Briefings in Bioinformatics Advance Access published March 29, 2006

doi:10.1093/bib/bbl003

Advances in the Exon–Intron Database (EID)

Valery Shepelev and Alexei Fedorov

Abstract

Investigation of exon–intron gene structures is a non-trivial task due to enormous expansions of the eukaryotic genomes, great variety of gene forms, and the imperfectness in sequence data. A number of available informational systems on various gene characteristics complement each other and are indispensable for many genomic studies. Among them, the Exon-Intron Database (EID) is a good choice for large-scale computational examination of exon/intron structure and splicing. It has many internal filters that control for sequence quality, consistency of gene descriptions, accordance to standards, and possible errors. New innovations in EID are described. The collection of exons and introns has been extended beyond coding regions and current versions of EID contain data on untranslated regions of gene sequences as well. Intron-less genes are included as a special part of EID. For species with entirely sequenced genomes, species-specific databases have been generated. A novel Mammalian Orthologous Intron Database (MOID) has been introduced which includes the full set of introns that come from orthologous genes that have the same positions relative to the reading frames. Examples of statistical analyses of gene sequences using EID are provided. We present the latest data on our comparison of intron positions in 11 025 orthologous genes of human, mouse and rat, and find no convincing cases of intron gain. We discuss relevant data-quality issues of genomic databases. In particular, 5% of genes in genomic databases contain internal stop codons. This fact is due to a combination of biological reasons and also to errors in sequence annotations. The EID is freely available at www.meduohio.edu/bioinfo/eid/.

Keywords: *exons; introns; splicing; genomics; bioinformatics; computational biology*

INTRODUCTION

The arrangement of exons and introns in split genes is an extensive field of investigation, which has produced many amazing discoveries and intriguing questions. To facilitate these studies, many advanced databases describing exon and intron sequences have been created. These databases present diverse information on the genes and also provide web-based interfaces for quick and simple access to the data. Popular information systems on the Internet that characterize exon–intron structures are the following: Entrez Gene [1], Ensembl Genome Browser [2], UCSC Genome Browser [3], SpliceNest [4], Xpro [5], ISIS [6], ExInt [7] and others, a fraction of which can be found in Galperin's biological database repository [8]. In addition, several databases are specifically devoted to alternative splicing: ASD [9], SpliceInfo [10], ECgene [11], EASED [12]. Here, we describe important improvements in our Exon–Intron Database (EID), publicly available since 2000 [13]. The primary goal of EID is to offer a comprehensive and convenient dataset of sequences for computational biologists who study exon-intron gene structures and pre-mRNA splicing. The main advances in the recent EID releases include: (i) generation of species-specific databases for those organisms whose genomes have been entirely sequenced and annotated; (ii) addition

Corresponding author: Alexei Fedorov, Department of Medicine and Program in Bioinformatics and Proteomics/Genomics, Medical University of Ohio, Toledo OH 43614, USA. Tel: (419)-383-5270; Fax: (419)-383-3102; E-mail: afedorov@meduohio.edu

Valery A. Shepelev, PhD is a Research Scientist in the Department of Bioinformatics, Institute of Molecular Genetics, Russian Academy of Sciences. His research interests include understanding principles of genome organization and molecular evolution of primates, development of biomolecule primary structure databases and elaboration of algorithms for sequence analysis and DNA–protein interaction.

Alexei Fedorov is an Assistant Professor and Head of Bioinformatics Laboratory in the Department of Medicine, and Program in Bioinformatics and Genomics/Proteomics at Medical University of Ohio. His research interests include investigation of evolution and roles of exon–intron gene structures in different organisms.

