BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 9 2002
Pages 1266–1267

SEGE: A database on 'intron less/single exonic' genes from eukaryotes

Meena K. Sakharkar[1, 2,*], Pandjassarame Kangueane[2], Dmitri A. Petrov[3], A. S. Kolaskar[4] and S. Subbiah[5]

[1]Department of Applied Physics, Stanford University, CA 94305, USA, [2]Bioinformatics Group, MPE, NCSV, NTU, 639798, Singapore, [3]Department of Biological Sciences, Stanford University, CA 94305, USA, [4]BioInformatics Centre, University of Pune, 411007, Pune, India and [5]BIC, NUS, 119260, Singapore

Received on December 31, 2001; revised and accepted on April 4, 2002

ABSTRACT

Summary: Eukaryotes have both 'intron containing' and 'intron less' genes. Several databases are available for 'intron containing' genes in eukaryotes. In this note, we describe a database for 'intron less' genes from eukaryotes. 'Intron less' eukaryotic genes having prokaryotic architecture will help to understand gene evolution in a much simpler way unlike 'intron containing' genes.

Availability: SEGE is available at http://intron.bic.nus.edu.sg/seg/

Contact: mmeena@ntu.edu.sg

INTRODUCTION

The discovery of introns obscured the fact that eukaryotic genes actually have a prokaryotic architecture (Gilbert, 1978). Most eukaryotic genes are 'multi exonic' with their gene structure being interrupted by introns. Introns account for a major proportion in many eukaryotic genomes. For example, the human genome is proposed to contain 24% introns and only 1.1% exons (Venter *et al.*, 2001). Although most genes in eukaryotes contain introns, there are many reports on 'intron less' genes. Databases are available for 'intron containing' genes from eukaryotes (Sakharkar *et al.*, 2000; Saxonov *et al.*, 2000; Schisler and Palmer, 2000). However, there is no database for 'intron less' genes to study them in a concerted manner. Such a database will be useful to identify commonalities in genes having single exon structure and hence enhance our understanding of gene evolution. The availability of annotated sequence data in GenBank (Benson *et al.*, 2000) makes it possible to study these genes in greater detail. In this note, we describe a database of 'intron less' genes in eukaryotes. This database, which we call SEGE, is a collection of gene sequences that are annotated to be 'intron less' with 'single exon' structure.

*To whom correspondence should be addressed.

SYSTEMS AND METHODS

The eukaryotic subdivision files from GenBank release 125 were used to create a dataset containing entries that are reservedly considered as 'single exonic' genes according to the 'CDS' FEATURE convention. By definition, we consider an entry to be putatively 'single exonic' in gene structure if it contains the following description patterns in the corresponding GenBank lines.

1. Contain the word 'DNA' in the LOCUS line at positions 48–53 as per the new locus line format.
2. Contain the pattern 'CDS' in the FEATURES.
3. The 'CDS' line in the FEATURES should contain a continuous span of bases indicated by the number of the first and the last bases in the range separated by two periods (e.g. 23..78). If symbols '<' or '>' are indicated at the end points of the range, the entry is discarded because the range is beyond specified base number in such cases. When operators such as 'complement (location)' are used in the 'CDS' line, the feature is read as complementary to the location indicated and therefore the complementary strands are read from 5′ to 3′.

IMPLEMENTATION

Dataset

The GenBank sequences thus obtained represent a dataset (SEGE) of genes that are putatively considered as eukaryotic 'intron less'. Subset sequences are generated for each of the eukaryotic GenBank subdivisions. 'Intron less' sequences are also generated using annotations in the ORGANISM line for five model organisms—*Homo sapiens*, *Drosophila melanogaster*, *Saccharomyces cerevisiae*, *Caenorabditis elegans* and *Arabidopsis thaliana*.

Caveat

It should be noted that our approach does not include a fraction of eukaryotic 'intron less' genes that do not follow the 'CDS' feature convention. We also do not consider entries that are annotated as 'NA', RNA, 'mRNA', 'tRNA', 'rRNA', 'uRNA', 'snRNA' or 'snoRNA' in the LOCUS line at positions 48–53. Some of these entries might naturally be 'intron less' genes but no robust methodology is available for identifying these entries from GenBank data.

DISCUSSION AND CONCLUSION

The proportion of 'intron containing' and 'intron less' genes in eukaryotes complement each other in different species. The varying proportion is related to the degree of genome complexity. The subtle interplay between their proportions might aid in efficient genome organization during evolution. SEGE will help to take an alternate approach of using 'intron less' genes as a mode for identifying unique features in these genes and hence understand the role of introns in genome organization and gene evolution. 'Intron less' genes circumvent alternative splicing that is frequent in 'intron containing' genes. Therefore, 'intron less' genes in human can be considered for drug targets with less caution. We propose to update the database on a quarterly basis.

REFERENCES

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.

Gilbert,W. (1978) Why genes in pieces? *Nature*, **271**, 501.

Sakharkar,M.K., 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.

Venter,C.J., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A. Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.