Nucleic Acids Research, 2002, Vol. 30, No. 1 **191–194**

ExInt: an Exon Intron Database

M. Sakharkar[1], F. Passetti, J. E. de Souza, M. Long[2] and S. J. de Souza*

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

Received September 19, 2001; Accepted September 26, 2001

ABSTRACT

The Exon/Intron Database (ExInt) stores information of all GenBank eukaryotic entries containing an annotated intron sequence. Data are available through a retrieval system, as flat-files and as a MySQL dump file. In this report we discuss several implementations added to ExInt, which is accessible at http://intron.bic.nus.edu.sg/exint/newexint/exint.html.

INTRODUCTION

The exponential growth of sequence databases, especially due to genome and EST sequencing, has generated a parallel increase in the amount of sequences showing an intron/exon organization. We have recently developed a database containing all sequences in GenBank bearing in their annotation at least one exon/intron boundary (1). This, and other related databases (2,3), has been used in several studies approaching issues related to the exon/intron organization of eukaryotic genes (4,5).

In this report, we describe a series of implementations to the Exon/Intron Database (ExInt) as follows:

1. Relational database: data are now stored in a relational database (MySQL). The table structure is presented in Figure 1. Data from the database tables can be downloaded in a dump format, which allows direct incorporation in other MySQL relational databases.
2. Purged database: it is known that GenBank is extremely redundant. To avoid any potential bias, we have made available in this latest version of ExInt a non-redundant set of the data. Overall analysis of both redundant and non-redundant sets confirmed that most of the sequences (>80%) are redundant in current databases. Both datasets are available for download as Fasta libraries. They are also searchable using ExInt Blast engine.
3. Statistics link: several statistical features (for the whole database and models species) are available, such as number of genes, exons and introns before and after purging (Table 1); exon length distribution (Fig. 2); intron length distribution (Fig. 3) and intron phase distribution (Table 2).
4. Validation of predicted gene structure using EST data: we provided a validated subset for genes predicted in seven species: *Homo sapiens*, *Mus musculus*, *Rattus* sp., *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana* and *Saccharomyces cerevisae* (Table 3).

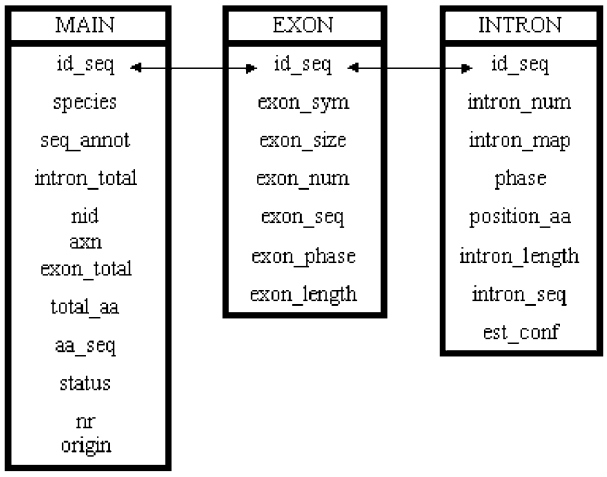

Figure 1. Description of the ExInt relational database.

Table 1. Gene, exon and intron number for whole ExInt and subdivisions

	Gene number	Exon number	Intron number
Whole ExInt	94 615	518 169	525 870
Non-redundant ExInt	15 271	113 457	128 065
Rattus norvegicus	835	4889	7191
Homo sapiens	8287	60 499	43 127
Mus musculus	3044	18 920	15 407
Drosophila melanogaster	15 220	64 271	89 969
Caenorhabditis elegans	18 924	121 708	108 803
Arabidopsis thaliana	25 216	158 629	127 386
Saccharomyces cerevisiae	589	1695	1438

METHODOLOGY

We have used GenBank release 122 to construct a raw database containing all eukaryotic sequences with an exon/intron organization. The approach used to identify all intron-containing sequences in GenBank has been described previously (1). The same is true for the methodology used to construct the following derived databases: predicted introns, experimentally defined introns, organelle and nuclear genes (1). A purged database was constructed using a modification