of sequences for untranslated regions (UTRs) in exons; (iii) addition of intron sequences located outside protein coding regions (CDS); (iv) addition of genes that do not have intron(s) inside their CDS, but only in their UTRs (novel UTR Intron Database, or UID); (v) addition of intronless genes (novel Intron-Less Database, or ILD); (vi) inclusion of information about alternatively spliced gene isoforms, which is available from GenBank genomic annotations and (vii) generation of novel Mammalian Orthologous Intron Database (MOID). Currently, species-specific sets of all introns and all exons are available for human, mouse, rat, dog, chicken, zebrafish, fruit fly, worm (*Caehorhabditis elegans*), and mouse-ear cress (*Arabidopsis thaliana*). This list will be extended on a monthly basis in accordance with GenBank updates. EID is freely available at www.meduohio.edu/bioinfo/eid/ and the documentation about the described EID innovations are presented in the 'README_Sept05' file on the EID website. EID is currently available as a flat-file, fasta-formatted database. While this version has thus far proven convenient for large-scale bioinformatics analysis, the authors plan to extend the usability and range of scientific questions that can be investigated with this database. To this end, a web-accessible, relational database version of EID is under construction, which will provide a more user-friendly interface and enable researchers to make dynamic queries on various aspects of EID. Both versions of EID will be subsequently maintained and updated, giving the researcher flexibility when analyzing this specialized dataset.

Here, we describe novel features of EID, demonstrate examples of EID usage, and illustrate several problems that current genomic databases pose to researchers who study gene structures.

EID CONTENTS

Novel genomic exon–intron database

The new version of EID consists of eight files described in Table 1. The name of each file contains information about the species and GenBank release it was generated from [14], while the file extension shows the type of data. For instance, mm34p1.dEID presents the database for *Mus musculus* prepared from GenBank Build 34.1, and contains the 'DNA-form' of gene sequence representation. In this format, exon sequences are shown in upper case and introns in lower case, as described by Saxonov and co-authors [13]. In addition to the previously described dEID, pEID and hEID formats, newer releases contain five novel file types. Three of these: mrnaEID, exEID and intrEID, present sequences of mRNA, individual exons and individual introns, respectively. The informational fasta-formatted line in these files is the same as in the dEID file, with the exception that in exEID and intrEID files, the consecutive number of exons or introns is shown at the beginning of this line. The main statistics for exons and introns in EID are summarized in the files with extension 'sEID'. The content of this file is demonstrated in Table 2 for human, mouse and rat genomes. Finally, the file with extension 'tEID' represents the technical records containing data from the toolkit computations.

Table I: Characterization of EID files

File extension name	Description of the file
dEID	Fasta-formatted database of gene sequences as described in Saxonov *et al*. [13]
pEID	Fasta-formatted database of protein sequences as described in Saxonov *et al*. [13]
hEID	Fasta-formatted database of header information as described in Saxonov *et al*. [13]
mrnaEID	New fasta-formatted database of mRNA sequences
exEID	New fasta-formatted database of exon sequences
intrEID	New fasta-formatted database of intron sequences
tEID	New technical file containing full report on the construction of the EID
sEID	New file with main statistics on the current version of EID (Table 2)

Information line

Following is an example of a single information line from the human dEID, or mrnaEID files. Due to its length this single line is wrapped into multiple lines in this example.

>30A_NT_077913 protein_id:NP_057260.2; Homo sapiens chromosome 1 genomic contig./ gene = "Cab45"; intron(phase:u21110,size:2945, 4499,474,4316,135,769,intr_sum:13138); exon(size: 110,479,137,114,159,176,780, ex_sum:1955); {spli ce:gtag,gtag,gtag,gtag,gtag,gtag}; CDS_start = 3209, CDS_end = 14490, CDS_len = 1089

This line starts with the EID serial number of the gene (30 in this instance). The optional capital letter(s) after the serial number shows that there are several alternative isoforms in the GenBank Feature

Table 2: General statistics on mouse, rat and human exons and introns provided from the EID files mm34pl.sEID, rn3pl.sEID and hs35pl.sEID

Description of data	Mouse	Rat	Human
I General			
Total number of gene blocks in GenBank	27 097	25 620	26 773
Total number of protein-coding genes	24 888	22 624	23 630
Total number of protein-coding genes having intron(s) within CDS region	20 127	19 146	20 342
Total number of genes without alternative splicing	19 551	19 100	17 903
Total number of genes with alternative splicing	576	46	2439
Total number of alternatively spliced isoforms	1339	97	6638
Total number of overlapped protein-coding genes in EID	367	101	563
II Problematic genes			
Number of genes with stop codons inside CDS (for genes with alternative splicing only the case when all isoforms have stop codons inside CDS counts)	976	1038	833
Number of CDS starting not from ATG codon	409	246	282
Number of genes with invalid/unidentified codon(s)	115	425	10
III Exons and introns			
Total number of introns (for genes with AS only one isoform with maximum number of introns counts)	181 865	185 689	189 191
Total number of exons (for genes with AS only one isoform with max number of exons counts)	201 992	204 835	209 533
Number of non-canonical introns (non TG . . . AG termini for genes with AS only one isoform with max number of non-canonical introns counts)	3052	4372	3233
Number of (AT . . . AC) introns (for genes with AS only one isoform with max number of AT . . . AC introns counts)	211	260	218
Number of extra-short introns (‹30 bp; for genes with AS only one isoform with maximal number of extra-short introns counts)	22	101	47
Number of extra-long introns (›100 000 bp; for genes with AS only one isoform with maximal number of extra-long introns counts)	760	802	1262
Number of introns with unidentified ends	368	392	46

