Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W271–W273***
doi:10.1093/nar/gki589

SVC: structured visualization of evolutionary sequence conservation

S. Roepcke*, P. Fiziev, P. H. Seeburg[1] and M. Vingron

Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany and
[1]Max Planck Institute for Medical Research, Jahnstrasse 73, 69120 Heidelberg, Germany

Received February 14, 2005; Revised and Accepted April 29, 2005

ABSTRACT

We have developed a web application for the detailed analysis and visualization of evolutionary sequence conservation in complex vertebrate genes. Given a pair of orthologous genes, the protein-coding sequences are aligned. When these sequences are mapped back onto their encoding exons in the genomes, a scaffold of the conserved gene structure naturally emerges. Sequence similarity between exons and introns is analysed and embedded into the gene structure scaffold. The visualization on the SVC server provides detailed information about evolutionarily conserved features of these genes. It further allows concise representation of complex splice patterns in the context of evolutionary conservation. A particular application of our tool arises from the fact that around mRNA editing sites both exonic and intronic sequences are highly conserved. This aids in delineation of these sites. SVC is available at http://svc.molgen.mpg.de.

INTRODUCTION

Evolutionary conservation is a powerful feature to identify functionally important regions of a genome. Among the most highly conserved are the protein-coding parts. Furthermore, RNA transcripts and regulatory regions show a high degree of sequence similarity. Our starting point for the design of the SVC tool was a collaboration in which we have searched for novel candidate sites subject to RNA editing (1). As shown recently, exceptionally conserved exonic sequence guide us to editing sites in the transcript (2). For visual exploration and detailed analyses of alignments of evolutionarily related sequences versatile software tools are indispensable (3,4). The VISTA tools, for example, provide the user with the ability to visualize and explore the degree of conservation and the alignments along genomic regions. One main difficulty in detailed genome analyses of single loci is the variability of the scales. In particular, many mammalian genes consist of many exons of a few hundred bases or even fewer that are interrupted by large introns of up to several hundred thousand bases. So, if drawn to scale, the usual visualization of such a gene in a genome browser results in a line representing the genomic sequence with the exons appearing as thin vertical marks (5). The detailed investigation of sequence conservation in the genomic context, at splice sites for example, usually requires a lot of zooming in and out or moving along the genomic sequence.

We introduce another approach that is tailored to investigating the conservation of the gene structure and regulatory sequences. We exploit the fact that distant homology can more easily be identified between protein-coding sequences. Mapping the aligned sequences back onto their genomes provides a scaffold of the conservation pattern at this orthologous region. Pairwise nucleotide sequence alignments of the exons and introns are performed and fitted into the scaffold. Our aim is to display evolutionary conservation in a more structured way. Especially for researchers who are interested mainly in alternative gene products and their regulation, our tool helps to focus on relevant regions and to explore the genome more efficiently.

MATERIALS AND METHODS

As primary data source we use the following Ensembl databases (6): homo_sapiens_core_30_35c, mus_musculus_core_30_33f, rattus_norwegicus_core_29_3f, fugu_rubripes_core_30_2e, danio_rerio_core_30_4c. The lists of orthologous genes for each pair of organisms are obtained from Ensembl via EnsMart (7) and are based on the dataset ensembl_compara_27_1. For the identification of conserved sequences between human and puffer fish exonic sequence we used the following sequence file: Fugu_rubripes.FUGU2.jul.dna_rm.scaffold.fa.

Our application SVC performs extensive pairwise sequence alignments. For this we employ the BLAST family of alignment programs obtained from the NCBI website (8,9).

*To whom correspondence should be addressed. Tel: +49 30 84131159; Fax: +49 30 84131152; Email: roepcke@molgen.mpg.de

As input the user enters a gene name or identifier, a text string for an advanced keyword search, or uploads a list of gene identifiers from file. The advanced search scans the available gene descriptions in the Ensembl database and returns all genes that match the query. The first step in the analysis pipeline consists of the collection of all the necessary gene, transcript and exon identifiers. Then the orthologs in each organism of interest are identified and the sequences and the annotation are downloaded from Ensembl.