*To whom correspondence should be addressed. Tel: +55 11 32074922; Fax: +55 11 32077001; Email: sandro@compbio.ludwig.org.br

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

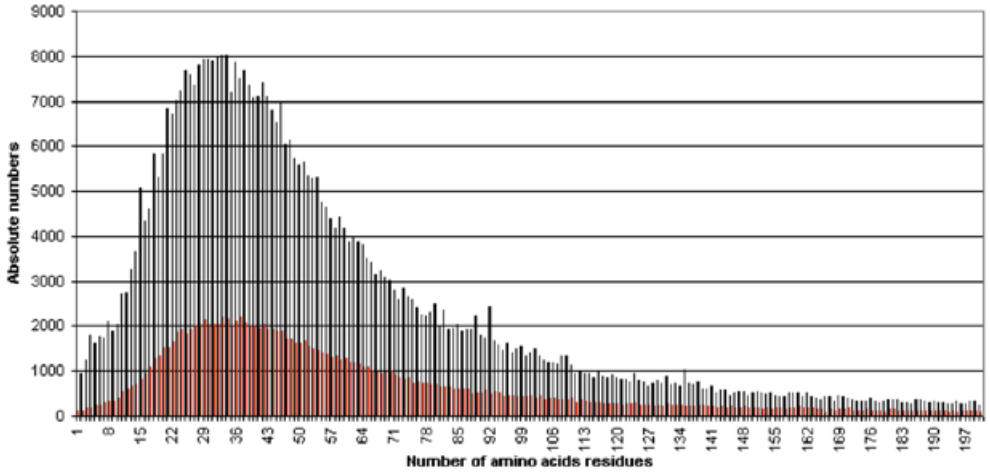

Figure 2. Exon size distribution. The complete database is shown in black, a non-redundant set is shown in red.

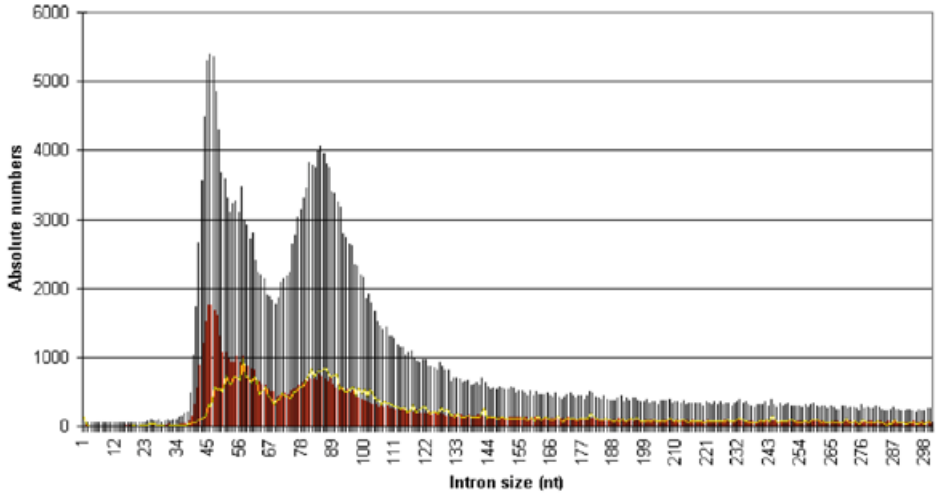

Figure 3. Intron size distribution. The complete database is shown in black, a non-redundant set is shown in red. The yellow line corresponds to experimentally defined introns.

Table 2. Intron phase distribution

	0	1	2
All ExInt	257 713 (49%)	147 625 (28%)	120 532 (23%)
Non-redundant	60 979 (48%)	35 438 (28%)	31 608 (24%)
Rattus norvegicus	2842 (39%)	2365 (33%)	1384 (28%)
Mus musculus	6703 (44%)	5921 (38%)	2783 (18%)
Caenorhabditis elegans	51 251 (47%)	28 553 (26%)	28 999 (27%)
Homo sapiens	19 102 (44%)	15 423 (36%)	8602 (20%)
Arabidopsis thaliana	71 958 (56%)	28 178 (22%)	27 250 (22%)
Drosophila melanogaster	38 101 (42%)	28 896 (32%)	22 972 (26%)
Saccharomyces cerevisiae	641 (45%)	428 (30%)	369 (25%)

of the method of Long *et al.* (6), as follows. We performed an all-against-all protein sequence comparison using a PVM-version of Fasta in an eight-node cluster of PCs running Linux. When two protein sequences have an identity level ≥25% over at least 70% of the length of the shorter sequence, just one sequence is kept. These comparisons are exhaustive until a complete non-redundant database is obtained. As a representative of the gene cluster we have taken the sequence with the largest number of exons and introns.