AS = alternative splicing.

Table records for this gene. Each isoform in our database has a unique letter code starting from 'A', and continuing as follows: {A, B, C, . . ., Z, AA, AB, . . ., etc.}. The order of genes in the EID strictly follows those of GenBank, and therefore corresponds to the physical order of genes in chromosomes. For the human EID, gene presentation proceeds as: chr1, chr2, . . ., chr22, which is followed by chrX and chrY. Thus, neighboring genes in the genome always have consecutive numbers in EID. Following the EID gene number and underscore character ('_'), is the name of the contig to which this gene belongs (NT_077913 in this instance). This is followed by the protein identifier (NP_057260.2), the species and chromosome information, and the common gene name (Cab45 in this case). These data are taken from the corresponding GenBank record. Information about intron phases, intron sizes (in nucleotides), total size of all introns, exon sizes, total size of all exons and splice sites (as described previously [13]) are given as well. The first line may additionally include five optional tags at the very end, as listed in Table 3.

Table 3: Description of optional tags in the informational line

Optional tags	Description
STOP_CODON	in-frame stop codon encountered
UTR_AMBI	UTR is ambiguous since several suitable mRNA found
UTR_NF	UTR not found since no suitable mRNA found
CDS_incomplete	GenBank reports that current CDS annotation represents only part of coding region
PSEUDO	CDS feature in GenBank has a /pseudo tag

UTR introns

The new version of EID contains introns that are outside of CDS regions and that disrupt the UTRs of genes. These are denoted 'UTR introns'. UTR introns do not have a phase, yet in the intron phase records these introns are denoted as 'u' (unidentified). In the example fasta-formatted informational line shown earlier, the intron phase record is 'phase:u21110'. This indicates that the first intron is in the 5′-UTR gene region. Since the new version

of EID contains UTR sequences, the beginning, end and total length of the CDS is included at the end of the information line (CDS_start = 3209, CDS_end = 14490, CDS_len = 1089).

The protein format and heading format of EID (pEID and hEID, respectively) contain the information line as described in Saxonov and others [13].

UTR-intron Database (UID)

Traditionally, EID has presented genes that possess introns within their coding regions. There exists, however, a subset of genes that have only UTR introns and that do not have a single intron interrupting their CDS regions. These genes are stored in a separate UID. UID consists of the same eight files described in Table 1, yet these have 'UID' in their extension rather than 'EID'. Differentiation between 5′- and 3′-UTR introns is performed by appending a hyphen ('-') to the end or beginning of the phase description, respectively. For example, 'phase:uu-' means that this gene has two introns in its 5′-UTR, while 'phase:-u' means that this gene has a sole UTR intron in its 3′-end. Additionally, 'phase:uuu-uu' means that this gene has three introns in the 5′-UTR and two introns in the 3′-UTR. The current release of this database contains 1404 human genes, 1857 mouse genes and 796 rat genes.

Intron-less Database (ILD)

Finally, we created a database for intron-less genes that contains all genes without introns. Due to the nature of this database, the file intrILD would have no data and therefore is absent, while exILD, mrnaILD and dILD contain the same intron-less sequence. Consequently, there are five file types in this database, with the extensions 'dILD', 'pILD', 'hILD', 'sILD', and 'tILD'. The current release of this database contains 1760 human genes, 2939 mouse genes and 2,683 rat genes.

