



Structural exon database, SEDB, mapping exon boundaries on multiple protein structures

Chesley M. Leslin, Alex Abyzov and Valentin A. Ilyin*

Department of Biology, Northeastern University, 360 Huntington Avenue, Boston, MA 02115, USA

Received on December 19, 2003; revised on February 10, 2003; accepted on February 16, 2004
Advance Access publication February 26, 2004

ABSTRACT

Summary: Comparative analysis of exon/intron organization of genes and their resulting protein structures is important for understanding evolutionary relationships between species, rules of protein organization and protein functionality. We present Structural Exon Database (SEDB), with a Web interface, an application that allows users to retrieve the exon/intron organization of genes and map the location of the exon boundaries and the intron phase onto a multiple structural alignment. SEDB is linked with Friend, an integrated analytical multiple sequence/structure viewer, which allows simultaneous visualization of exon boundaries on structure and sequence alignments. With SEDB researchers can study the correlations of gene structure with the properties of the encoded three-dimensional protein structures across eukaryotic organisms.

Availability: SEDB is publicly available at <http://glinka.bio.neu.edu/SEDB/SEDB.html>

Contact: ilyin@neu.edu

Supplementary information: On the SEDB Web site.

Finding relationships between exon/intron composition in genes, along with the structural properties of the encoded proteins, can help elucidate the evolution of diverged protein superfamilies. Expedient progress in the sequencing of many genomes and the determination of protein structures, along with the advent of experimental and computational methods to detect genes and their exon/intron boundaries, allows researchers to study the relationship between the gene structure and the tertiary structure of the encoded protein. Since the discovery of the linear organization of eukaryotic genes (Berget *et al.*, 1977; Broker *et al.*, 1978), some resources have been developed to study exon/intron boundaries: text-based database EID (Saxonov *et al.*, 2000), relational database ExInt (Sakharkar *et al.*, 2002) and structural viewer XdomView (Vivek *et al.*, 2003), which provide mapping of exon boundaries on a single protein structure. Presented here, Structural Exon Database (SEDB) aims for multiple structural comparisons of exon boundaries, visualizing structural

properties of exons themselves and integrated evolutionary analysis including sequence similarity, structure similarity and conservation of the gene structure. SEDB is a cross database, integrating sequence data from GenBank (Benson *et al.*, 2002), with the search engine BLAST (Altschul *et al.*, 1997); structural neighbor alignments from combinatorial extension (CE; Shindyalov and Bourne, 2001) along with protein structures from the Protein Data Bank (PDB; Berman *et al.*, 2000) and exon/intron information from an in-house version of EID (Saxonov *et al.*, 2000). SEDB's outputs are visualized on Web pages. In addition to the usual Web forms in text/tables/plots, the spatial localization can be viewed in our integrated structure-sequence viewer Friend (Abyzov *et al.*, 2003, <http://mozart.bio.neu.edu/friend>), an extended standalone version of ModView (Ilyin *et al.*, 2003). SEDB provides the ability for comprehensive comparative analysis of exon boundaries/intron positions in multiple structurally similar proteins, which will provide new, additional insight to the previous (Gilbert, 1987) and many current studies on the role and function of exon/intron boundaries in protein structure organization, evolution of gene structure and their correlation between species (references can be found in the SEDB Web page). The current release of SEDB contains 158 958 proteins with exon/intron data. An example of SEDB's outputs is shown in Figure 1.

SEDB has two main search criteria, the first search is based on structural similarity and the second search is based on sequence similarity. When performing a structural search, SEDB accepts a PDB code, chain identifier, RMSD and Z-score cutoffs to create a table to display structurally similar proteins (A1). For each structural neighbor in this table the related sequences with exon data are collected from a local EID and presented in the table. To speed up the data retrieval, the alignments between PDB structures and EID (GenBank) sequences with annotated gene structure have been pre-calculated using BLAST. The following thresholds have been applied: sequence similarity of at least 90% and alignment length of at least 50% of either sequence length, but not smaller than 30 amino acids. The user can check which proteins to show in the structural alignment, and then obtain the alignment with EID sequences aligned to the PDB sequences. It should

*To whom correspondence should be addressed.