To validate the predicted gene structures, we take the predicted cDNA structure (keeping the positional information of all predicted introns) for all genes within seven model species and used Blast (7) to search them against the respective (same species) EST datasets. A script in PERL was written to parse the Blast output looking for cases where a predicted exon/exon boundary (by that we mean a region in the cDNA where a predicted intron is present at the genomic level) was confirmed by at least one EST.

RESULTS AND DISCUSSION

ExInt contains a wealth of relevant biological information. Here, we present some statistics that are important to the database construction and for a general evaluation of the data.

Table 3. Predicted introns confirmed by EST

	GenBank ID with predicted introns	GenBank ID with confirmed predicted introns	Predicted introns	Number of ESTs	Predicted introns confirmed by ESTs
Rattus norvegicus	23	10	183	273591	31 (17%)
Mus musculus	137	73	1704	1 296 332	389 (23%)
Caenorhabditis elegans	3016	2283	100 977	58 367	17 454 (17%)
Homo sapiens	1852	1149	23 235	3 406 430	6013 (26%)
Arabidopsis thaliana	1592	1438	125 567	112 999	31 873 (25%)
Drosophila melanogaster	703	542	52 639	116 099	10 278 (20%)
Saccharomyces cerevisiae	317	38	1024	11 159	38 (4%)

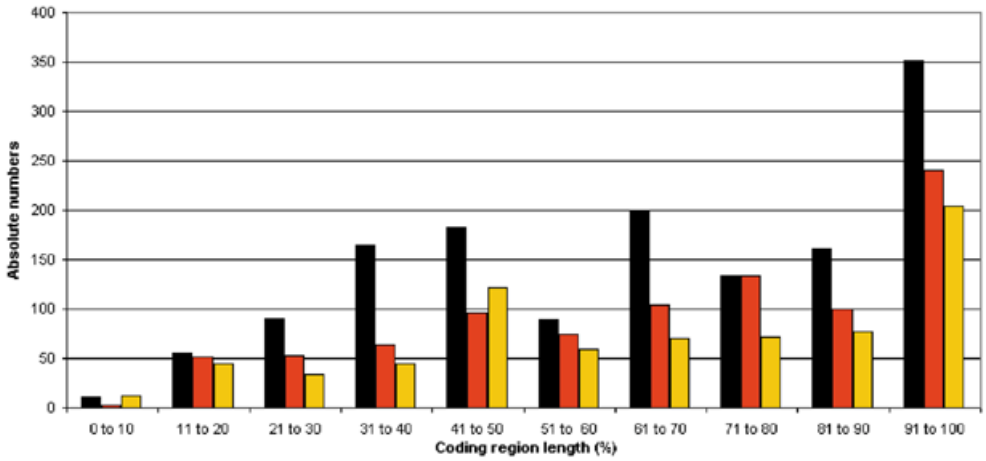

Figure 4. Intron phase distribuition along the cds. Black, introns phase 0; red, introns phase 1; yellow, introns phase 2.