Original version of EID

We continue to generate updates of the original versions of the EID, representing introns from all species, as described in Saxonov *et al.* [13]. This version of EID is constructed based on the individual records from the following GenBank files: gbinv*N*.seq, gbmam*N*.seq, gbpln*N*.seq, gbpri*N*.seq, gbrod*N*.seq, and gbvrt*N*.seq, where '*N*' represents a number indicating a part of the database. The current release of this database (gb149EID) contains all innovations described above for genomic versions of EID. It also consists of the eight files shown in Table 1.

Mammalian Orthologous Intron Database (MOID)

Based on the genomic EID, we created the MOID, comprising human, mouse and rat sequences. We define 'orthologous introns' as introns from orthologous genes that also have the same position relative to the two coding sequences. Since there were no cases of intron gain and only solitary cases of intron loss in mammals [15], orthologous introns most likely descended from the corresponding intronic sequence of the last common ancestor for the taxon. The primary goal of MOID is to identify conserved functional motifs or non-coding genes inside introns. An example of successful utilization of MOID for the characterization of mammalian snoRNA genes has been demonstrated [16].

For MOID generation we imitated the authors of the Clusters of Orthologous Groups (COG) Database [17], and used the best hit (BeT) approach to define orthologous genes. Every protein sequence of species *X* was compared with all protein sequences of species *Y*, and vice versa, using the program blastp [18]. If gene A_X of species *X* matched protein B_Y of species *Y* in both comparisons (*X* versus *Y* and *Y* versus *X*), we treat them as orthologous. To compare intron positions in orthologous genes we used the program CIP.pl [19]. MOID contains three tables of orthologous introns: (i) mouse and human (file MOID9.05_Mm_Hs includes 116 746 orthologous intron pairs); (ii) human and rat (file MOID9.05_Rn_Hs, 107 843 orthologous intron pairs) and (iii) rat and mouse (file MOID9.05_Rn_Mm, 110 650 orthologous intron pairs). These three files represent tables of identifiers for orthologous introns taken from the 'intronic' form of EID (intrEID). Each line in this table represents a pair of orthologous intron identifiers. An example of one line from the mouse–human MOID is shown below.

```
INTRON_1 7184_NT_025741
INTRON_1 10787_NT_039491
```

It demonstrates that the first intron of the mouse gene with EID identifier 7184_NT_025741 is orthologous to the first intron of the human gene 10787_NT_039491. Finally, we generated a table of orthologous introns for three species (file MOID9.05_Hs_Mm_Rn, representing 87 843

triplets of orthologous introns of human, mouse and rat). These triplets correspond to the simplest orthologous triangle patterns according to Tatusov and co-authors [20].

Statistics for genomic EID

The pertinent statistics for the genes in our database are presented in the files with extension 'sEID', 'sUID', and 'sILD', respectively. An example of the content of 'sEID' files for human, mouse and rat is shown in Table 2. It consists of three sections: (A) general information about genes in GenBank genomic records; (B) possible problematic issues and (C) statistics for exons and introns. Section B shows that there are 833 human, 976 mouse and 1038 rat genes with internal stop codons. The most frequently asked question about EID that we have received is essentially, 'What are the possible reasons for the high occurrence of internal stop codons?' Indeed, 4–5% of mammalian genes have stop codons inside their CDS. This value is at least one order of magnitude greater than expected on the basis of the declared sequencing error rate of the human genome (1 per 100 000 bases) according to the Human Genome Consortium [21]. All of these genes with internal stop codons are annotated in the GenBank Feature Table. The annotation commonly used in GenBank for such cases is (/note = 'overriding stop codons') or (/note = 'unclassified translation discrepancy'). Internal stop codons occur with the same frequency in intron-containing genes as in intronless genes. Three alternatives potentially responsible for the appearance of internal stop codons are: (i) natural biological reason(s), (ii) sequencing/annotation errors and (iii) pseudogenes treated as real genes in GenBank. Among biological reasons, it could be either RNA-editing [22, 23]; stop codon suppression that makes them translated as sense codons (also known as 'translational readthrough') [24, 25]. A recent large-scale bioinformatic study of RNA editing in the human transcriptome showed that RNA editing was predominantly in intronic and intergenic regions, while few editing sites were found in translated exons [23]. On the other hand, while the read through of stop codons has been described in individual mammalian genes [26, 27], it seems that this process is infrequent in this taxon, according to the literature. Finally, there are only 108 human intron-containing pseudogenes in the GenBank Feature Table which could not explain the observed 833 genes with internal stop codons. Consequently, biological reasons are unlikely the sole cause of hundreds of internal stop codons in mammals. Investigation of 15 genes which were obtained randomly from the 833 sample, showed that in most cases multiple internal stop codon appearance is accounted for invalid annotation in the human genome Build 35.1 and is absent in RefSeq mRNA records. These annotation errors include the following: (i) invalid CDS start (two thirds of cases); (ii) extra nucleotides in CDS causing frameshift and (iii) invalid annotation for exon/intron structure.

