file result. Abbreviations: EST, expressed sequence tag; SNP, single nucleotide polymorphism.

Accuracy of candidate AS sites

A filtering system is employed to address problems arising from pseudogenes and close homologues. The BLAST results are filtered by default at 95% identity over 100 bp or 30 amino acids. Using check boxes, the user can change these values. The AS finder tool compares the start and end of each high-scoring sequence alignment between query mRNA and EST. The program also filters out AS candidates with internal protein repeats that can be confused in the BLAST result file as possible AS forms. It looks for loss or addition of sequence present in query and not in the EST or vice versa. Once a difference is located and the filter criteria are met, the particular BLAST result is highlighted in red and the result reported (Fig. 2). The user is given the information in table form with type of AS form (insert or deletion), the position of the AS form in query and EST, and the percentage identity of the EST match. As with the SNP finder, from the table of results 'at least two different ESTs' rule can also be applied, giving greater confidence to the prediction. In association with this tool, we have created a database of over 2747 AS forms from 1797 published human mRNAs and SWISS-PROT disease-associated proteins where candidate AS were detected and reported in the same manner.

Limitations of the tool

Evidently, the WWW tool can only highlight candidate AS forms. Some of the EST splice forms will represent mutations leading to truncated proteins. We have found such examples in our own studies, and we are now studying these in relation to disease phenotypes. At present, our tool makes no distinction between splice forms leading to truncated proteins and splice forms leading to in-frame changes in amino acid sequence. However, a simple check of the reading frame can identify stop codons. There may be contamination of the database with pre-mRNA or genomic DNA, giving rise to false positive inserted intronic sequence. There are also a number errors of annotation in the EST collections. For this reason, we recommend resequencing the EST clone to make sure the correct sequence corresponds to stated EST accession number. We normally follow this by a multiple tissue cDNA expression check to identify both forms.

Other databases

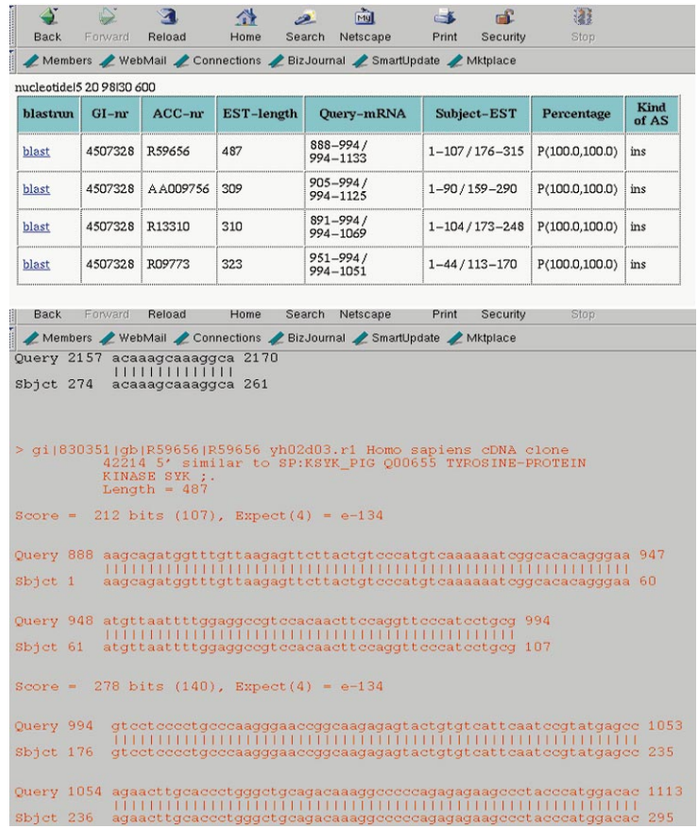
There are several public-access SNP databases and AS databases. At present, both tools have links to these databases so that a user can check an SNP or alternative splice form against those already published. Useful sites include: HGBASE (http://hgbase.interactiva.de/); dbSNP (http://www.ncbi.nlm.nih.gov/SNP/); ASDB (http://cbcg.nersc.gov/asdb); the database for our own group (ftp://ftp.bioinf.mdc-berlin.de/Pub/database/SPLICE_SITE/MRNA/humrna.html); and the Max Delbrueck Center (MDC) bioinformatic group homepage, which includes both tools described here plus other bioinformatic resources (http://mahe.bioinf.mdc-berlin.de/home.html).