For each pair of orthologous genes we compute pairwise local alignments between the exons to obtain the best reciprocal hits. In this procedure we apply TBLASTX to gain sensitivity for the coding exons. This generates the scaffold of the conservation of the gene structure. The exonic as well as the intronic sequences are aligned at the nucleotide level using BLASTN in order to determine evolutionarily highly conserved regions. Then we perform a sequence search of the exons against the introns of the orthologous genes in the other species using TBLASTX again. The purpose of this step is the identification of alternative and conserved exons that were observed in one organism but not in the other. After some trials we decided to choose an *E*-value cutoff of 10^{-4} in all the BLAST analyses.

We have also adapted our analysis pipeline to search for candidate RNA editing sites. As an additional step, we scan the protein-coding exons for highly similar sequence windows. The length of the window and the required percentage identity can be specified by the user. For a human–mouse comparison we recommend 80 bp window length and a cutoff of 95% identity. As supporting evidence we align the flanking introns to identify highly similar sequence regions, and we compare the exonic windows to the puffer fish sequence to illustrate the degree of conservation.

To get an overview of complex gene structures we developed a new schematic visualization. Exons are drawn as boxes and introns as horizontal lines between them. Arcs indicate the splicing events of the adjoint exons. Arcs belonging to alternative transcripts are coded in different colours. It is important to emphasize that exons and introns are not drawn to scale, but all at fixed lengths. Sequence similarities between introns and between introns and exons are displayed by blue and green boxes between the genes, respectively. All these data can be downloaded from the SVC server for further analysis. A detailed description of the options, the visualization and the output has been included in the online help page.

RESULTS AND DISCUSSION

The main focus of our work is the investigation of alternative gene products and their differential regulation. To support this, we have developed the web tool SVC. Given two orthologous genes, SVC facilitates a detailed analysis of the evolutionary conservation of their structural elements. Structural elements in our nomenclature are exons and regulatory sequences. In this work we focus on a detailed comparison of complex orthologous genes. The detection of the orthologs is not part of this work, and for this we rely on the Compara database (10). We chose not to align the whole genomic sequence but to concentrate on the protein-coding part of orthologous genes first. This allows us to identify more distant sequence homology, even between small exons. If the structure of the two genes is found to be conserved we obtain a scaffold to map the orthologous exons and introns onto the underlying genomes. SVC presents the available information on the gene products in a clearly arranged way. Sequence similarity between exons and introns is emphasized by colour bars.