EXAMPLES OF EID USAGE

Flat-file databases represent structured texts written in ASCII code. Minimal programming skills are required for working with such databases. We make publicly available all of the programs that we have created for EID data mining (all of them are written in PERL). For example, the SNO.pl program from our group which was used for the detection of evolutionary conserved structures of snoRNA genes in mammalian introns [16], is available at our website (http://www.meduohio.edu/eid). A more sophisticated PERL pipeline package which was used for the comparison of 6 million human EST sequences with all human exon and intron sequences from EID in order to find splicing abnormalities [28], can be found at: http://www.meduohio.edu/bioinfo/software.html.

Recently, EID has been used for a number of computational projects worldwide, such as: the characterization of sequence information required for the splicing of human pre-mRNA [29]; splice site modeling and prediction [30]; generation of a structural exon database (SEDB) showing exon boundaries on protein multiple alignments [31]. Exon and intron sequences from EID were used for several statistical analyses as well. These include the examination of: (i) DNA periodicities in exons [32]; (ii) the distribution of AG and GT dinucleotides near 5′- and 3′-ends of exons [33]; (iii) periodicities of dinucleotides in the vicinity of exon/intron junctions and their correlation with nucleosome positions [34]; (iv) biased codon usage near intron-exon junctions [35] and (v) statistics of intron phase in the vicinity of signal peptide cleavage site [36]. Also, EID was used as a source of data in the study of introns and splicing elements in fungi [37]; in testing hypotheses of correlation between exons and protein domains [38]; in the explanation of intron phase bias

A

```
/gene="Phox2b"
CDS join(9904549..9904789,9905639..9905826,9906748..9907024, 9907101..9907339)
```

B

```
9906961 ggcggagggc ccagcccagc cggagctccg ggggcggcgg gcccgggggg cccgggaggc
9907021 gaacccggca agggcggggc agctgcggcg nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn
9907081 nnnnnnnnnn nnnnnnnnnn ccggcaaggg cggggcggct gcggcggctg ctgcagcggc
9907141 ggcggcggca gcggcggctg cggccgcggc ggcgggaggc ctggctgcgg ccgggggccc
```

C

```
                                  *
     ggaggcgaacccggcaagggcggggcagctgcggcgnnnnnnnn 5'-end of intron
               ::::::::::::::::: :::::::::
     nnnnnnnnnnccggcaagggcggggcggctgcggcggctgctgc 3'-end of intron
913  ggaggcgaacccggcaagggcggggcggctgcggcggctgctgc mRNA (XM_344239)
```

Figure 1: Detailed investigation of the exon–intron structure in the rat homeobox 2b gene. (**A**) Information on exon-intron positions in the rat gene from the Feature Table of the GenBank file 'rn_ref_chr14.gbk', contig NW_047425, Build 3.1. The third rat intron located in the region 9907025–9907100 is absent in mouse and human orthologues. (**B**) Sequence fragment of the rat contig NW_047425 from the GenBank file 'rn_ref_chr14.gbk' that contains the intron examined. 'Intron' is underlined, repetitive sequences in the vicinity of this intron are shown in italic and bold. (**C**) Sequence alignment of homeobox 2b gene sequences in the vicinity of the third rat intron and its mRNA transcript. Single mismatch within repetitive region is shown with asterisk (*) above this position.

by codon usage bias [39]; in an algorithm construction for recognition of short exons in human genes [40]; and in the analysis of splicing acceptor sites [41].

All in all, the latest improvements in the EID and novel MOID that are presented here, are designated to facilitate computational investigations of genes and improve the quality of mining this data.