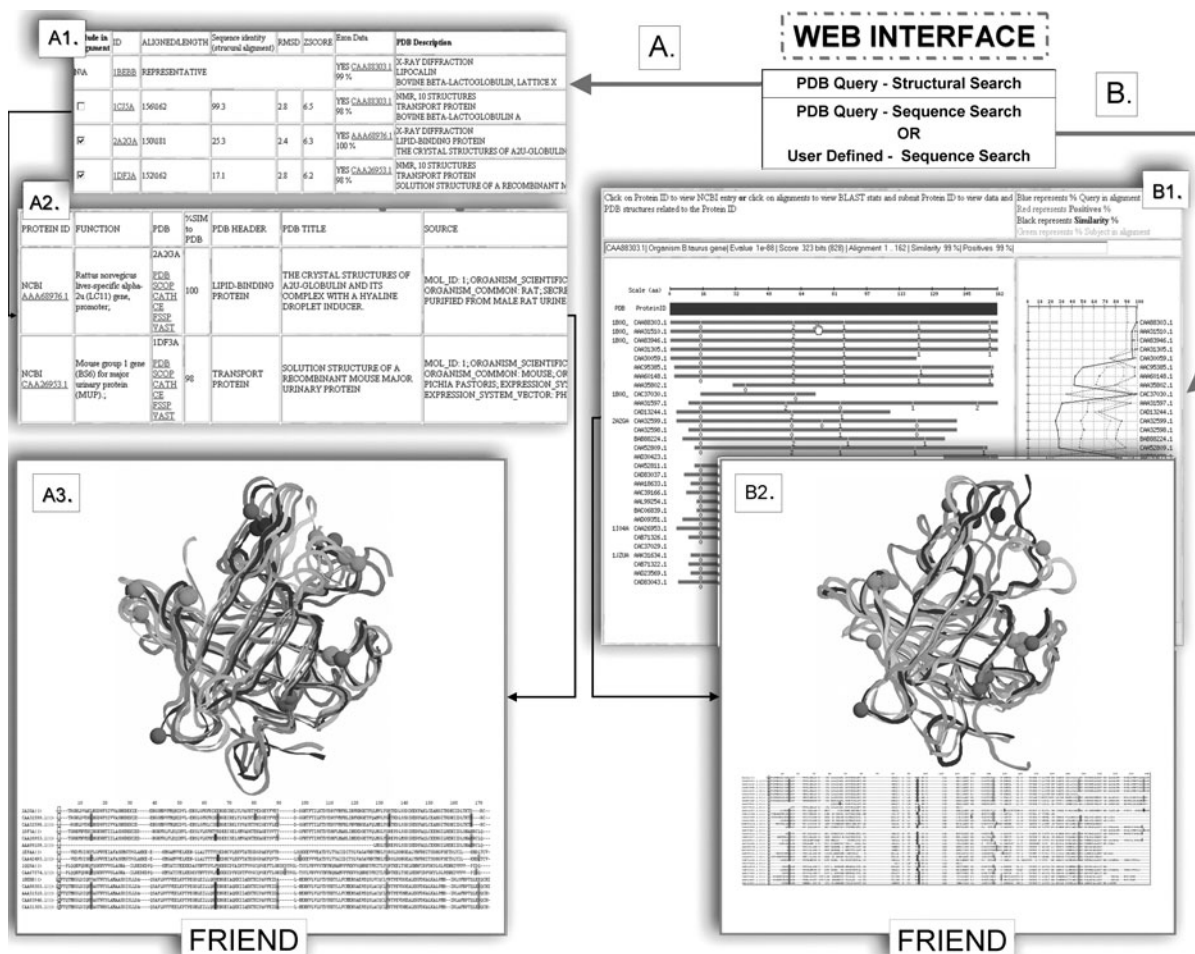


Fig. 1. An example of SEDB Web interface and query results.

be noted that aligning the EID (GenBank) sequences does not change the structural alignment. A file is created in a format that allows Friend to display both the three-dimensional structures superimposed by the structural alignment in the structure window and the aligned sequences in the sequence window (A3). The exon boundaries and intron phases are denoted by color, red = phase 0, green = phase 1, blue = phase 2.

The second main search is a sequence search that uses BLAST to find similar EID sequences. When performing this type of search a user can specify a PDB code, chain identifier, the number of output alignments, un/gapped BLAST, name of the database (experimental or non-experimental) and the number of high-scoring segment pairs (HSPs) to display. Once a search is completed, a Web page is displayed (B1) showing hits in a graphical output, along with a plot displaying the scoring information about the hits. From this page users can visualize data in the Friend viewer similar to the output from the structural search. The sequence window (B2) is displayed in a multiple pairwise alignment format, where all hits are aligned to the query sequence.

Another feature of SEDB (query by Protein ID) provides users with the ability to map exon data for a protein sequence with unknown structure on similar proteins with known structures. This allows spatial visualization of exon/intron boundaries in a protein without an experimental structure.

