Exon-Intron Structure Database: a tool for intron analysis.

I.V. Poverennaya[1,2],

[1] *Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia;*

[2] *Vavilov Institute of General Genetics RAS, Moscow, Russia;* ipoverennaya@gmail.com,

D.D. Gorev[3], V.V. Yacovlev[4,5], M.A. Roytberg[3,4]

[3] *Moscow Institute of Physics and Technology, Moscow, Russia;*

[4] *Institute of Mathematical Problems of Biology RAS, Pushchino, Russia;*

[5] *Higher School of Economics, Moscow, Russia;*

gorev.d@phystech.edu, v.yakovlev@gmail.com, mroytberg@lpm.org.ru

Analysis of exon-intron structures of eukaryotic genes is essential for many biological fields such as evolutionary genomics, gene engineering, molecular biology, etc. Unfortunately, it is complicated by imperfectness of sequencing data and, thereby, gene annotation problems. To handle this, several separate curated databases about genes and its elements in different organisms had been developed (e.g., EID[1]). However, due to the absence of updates, the wide expansions of sequencing data, and the new raised questions in the gene evolution, these databases are mainly out of the date. The most recent ones are JuncDB[2] and PIECE 2.0[3], but the first one contains only orthologous gene isoforms and lacks the majority of intron data, and the second one focuses only on plants. The primary goal of our new exon-intron gene structure database (EIS DB) is to provide an easy access of comprehensive data of well-annotated genes in more than 100 eukaryotic genomes from different taxonomic groups to the scientists who study gene or intron evolution, and alternative splicing.

As the main source of gene sequences and annotations, we have used RefSeq[4] genome assemblies, which versions were current to March 2017. The corresponding gbk-files were downloaded from ftp-server of NCBI[5]. In addition, for each organism we also prepared extra input files with data about common names, taxonomy, chromosome number, etc. In total, there are 112 eukaryotic organisms in the current EIS DB version that embraces such taxonomic categories as Vertebrates (birds, fishes, mammals, etc.), Invertebrates (insects, roundworms), Plants, and Fungi. To parse gbk-files, load and verify the obtained transformed

data into EIS DB, a special software written on C++ has been developed. It works directly with the database control system.

EIS DB is a relational database managed by PostgreSQL. Structurally, it contains 15 tables. The main ones are ‘Organisms’, ‘Genes’, ‘Isoforms’, ‘Exons’ and ‘Introns’. The others contain auxiliary data, e.g., taxonomy and intron types. The gene isoforms (and the corresponding exons and introns) that failed the verification are still kept in database, but they are labeled and available only by special request. Besides, the database includes fasta-files with sequences, references to which are present in ‘Gene’, ‘Exon’ and ‘Intron’ tables.

Unlike earlier developed exon-intron structure databases, EIS DB allows working with either full set of isoforms or one ‘canonical’ isoform (the isoform with the biggest exon-intron structure or with the longest protein product).

As intron section of database is quite detailed – there are data about intron length, phase, sequence, splice sites, etc. – it makes EIS DB especially appealing for studying various intron features in different organisms such as phase vs length correlations, non-canonical splice sites distribution and etc.

The web interface of EIS DB will be soon publicly available. Apart from web-version, the database could be downloaded to user’s PC for more advanced requests.

1. Shepelev, V. & Fedorov, A. Advances in the Exon-Intron Database (EID). *Brief. Bioinform.* **7,** 178–185 (2006).
2. Chorev, M., Guy, L. & Carmel, L. JuncDB: an exon-exon junction database. *Nucleic Acids Res.* 44**,** D101-9 (2016).
3. Wang, Y. *et al.* PIECE 2.0: an update for the plant gene structure comparison and evolution database. *Nucleic Acids Res.* 45**,** 1015–1020 (2017).
4. NCBI RefSeq Database. http://www.ncbi.nlm.nih.gov/refseq/
5. https://ftp.ncbi.nlm.nih.gov/genomes/