Examination of intron gain illustrates problems in intron-sequence quality

Despite knowing millions of split gene structures in dozens of species, the process of intron gain and loss during the course of evolution still remains a mystery [42]. There are alternative views on the processes of intron acquisition and loss, most recently discussed in Rogozin and co-authors [43] and in Roy and Gilbert [44]. Scientists still doubt whether or not new introns have been acquired during the evolution of mammals. In 1998, O'Neill and colleagues [45] experimentally characterized a novel intron inside the SRY sex-determining gene of marsupials that is absent in other mammals. However, a detailed analysis of the marsupial SRY gene shows that immediately downstream of the new intron position, the gene loses any similarity with homologs from other species. Therefore, alternative explanations are possible for the appearance of this intron. We would rather interpret this event as a complex rearrangement (gene fusion) and not as a simple insertion of a novel intron. A large scale analysis of 10 000 orthologous introns of human, mouse and rat from the EID database resulted in the absence of any intron acquisition events [46]. On the other hand, last year, using the database of evolutionary distances, Veeramachaneni and Makalowski [47] reported a novel intron in the rat homeobox 2b. This novel intron is the third one in the rat gene and it is in complete accordance with the exon–intron description from the GenBank Feature Table of the rat genome (Figure 1A). However, the original GenBank genomic sequence (Figure 1B) demonstrates that there is no strong reason to believe that it is a real intron. This intron predominantly consists of unknown bases '*n*'. In addition, the intron termini (5′-CC and NN-3′) do not correspond to the standard intron termini (5′-GT and AG-3′). A more close inspection shows that a sequence gap (string of *n*'s letters) within the intron is flanked by near perfect direct repeats that are shown in Figure 1B and 1C. Figure 1 suggests that, most likely, there is no gap in the rat genomic sequence at all. If a base A (marked by an asterisk near the donor site of splicing, Figure 1C) would change to base G, this would restore the genomic contig by precise removal of this intron. We found no supporting evidence for this intron to be present in genomic sequences among all mammalian non-redundant, EST, GSS and WGS GenBank databases. Our example shows that some genomic sequences could be imperfect. Also, all available gene

Table 4: Large-scale characterization of possible cases of intron gain using 80 000 introns form MOID

Case	Species	Gene name; Protein id	Intron position; length	No. of unknown bases '*n*'	5′- and 3′-ends
1	Mouse	Hpn; NP.032307.1	9th; 105 bp	100	5′-nn ... tc-3′
2	Mouse	Pkpl; NP.062619.1	10th; 134 bp	100	5′-at ... ac-3′
3	Mouse	Atp8bl; NP.001001488.1	7th; 2573 bp	100	5′-aa ... nn-3′
4	Rat	LOC287607; XP.213426.2	8th; 1 bp	0	5′-g-3′
5	Mouse	Eif2sl; NP.080390.1	3rd; 129 bp	100	5′-an ... ac-3′
6	Mouse	1110008L16Rik; XP.126928.3	2nd; 1 bp	0	5′-t-3′
7	Mouse	Xpnpepl; NP.573479.2	11th; 137 bp	100	5′-ct ... ca-3′
8	Mouse	Rbp3; NP.056560.1	1st; 110 bp	100	5′-ac ... nn-3′

prediction toolkits occasionally produce false results. There are other problems in the characterization of intron acquisitions [46] and thus, each putative case of intron gain should be investigated in detail.

Table 4 illustrates the results of our recent extended search for newly gained introns in human, mouse and rat using the same algorithm as described by Roy and co-authors [46]. In this research more than 80 000 orthologous intron triplets of human, mouse and rat from the first release of MOID [16] were studied. Altogether, 11 025 orthologous genes of these three organisms were examined. We detected several cases of putative intron gain whose positions correspond to undisputable alignments of protein sequences. However, individual examination of all these 'newly acquired' intronic sequences revealed serious problems in their structures as described in Table 4, for example, one-nucleotide long introns (cases 4 and 6) or predominantly uncharacterized intronic sequences comprised of unknown nucleotides in addition to non-standard intron termini. These results strongly suggest that no single case of intron gain occurred in the lineages of rodent and human. Yet, it is still premature to extrapolate this conclusion on all mammals. We cannot rule out that some specific lineages of these species could have recent alterations in split gene organization. It was shown, that intron gain occurred in fish [48], insects [49] and fungi [50].