Table 4. Frequency of exon symmetry

	0.0	0.1 + 1.0	1.1	1.2 +2.1	2.2	0.2 + 2.0
Whole ExInt	111 959	97 398	37 923	50 644	24 475	92 348
Non-redundant	26 878	25 491	9729	14 004	7290	24 803
Rattus norvegicus	497	448	1474	1017	62	1855
Mus musculus	3037	3037	2264	1547	470	2189
Caenorhabditis elegans	20 422	20 814	7009	11 898	7022	21 448
Homo sapiens	8552	8989	5289	5072	1645	6581
Arabidopsis thaliana	35 951	22 991	5623	9216	5245	24 420
Drosophila melanogaster	11 898	15 756	6821	10 083	4967	13 691
Saccharomyces cerevisiae	99	84	27	79	24	86

Table 1 shows the number of genes, exons and introns for the redundant and non-redundant datasets and for seven model species. We note that there are, on average, 5.48 exons per gene with AL445795 having the higher number (96). Figures 2 and 3 show the exon and intron length distributions, respectively. We confirm an observation from Deutsch and Long (8) that invertebrate introns are on average smaller than human introns. As also seen by Deutsch and Long (8), we have observed a bimodal distribution of intron length for the whole dataset, which does not seem to be due to predicted introns, since the same pattern is also observed for the confirmed introns (Fig. 3). Positioning of introns along the coding region (Fig. 4) shows a bias distribution towards the C-terminal half of the protein molecule. This piece of information is important for interpretation of data related to gene structure. For example, it has recently been suggested that alternative

splicing events are more frequent on the C-terminal half of proteins (9), a bias that can be due to the distribution shown in Figure 4.

The validation of predicted gene structures is probably the most important implementation to ExInt. It has been shown that gene prediction programs may generate a large amount of artefactual gene structures, and analysis using these datasets may draw incorrect conclusions (10). We have made use of the large amount of EST data available in dbEST to validate the predicted gene structure for sequences of seven different model species, *H.sapiens*, *M.musculus*, *Rattus* sp., *C.elegans*, *D.melanogaster*, *A.thaliana* and *S.cerevisae*. This validation step creates a sub-set of 'trusted' predicted gene structure that may be important in a number of biological queries. The sub-set of validated intron/exon boundaries may also constitute a useful resource for developers of gene prediction programs. It is important to emphasize that the absence of validation does not imply that the predicted gene structure is wrong, since the coverage of the transcriptome by ESTs is not yet complete.

AVAILABILITY

ExInt is accessible via a World Wide Web interface at http://intron.bic.nus.edu.sg/exint/newexint/exint.html. Different features can be used as a query element such as: NID, locus name and keyword. The whole database, as well as derived databases, is available for download. Derived databases include: purged database, predicted intron, experimentally defined introns, organelle genes and nuclear genes. Users can also search all databases with a query sequence using Blast. ExInt will be updated twice a year.

ACKNOWEDGEMENT

F.P. is supported by Fapesp (00/02228-9).

REFERENCES

1. Sakharkar,M., Long,M., Tan,T.W. and de Souza,S.J. (2000) ExInt: an Exon/Intron database. *Nucleic Acids Res.*, **23**, 191–192.
2. Saxonov,S., Daizadeh,I., Fedorov,A. and Gilbert,W. (2000) EID: the Exon–Intron Database—an exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Res.*, **28**, 185–190.
3. Schisler,N.J. and Palmer,J.D. (2000) The IDB and IEDB: intron sequence and evolution databases. *Nucleic Acids Res.*, **28**, 181–184.
4. Sakharkar,M., Kangueane,P., Woon,T.W., Tan,T.W., Long,M., Kolatkar,P.R. and de Souza,S.J. (2000) IEKb—an exon intron knowledge base from databases. *Bioinformatics*, **16**, 1151–1152.
5. Sakharkar,M., Tan,T.W. and de Souza,S.J. (2001). Generation of a database containing discordant intron positions in eukaryotic genes (MIDB). *Bioinformatics*, **17**, 671–675.
6. Long,M., Rosenberg,C. and Gilbert,W. (1995) Intron phase correlations and the evolution of the intron/exon structure of genes. *Proc. Natl Acad. Sci. USA*, **92**, 12495–12499.
7. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
8. Deutsch,M and Long,M. (1999) Intron–exon structures of eukaryotic model organisms. *Nucleic Acids Res.*, **27**, 3219–3228.
9. Modrek,B., Resch,A., Grasso,C. and Lee,C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.*, **29**, 2850–2859.
10. Rogic,S., Mackworth,A.K. and Ouellette,F.B. (2001) Evaluation of gene-finding programs on mammalian sequences. *Genome Res.*, **409**, 685–690.