 Nucleic Acids Research, 2000, Vol. 28, No. 1 **191–192**

ExInt: an Exon/Intron database

M. Sakharkar, M. Long[1], T. W. Tan and S. J. de Souza[2,*]

Bioinformatics Center, National University of Singapore, Singapore, [1]Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA and [2]Laboratory of Computational Biology, Ludwig Institute for Cancer Research, Sao Paulo branch, Rua Prof. Antonio Prudente 109, 4 andar, 01509-010 Sao Paulo, Brazil

Received August 31, 1999; Revised and Accepted October 17, 1999

ABSTRACT

The Exon/Intron (ExInt) database incorporates information on the exon/intron structure of eukaryotic genes. Features in the database include: intron nucleotide sequence, amino acid sequence of the corresponding protein, position of the introns at the amino acid level and intron phase. From ExInt, we have also generated four additional databases each with ExInt entries containing predicted introns, introns experimentally defined, organelle introns or nuclear introns. ExInt is accessible through a retrieval system with pointers to GenBank. The database can be searched by keywords, locus name, NID, accession number or length of the protein. ExInt is freely accessible at http://intron.bic.nus.edu.sg/exint/exint.html

INTRODUCTION

Rapid progress in genome research has produced and will continue producing an unprecedented amount of nucleotide sequence data. The exponential growth of the Expressed Sequence Tag (EST) (1) database (dbEST), for example, has contributed to the identification and mapping of the whole set of human genes, which is believed to range from 80 000 to 100 000. The mapping of ESTs still provides information about the exon/intron organization of the corresponding gene. Programs like GRAIL (2) and GENESCAN (3) have also contributed to the understanding of the organization of the human genome. Because of the accumulation of sequence data, information on the exon/intron organization of eukaryotic genes is becoming widely available. However, the retrieval of this information, particularly on a large scale basis, is a difficult task. In the present report, we present a database of all intron-containing genes from eukaryotes present in GenBank. This exon/intron database, which we call ExInt, collects information about the exon/intron organization of eukaryotic genes present in GenBank and organizes the data in a retrieval form available on the WWW.

Since there is an error rate in the prediction of the intron positions by computer programs, we decided to create subsets of ExInt, one containing all entries where the introns were predicted and another one containing ExInt entries where the introns were characterized by experiments. ExInt was also divided into two other independent subsets containing the entries corresponding to organelle and nuclear genes.

METHODOLOGY

The ExInt database is built as a text file in FASTA format. We have used GenBank release 113 to construct a raw database containing all eukaryotic sequences with an exon/intron organization. We first identified all the GenBank entries that had the word 'DNA' in the identification line (ID). We then searched for the word 'intron' and used the 'cds.....join' and/or the 'cds.....complement join' lines to identify introns interrupting the corresponding coding region. Introns interrupting the 5′ UTR and 3′ UTR are also shown in the final ExInt entry. We also used those lines to calculate the phase and position of the introns as well as the number and size of the corresponding exons. Entries containing predicted introns were identified by searching ExInt for the words 'cosmid', 'BAC', 'PAC' or 'chromosome'. By searching the 'ORGANISM' line in the GenBank entry for words like 'mitochondrion', 'chloroplast' and 'plastid', we identified all entries containing organelle introns.

DESCRIPTION OF THE DATABASE

Each entry in ExInt contains the following information:

- Locus: locus name.
- Description: description as in GenBank.
- Accession number and NID: accession number and NCBI ID.
- Introns: phase and position for introns interrupting the coding sequence.
- Exons: number, size and length (in amino acids) of the protein encoding exons.
- Nucleotide position for the introns: (i_1.....j_1, i_2.....j_2,.....,i_n....j_n) where i is the first nucleotide and j is the last nucleotide of the intron.
- Genome: organelle or nuclear.
- Protein sequence: one letter amino acid sequence for the protein.

The ExInt database constructed from GenBank release 113 has 51 651 entries. Entries containing experimentally-defined introns correspond to 44.3% of ExInt. The number of intron positions for the whole database is 235 590 which gives an average of 4.56 introns per entry. Intron phase distribution for nuclear introns is as follow: 48.6% of phase 0 introns, 28.9% of phase 1 introns and 22.5% of phase 2 introns. Similar numbers were found by Long *et al.* (4), using an earlier release of GenBank.

*To whom correspondence should be addressed. Tel: +55 11 270 4922; Fax: +55 11 270 7001; Email: sandro@compbio.ludwig.org.br

WWW ACCESS AND AVAILABILITY

ExInt is accessible via a WWW interface (http://intron.bic.nus.edu.sg/exint/exint.html). Queries can be performed by locus name, NID, accession number or length of the protein. The user can choose to search the whole database, the predicted intron database, the experimentally-defined intron database, the organelle intron database or the nuclear intron database. Users can also search these five databases with a query sequence using BLAST (5). The result of all these searches will be an object report from ExInt including appropriate annotation and sequence information among other features. By clicking on hyperlinks the user can easily navigate to connected objects or other databases. All five databases are also available for downloading as text files. ExInt will be updated within 20 days of a new release from GenBank.

FINAL COMMENTS

ExInt can be of potential use for biologists studying genomics and gene evolution as well as for molecular biologists interested in a specific gene family. It should be taken into consideration, however, that ExInt does not contain intron-lacking genes. Therefore, ExInt users must be cautious when deriving overall conclusions about the structure and evolution of eukaryotic genes.

SUPPLEMENTARY MATERIAL

Supplementary figures are available at NAR Online.

REFERENCES

1. Adams,M., Dubnick,M., Kerlavage,A.R., Moreno,R., Kelley,J.M., Utterback,T.R., Nagle,J.W., Fields,C. and Venter,J.C. (1992) *Nature*, **355**, 632–634.
2. Uberbacher,E.C. and Murad,R.J. (1991) *Proc. Natl Acad. Sci. USA*, **88**, 11261–11265.
3. Burge,C. and Karlin,S. (1997) *J. Mol. Biol.*, **268**, 78–94.
4. Long,M., Rosenberg,C. and Gilbert,W. (1995) *Proc. Natl Acad. Sci. USA*, **92**, 12495–12499.
5. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.