Key Point

- Innovations in Exon–Intron Database (EID) are described which significantly extend possibilities in the study of exon–intron gene structures and splicing. Among them is generation of species-specific databases for entire sequenced genomes and introduction of a novel Mammalian Orthologous Intron Database (MOID). Data-quality problems in genomic databases are reviewed.

Acknowledgements

Support for this work was provided by the Medical University of Ohio Foundation and the Stranahan Foundation, through the Program in Bioinformatics and Proteomics/Genomics. We would like to thank Robert Blumenthal and Peter Bazeley, Medical University of Ohio, for discussion and suggestions on our manuscript.

References

1. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res* 2005; **33**:D54–8.
2. Birney E, Andrews D, Caccamo M, *et al*. Ensembl 2006. *Nucleic Acids Res* 2006;**34**:D556–61.
3. Hinrichs AS, Karolchik D, Baertsch R, *et al*. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* 2006;**34**:D590–8.
4. Coward E, Haas SA, Vingron M. SpliceNest: visualization of gene structure and alternative splicing based on EST clusters. *Trends Genet* 2002;**18**:53–5.
5. Gopalan V, Tan TW, Lee BTK, Ranganathan S. Xpro: database of eukaryotic protein-encoding genes. *Nucleic Acids Res* 2004;**32**:D59–63.
6. Croft L, Schandorff S, Clark F, *et al*. ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nature Genetics* 2000;**24**: 340–1.
7. Sakharkar M, Passetti F, de Souza JE, *et al*. ExInt: an Exon Intron Database. *Nucleic Acids Res* 2002;**30**:191–4.
8. Galperin MY. The Molecular Biology Database Collection: 2006 update. *Nucleic Acids Res* 2006;**34**:D3–5.
9. Stamm S, Riethoven JJ, Le Texier V, *et al*. ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res* 2006;**34**:D46–55.
10. Huang HD, Horng JT, Lin FM, *et al*. SpliceInfo: an information repository for mRNA alternative splicing in human genome. *Nucleic Acids Res* 2005;**33**:D80–5.
11. Kim P, Kim N, Lee Y, *et al*. ECgene: genome annotation for alternative splicing. *Nucleic Acids Res* 2005;**33**:D75–9.
12. Pospisil H, Herrmann A, Bortfeldt RH, Reich JG. EASED: extended alternatively spliced EST database. *Nucleic Acids Res* 2004;**32**:D70–4.
13. Saxonov S, Daizadeh I, Fedorov A, Gilbert W. EID: the Exon-Intron Database: an exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Res* 2000;**28**: 185–90.