APPLICATIONS

Structurally related proteins may not share any detectable sequence similarity (Sander and Schneider, 1991). In order to study the correlation of gene structures in a family of proteins both sequence and structure approaches can be applied. For example, lipocalin proteins show low sequence similarity falling below the commonly accepted threshold for a reliable sequence alignment of 30%. The lipocalin protein (PDB 1BEB chain B) was analyzed using the structural similarity search in SEDB (Fig. 1). The search identified three structurally related proteins and their relatives from EID (1A). These proteins share low sequence similarity with lipocalin (which are 20.26, 13.92 and 15.62% for the

corresponding Protein IDs, CAA42493.1, CAA67574.1 and CAA26553.1, respectively), but their exon boundaries are remarkably similar (A3).

Since SEDB is a Web-based application, all data can be linked to other forms of information using HTML embedded JavaScript. Taking into account the current uncertainty in the annotation of gene structure in the existing databases, all Protein IDs are directly linked to NCBI to verify whether the gene structure is from experimental or predicted data. Other links for quick retrieval of information include: PDB (Berman *et al.*, 2000), SCOP (Murzin *et al.*, 1995), CATH (Orengo *et al.*, 1999), CE (Shindyalov and Bourne, 2001), FSSP (Holm and Sander, 1996) and MMDB databases (Gibrat *et al.*, 1996).

SEDB utilizes protein encoding, intron containing genes from GenBank release 137 (158 958 proteins). The dataset was compiled using modified PERL scripts from EID (Saxonov *et al.*, 2000). Structural relationships are obtained from the CE database (Shindyalov and Bourne, 2001). SEDB uses PERL scripts on a relational MySQL database, running on a LINUX server (RedHat 8.0, Apache version 2.0.40). To view structural superimpositions the application Friend has to be installed on the local machine, which is publicly available at <http://mozart.bio.neu.edu/friend>.

Future developments include extending the structural alignments to other structural databases, gathering exon boundaries directly from cDNA data and using SEDB for large-scale analysis of structural exons.

REFERENCES

- Abyzov, A., Leslin, C. and Ilyin, V.A. (2003) An integrated analytical front-end application for bioinformatics, Friend, CSBi, Computational and Systems Biology Conference, January 9–10, Cambridge, MA.
- 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.
- Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. and Wheeler, D.L. (2002) GenBank. *Nucleic Acids Res.*, **30**, 17–20.
- Berget, S.M., Moore, C. and Sharp, P.A. (1977) Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Rev. Med. Virol.*, **10**, 356–362.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
- Broker, T.R., Chow, L.T., Dunn, A.R., Gelin, R.E., Hassell, J.A., Klessig, D.F., Lewis, J.B., Roberts, R.J. and Zain, B.S. (1978) Adenovirus-2 messengers—an example of baroque molecular architecture. *Cold Spring Harb. Symp. Quant. Biol.*, **42**, 531–553.
- Gibrat, J.F., Madej, T. and Bryant, S.H. (1996) Surprising Similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.
- Gilbert, W. (1987). The exon theory of genes. *Cold Spring Harb. Symp. Quant. Biol.*, **52**, 901–905.
- Holm, L. and Sander, C. (1996) Mapping the protein Universe. *Science*, **273**, 595–603.
- Ilyin, V.A., Pieper, U., Stuart, A.C., Marti-Renom, M.A., McMahan, L. and Sali, A. (2003) ModView, visualization of multiple protein sequences and structures. *Bioinformatics*, **19**, 165–166.
- Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
- Orengo, C.A., Pearl, F.M., Bray, J.E., Todd, A.E., Martin, A.C., Lo, C.L. and Thornton, J.M. (1999) The CATH database provides insights into protein structure/function relationships. *Nucleic Acids Res.*, **27**, 275–279.
- Sakharkar, M., Passetti, F., de Souza, J.E., Long, M. and de Souza, S.J. (2002) ExInt: an exon-intron database. *Nucleic Acids Res.*, **30**, 191–194.
- Sander, C. and Schneider, R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
- 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.
- Shindyalov, I.N. and Bourne, P.E. (2001) A database and tools for 3-D protein structure comparison and alignment using the combinatorial extension (CE) algorithm. *Nucleic Acids Res.*, **29**, 228–229.
- Vivek, G., Tan, T.W. and Ranganathan, S. (2003) XdomView: protein domain and exon position visualization. *Bioinformatics*, **19**, 159–160.