Conclusion

One of the first priorities on completion of the human genome is to extract the coded genes from the raw genomic sequence. Primarily, this information will be provided by random ESTs and cDNA production. The percentage of AS found within this EST data will have a direct effect on the final number of human genes identified in the genome. In recent studies carried out by our group and others, 35% of genes examined by EST matching contained at least one alternative splice form[6,7]. Indeed, this is probably an underestimate as EST coverage is estimated to only 30% of all exons[4]. In addition, exons contained completely within published introns are not matched by mRNAs.

The tools described here should allow an easy exploitation of up-to-date EST information to reveal SNPs and AS in human genes. The accuracy is sufficient to study the candidates experimentally for various applications including SNP haplotype prediction and novel AS forms implicated in disease progression. The tools have simple cut-and-paste interfaces, direct access to public SNP and AS databases, plus helpful pointers to explain and evaluate the results.

FIGURE 2. The alternative splice finder tool: an example



A results file of the alternative splice finder tool is shown at the top of the figure for the Syk kinase mRNA (gi 4507328). Four ESTs are listed as containing a candidate alternative splice form. The 'Query-mRNA' column shows the mRNA sequence pieces aligned by the BLAST program; here, the ranges (e.g. 888–994/994–1133) indicate no gap. By contrast, in the 'Subject-EST' column, an insertion of 69 bp is indicated (e.g. 1–107/176–315). The corresponding BLAST results file is shown at the bottom of the figure. The matches with the ESTs containing candidate AS forms are shown in red. The figure shows the corresponding example from row 1 in the table above. Note the query and subject endings of the first alignment and the query and subject start of the second. The inserted AS form in Syk kinase has been detected in a basophilic leukaemia cell line[12]. Abbreviations: AS, alternative splicing; EST, expressed sequence tag; SNP, single nucleotide polymorphism.

References
1 Adams, M.D. *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 21, 1651–1656
2 Schuler, G.D. *et al.* (1997) Pieces of the puzzle: expressed sequence tags and the catalogue of human genes. *J. Mol. Med.* 75, 694–698
3 Boguski, M.S. and Schuler, G.D. (1995) ESTablishing a human transcript map. *Nat. Genet.* 10, 369–371
4 Sunyaev, S. *et al.* (1999) Prediction of nonsynonymous single nucleotide polymorphisms in human disease-associated genes. *J. Mol. Med.* 77, 754–760
5 Picoult-Newberg, L. *et al.* (1999) Mining SNPs from EST databases. *Genome Res.* 9, 167–174
6 Hanke, J. *et al.* (1999) Alternative splicing of human genes: more the rule than the exception. *Trends Genet.* 15, 383–427
7 Mironov, A. *et al.* (1999) Frequent alternative splicing of human genes. *Genome Res.* 9, 1288–1293
8 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
9 Ewing, B. *et al.* (1998) Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* 8, 186–194
10 Klamt, B. *et al.* (1998) Frasier syndrome is caused by defective alternative splicing of WT1 leading to an altered ratio of WT1+/− KTS splice isoforms. *Hum. Mol. Genet.* 4, 709–714
11 Qi, M. and Byers, P.H. (1998) PH Constitutive skipping of alternatively spliced exon 10 in the ATP7A gene abolishes Golgi localization of the menkes protein and produces the occiptal horn syndrome. *Hum. Mol. Genet.* 7, 465–469
12 Yagi, S. *et al.* (1994) Cloning of the cDNA for the deleted syk kinase homologous to ZAP-70 from human basophilic leukemia cell line (KU812). *Biochem. Biophys. Res. Commun.* 200, 28–34

Repbase Update
a database and an electronic journal of repetitive elements

Jerzy Jurka
jurka@charon.girinst.org

Genetic Information Research Institute, 1170 Morse Avenue, Sunnyvale, CA 94089-1605, USA.