14. Benson DA, Karsch-Mizrachi I, Lipman DJ, *et al*. GenBank. *Nucleic Acids Res* 2006;**34**:D16–20.
15. Roy SW, Fedorov A, Gilbert W. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc Natl Acad Sci USA* 2003;**100**:7158–62.
16. Fedorov A, Stombaugh J, Harr MW, *et al*. Computer identification of snoRNA genes using a Mammalian Orthologous Intron Database. *Nucleic Acids Res* 2005;**33**: 4578–83.
17. Tatusov RL, Fedorova ND, Jackson JD, *et al*. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 2003;**4**:41.
18. Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
19. Fedorov A, Merican AF, Gilbert W. Large-scale comparison of intron positions between plant, animal and fungal genes. *Proc Natl Acad Sci USA* 2002;**99**:16128–33.
20. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science* 1997;**278**:631–37.
21. International Human Genome Sequence Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 2004;**431**:931–45.
22. Athanasiadis A, Rich A, Maas S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol* 2004;**2**:2144–58 (e391).
23. Blow M, Futreal PA, Wooster R, Stratton MR. A survey of RNA editing in human brain. *Genome Res* 2004;**14**: 2379–87.
24. James CM, Ferguson TK, Leykam JF, Krzycki JA. The amber codon in the gene encoding the monomethylamine methyltransferase isolated from Methanosarcina barkeri is translated as a sense codon. *J Biol Chem* 2001;**276**:34252–8.
25. Beier H, Grimm M. Misreading of termination codons in eukaryotes by natural nonsense suppressor tRNAs. *Nucleic Acids Res* 2001;**29**:4767–82.
26. Howard MT, Anderson CB, Fass U, *et al*. Readthrough of dystrophin stop codon mutations induced by aminoglycosides. *Ann Neurol* 2004;**55**:422–6.
27. Chittum HS, Lane WS, Carlson BA, *et al*. Rabbit beta-globin is extended beyond its UGA stop codon by multiple suppressions and translational reading gaps. *Biochemistry* 1998;**37**:10866–70.
28. Shao X, Shepelev V, Fedorov A. Bioinformatic analysis of exon repetition, exon scrambling and trans-splicing in humans. *Bioinformatics* 2005;Nov 24; in press.
29. Zhang XHF, Heller KA, Hefter I, *et al*. Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. *Genome Res* 2003;**13**:2637–50.
30. Zhang L, Luo L. Splice site prediction with quadratic discriminant analysis using diversity measure. *Nucleic Acids Res* 2003;**31**:6214–20.
31. Leslin CM, Abysov A, Ilyin VA. Structural exon database, SEDB, mapping exon boundaries on multiple protein structures. *Bioinformatics* 2004;**20**:1801–3.
32. Eskesen ST, Eskesen FN, Kinghorn B, Ruvinsky A. Periodicity of DNA in exons. *BMC Molecular biology* 2004; **5**:12.
33. Eskesen ST, Eskesen FN, Ruvinsky A. Natural selection affects frequencies of AG and GT dinucleotides at the 5′ and 3′ ends of exons. *Genetics* 2004;**167**:543–50.
34. Kogan S, Trifonov EN. Gene splice sites correlate with nucleosome positions. *Gene* 2005;**352**:57–62.
35. Chamary JV, Hurst LD. Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else. *Trends Genet* 2005;**21**: 256–59.
36. Tordai H, Patthy L. Insertion of spliceosomal introns in proto-splice sites: the case of secretory signal peptides. *FEBS Letters* 2004;**575**:109–11.
37. Kupfer DM, Drabenstot SD, Buchanan KL, *et al*. Introns and splicing elements of five diverse fungi. *Eukaryot Cell* 2004;**3**:1088–100.
38. Nagarajan N, Yona G. Automatic prediction of protein domains from sequence information using a hybrid learning system. *Bioinformatics* 2004;**20**:1335–60.
39. Ruvinsky A, Eskesen ST, Eskesen FN, Hurst LD. Can codon usage bias explain intron phase distributions and exon symmetry? *J Molec Evol* 2005;**60**:99–104.
40. Gao F, Zhang CT. Comparison of various algorithms for recognizing short coding sequences of human genes. *Bioinformatics* 2004;**20**:673–81.
41. Zhao J, Zhu YM, Song PM, *et al*. Recognition of gene acceptor site based on multi-objective optimization. *Acta Biochim Biophys Sinica* 2005;**37**:435–9.
42. Fedorov A, Roy S, Fedorova L, Gilbert W. Mystery of intron gain. *Genome Res* 2003;**13**:2236–41.
43. Rogozin IB, Sverdlov AV, Babenko VN, Koonin EV. Analysis of evolution of exon-intron structure of eukaryotic genes. *Brief Bioinform* 2005;**6**:118–34.
44. Roy SW, Gilbert W. Rates of intron loss and gain: implications for early eukaryotic evolution. *Proc Natl Acad Sci USA* 2005;**102**:5773–8.
45. O'Neill RJ, Brennan FE, Delbridge ML, *et al*. De novo insertion of an intron into the mammalian sex determining gene, SRY. *Proc Natl Acad Sci USA* 1998;**95**: 1653–7.
46. Roy SW, Fedorov A, Gilbert W. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc Natl Acad Sci USA* 2003;**100**: 7158–62.
47. Veeramachaneni V, Makalowski W. DED: database of evolutionary distances. *Nucleic Acids Res* 2005;**33**:D442–6.
48. Venkatesh B, Ning Y, Brenner S. Late changes in spliceosomal introns define clades in vertebrate evolution. *Proc Natl Acad Sci USA* 1998;**96**:10267–71.
49. Coghlan A, Wolfe KH. Origins of recently gained introns in Caenorhabditis. *Proc Natl Acad Sci USA* 2004;**101**: 11362–67.
50. Nielsen CB, Friedman B, Birren B, *et al*. Patterns of intron gain and loss in fungi. *PLoS Biology* 2004;**2**:e422.