

## **IE-Kb: intron exon knowledge base**

Meena K. Sakharkar<sup>1,\*</sup>, Pandjassarame Kanguane<sup>1</sup>, Tong W. Woon<sup>1</sup>, Tin W. Tan<sup>1</sup>, Prasanna R. Kolatkar<sup>1</sup>, Manyuan Long<sup>2</sup> and Sandro J de Souza<sup>3</sup>

<sup>1</sup>Bioinformatics Centre, National University of Singapore, Singapore 119260, Singapore, <sup>2</sup>Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, USA and <sup>3</sup>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 on June 12, 2000; revised on July 31, 2000; accepted on August 8, 2000

### **Abstract**

**Summary:** *IE-Kb (Intron Exon-Knowledge base) illustrates the intron–exon dynamics in eukaryotic genes. We have developed three different knowledge sets, namely ‘Non-redundant ExInt’, ‘Non-redundant Pfam-ExInt complement’ and ‘Non-redundant GenBank eukaryotic subdivisional sets’ to understand this phenomenon. Statistical analysis is performed on each knowledge set and the results are made available online. The entries in knowledge sets are ranked based on their intron length, exon length and protein length with relational hyper-links to the corresponding intron phase, intron position, intron sequence, gene definition and parent GenBank entry.*

**Availability:** <http://intron.bic.nus.edu.sg/iekb/iekb.html>

**Contact:** [meena@bic.nus.edu.sg](mailto:meena@bic.nus.edu.sg)

GenBank contains more than 3.4 billion nucleotide bases from over 55 000 different organisms and doubles every 15 months (Benson *et al.*, 2000). The available knowledge in GenBank feature table has significantly improved the accuracy of computational gene-finding algorithms (Stormo, 2000). A normal mathematical formalism is often inadequate to develop models describing complex gene structures. Description of intron–exon architecture in genes to trace the connecting control elements for their functions in protein products is of interest. The structure of an eukaryotic gene described by assembling parameters (exons and introns) identified as predefined combination rules depends on the available knowledge and the established relationships between them. The combination rules required for describing gene structures are exponential and might exceed our realization that is sustained by taking together all available data. Eukaryotic genome data mining has helped to identify a few such rules (Deutsch and Long, 1999; Tomita *et al.*, 1996; Long *et al.*, 1995; Hawkins, 1988).

The manic growth in GenBank and the speculated complexities associated with intron–exon organization in eukaryotic genes provide impetus to systematically analyze genome data. Recently, ExInt (Sakharkar *et al.*, 2000), EID (Saxonov *et al.*, 2000) and IDB (Schisler and Palmer, 2000) were developed using GenBank data and they all try to describe intron–exon organization in eukaryotic genes. ExInt (<http://intron.bic.nus.edu.sg/exint/exint.html>) provides a WWW interface to study intron–exon structure in eukaryotic genes. This interface helps to navigate through intron–exon properties at specific locations in a protein sequence providing information on intron numbers, exon numbers, intron length, intron phase, exon length and gene definition. The attractive features in ExInt are the relational hyperlinks to the corresponding intron sequence and the parent GenBank entry. It should be also noted that EID provides flat files with extraction programs and IDB is available with a Macintosh based interface.

To further explore ExInt, we created a non-redundant ExInt using GenBank 116 data. To remove duplicate entries in ExInt we purged (Long *et al.*, 1995) the data at 100% identity level. We named the non-redundant ExInt as knowledge set 1. If introns are maintained in a dynamic steady state by the process of insertion and deletion then the statistical association of intron size and phase with the corresponding protein length will be of interest. Statistical analysis is performed for every entry in knowledge set 1 using parameters such as intron length, exon length, intron numbers, and exon numbers. Every entry in the knowledge set is ranked according to intron length, exon length and protein length in a table format. Each value in the table is hyperlinked to the corresponding ExInt entry. The observations from knowledge set 1 are: (1) Distribution of phase 0, 1, and 2 introns is 49.0, 28.4 and 22.6% respectively. The results are in accordance with earlier reports (Tomita *et al.*, 1996; Long *et al.*, 1995).