Interspersed repetitive elements from eukaryotic genomes represent inactive copies of contemporarily or historically active retroelements and DNA transposons known collectively as transposable elements (TEs). Retroelements, which include retroviruses and nonretroviral elements, reproduce via reverse transcription and genomic integration. By contrast, DNA transposons use a 'cut and paste' mechanism to take advantage of chromosomal duplication. Most TEs are nonautonomous; that is, their proliferation is controlled by a small number of retroelements or DNA transposons with intact protein machinery. This has important implications for understanding patterns of repeat insertions and the evolution of eukaryotic chromosomes[1]. Indeed, the majority of non-protein-coding chromosomal DNA, including most introns[2–4], might have been produced by waves of repeat insertions. A separate category of repetitive DNA that contributes to genomic structure and evolution is represented by tandem repeats of diverse origin, including telomeric repeats, centromeric satellites and interspersed simple repeats also known as mini- and microsatellites[5].

TABLE 1. The current content of Repbase Update (April 2000)

Type of repeat families	File name	Number of (sub) families
Human (primate)	humrep.ref	443
Processed pseudogenes (human)	pseudo.ref	28
Alu (primate)	humsub.ref	19
Rodent	rodrep.ref	252
Other mammalian	mamrep.ref	128
Zebrafish	zebrep.ref	4
Other vertebrate	vrtrep.ref	97
Caenorhabditis elegans	celrep.ref	96
Drosophila	drorep.ref	133
Other invertebrate (animal)	invrep.ref	192
Arabidopsis thaliana	athrep.ref	206
Other plant	plnrep.ref	111
Simple (microsatellites)	simple.ref	131
Total		**1840**
Unique		**1661**

A comprehensive resource

As genomic sequencing continues to accelerate[6–10], there is a growing demand for a comprehensive resource of sequence data and other basic information about TEs. The first such resource was established in 1992 and contained representative sequences and sequence fragments of 53 published human families of interspersed repeats[11]. It continued to grow and became widely known as 'Repbase'. In 1997, Repbase was succeeded by 'Repbase Update' (RU)[12] which, in addition to compiling known elements, began the electronic publication of TEs unreported elsewhere. (A comprehensive description of all original data systematically published in RU by invited contributors is beyond the scope of this short note, and will be reviewed elsewhere.)

To increase the involvement of the research community in organizing the rapidly expanding sequence data of TEs, a new, electronic, peer-reviewed journal is due to be launched later this year (follow web announcements at http://www.girinst.org). This companion publication will be released with each monthly issue of RU but, unlike RU, it will remain unchanged to preserve a permanent record of all original contributions, which can be referenced in the scientific literature.

There are 1661 unique families and subfamilies of TEs in the current release of RU (Table 1). This is an increase from 956 two years ago[12], and the number of families represented in RU is expected to double over the next 2–3 years. Each family ranges from less than one hundred to over a million elements per genome. Furthermore, some older families can be shared by the genomes of many different species. For example, all mammals share MIR elements that are at least 100 million years old. In addition to the files listed in Table 1, RU contains appendix files (humapp.ref, rodapp.ref, etc.) that serve only as archives documenting the history of sequence and annotation improvements over time.

Anatomy of a sequence entry

RU is available electronically (http://www.girinst.org) as a flat file in EMBL and FASTA formats, both in compressed (*.tar.gz), and uncompressed ASCII versions, and is updated monthly. Some of the software-specific versions will also be added in the near future.

An example of a recent entry for a SINE sequence called 'MIR3' is shown in Fig. 1. The name 'MIR3' (ID) reflects the sequence similarity to tRNA-related MIR elements and to autonomous L3 elements, which were discovered earlier. The original names