\*To whom correspondence should be addressed.

(2) Nearly 90 introns were present for every 100 exons. (3) The mean intron length is  $463.0 \pm 1885.5$  (SD). (4) The mean exon length is  $218.2 \pm 216.0$  (SD). (5) The standard deviation (1885.5) about the mean intron length explains the extreme variation in their length. (6) The *Homo sapiens* chromosome 8, Zinc finger gene (GenBank AC AF178030) and *Caenorhabditis elegans* ankyrin gene (GenBank AC U39847) was found to contain the longest intron and exon respectively.

To re-discover similar intron–exon properties, we created knowledge set 2 by a different approach using GenBank 116 data. GenBank complements of EMBL (Baker *et al.*, 2000) nucleotide entries that are cross-referenced to TrEMBL and SWISS-PROT (Bairoch and Apweiler, 2000) datasets were extracted for each Pfam-5.1 entry (Bateman *et al.*, 2000) for the construction of Pfam-ExInt complement set. The Pfam-ExInt complement set contains sequences that are duplicates within a protein family. Hence, we purged (Long *et al.*, 1995) the data at 100% identity level. Therefore, knowledge set 2 contains the ExInt complement for unique protein sequences in Pfam-5.1. The statistically important parameters such as phase-distribution, mean intron length, mean exon length, and the standard deviation about the mean are emphasized (details are available online). Thus, one can study the structure of genes whose protein sequences are characterized in Pfam.

To emphasize the dynamics of introns and exons in eukaryotic genes within GenBank subdivisions, we created knowledge set 3. Non-redundant knowledge sub-sets for the six eukaryotic GenBank sub-divisions were constructed by removing redundancy at 100% identity level. The sequence redundancy ranges from approximately 10–21% within sub-sets (details are available online).

The most interesting feature of IE-Kb is the systematic ranking of its content based on intron length, exon length and protein length. This approach will aid in discovering novel and critical parameters that best describe gene structures in known data. All derived knowledge on introns and exons stored in IE-Kb is based on the

‘cds...join’ statement in GenBank feature table. Hence, IE-Kb does not include information on single exonic genes and introns in the 5′ and 3′ UTRs. It should be noted that partial genes, alternatively spliced variants and pseudogenes are not clustered separately in IE-Kb. In future, we desire to archive information that helps in understanding the evolutionary aspects of introns and exons under IE-Kb.

## References

- Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
- Baker,W., Vanden,B.A., Camon,E., Hingamp,P., Sterk,P., Stoesser,G. and Tuli,M.A. (2000) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **28**, 19–23.
- Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.
- Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2000) GenBank. *Nucleic Acids Res.*, **28**, 15–18.
- Deutsch,M. and Long,M. (1999) Intron–exon structures of eukaryotic model organisms. *Nucleic Acids Res.*, **27**, 3219–3228.
- Hawkins,J.D. (1988) A survey on intron and exon lengths. *Nucleic Acids Res.*, **16**, 9893–9908.
- 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.
- Sakharkar,M., Long,M., Tan,T.W. and de Souza,S.J. (2000) ExInt: an exon/intron database. *Nucleic Acids Res.*, **28**, 191–192.
- 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.
- Schisler,N.J. and Palmer,J.D. (2000) The IDB and IEDB: intron sequence and evolution databases. *Nucleic Acids Res.*, **28**, 181–184.
- Stormo,G.D. (2000) Gene-finding approaches for eukaryotes. *Genome Res.*, **10**, 394–397.
- Tomita,M., Shimizu,N. and Brutlag,D.L. (1996) Introns and reading frames: correlation between splicing sites and their codon positions. *Mol. Biol. Evol.*, **13**, 1219–1223.