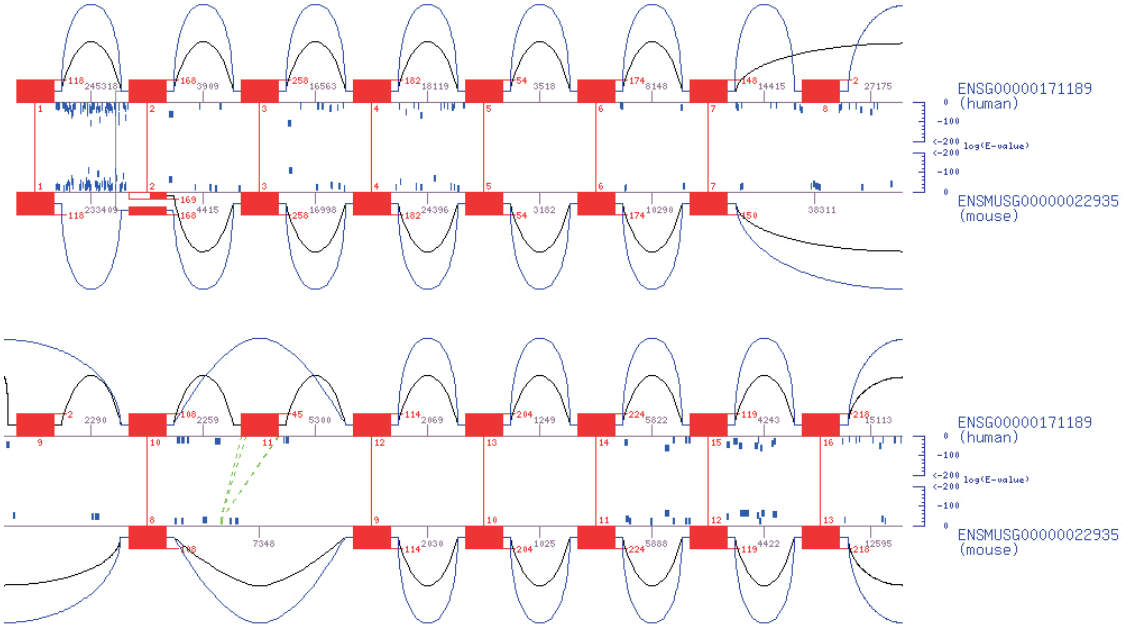

Figure 1. SVC graphical output for comparison of human and mouse glutamate receptor 5 (fragment). The known editing site is situated in human exon 15. The blue bars in the flanking introns illustrate high evolutionary conservation. Green dashed lines indicate similarity between exonic and intronic sequences. In mouse exon 2, a second transcription start site is annotated. The open red box indicates the 5′-UTR. The lengths of exons and introns are given in base pairs.

We know from single-gene studies that many genomic loci are not yet entirely resolved (11,12). Rare splice variants or alternative transcripts are often not represented in the available databases. If a splice variant is not displayed as such in SVC it still might be that there is already evidence for this transcript from large-scale efforts to investigate alternative splicing (13–15). Ideally, our web tool could access the detailed information of these projects. We believe that, if we do not display the exons and introns at their relative extensions, the visualization becomes more focused and conserved features stand out more clearly.

Sequence similarity in orthologous introns of diverged species suggests these regions have functional importance. This could stem from exons of alternative, but not yet observed, transcripts or regulatory sequences. Consider, for instance, an intronic region that shows sequence similarity to an exon in another organism. If one wished to test this in the lab, one could download the sequence and input it into a primer design program. Another interesting observation is that alternatively spliced exons often come with conserved regulatory regions in the flanking introns (16). If this is the case for an exon of interest, it should be apparent in the SVC visualization.

RNA editing is a phenomenon whereby single nucleotides in the mRNA are transformed—adenosine to inosine, for example (1). High evolutionary sequence conservation can indicate potential RNA editing sites (2). Applying our analysis pipeline, we have identified candidate editing sites in the human genome. Exonic sites are considered candidates if the surrounding 80 bp window is extraordinarily similar between the human and the mouse lineage and if we find conservation in the flanking introns. Additionally, we consider the alignments to the puffer fish genome. A number of identified candidates are currently being tested in the lab.

As a case study we present the results for human glutamate receptor 5 (GluR5) (17,18). The encoding gene is edited at a specific site in exon 14 of the Ensembl transcript ENST00000309434, which corresponds to exon number 15 in Figure 1. From the connecting red line and the blue bars in the vicinity it is obvious that the corresponding exons are conserved between human and mouse as well as many regions in the flanking introns. Another interesting observation is that human GluR5 exon 11 has not been observed in any mouse transcript in the Ensembl database. However, the sequence similarity to an intronic region at the right genomic location strongly suggests that this variant also exists in mice. The effect of drawing the introns not to scale but to standard length is also seen in Figure 1. The first intron is more then 200 kb long, which is more than 10 times the length of intron 4, for example. The conservation pattern of the first intron looks much denser, but this in fact might stem from its shear length.

In conclusion, we believe that SVC is a helpful tool to investigate sequence conservation in the context of structure in complex vertebrate genes. We have introduced a new visualization that displays the structural elements of a gene with standard size in order to emphasize evolutionarily conserved features.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by the Max Planck Institute for Molecular Genetics.

Conflict of interest statement. None declared.

REFERENCES

1. Seeburg,P.H. and Hartner,J. (2003) Regulation of ion channel/neurotransmitter receptor function by RNA editing. *Curr. Opin. Neurobiol.*, **13**, 279–283.
2. Hoopengardner,B., Bhalla,T., Staber,C. and Reenan,R. (2003) Nervous system targets of RNA editing identified by comparative genomics. *Science*, **301**, 832–836.
3. Frazer,K.A., Pachter,L., Poliakov,A., Rubin,E.M. and Dubchak,I. (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res.*, **32**, W273–W279.
4. Schwartz,S., Zhang,Z., Frazer,K.A., Smit,A., Riemer,C., Bouck,J., Gibbs,R., Hardison,R. and Miller,W. (2000) PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res.*, **10**, 577–586.
5. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.
6. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al.* (2005) Ensembl 2005. *Nucleic Acids Res.*, **33**, D447–D453.
7. Kasprzyk,A., Keefe,D., Smedley,D., London,D., Spooner,W., Melsopp,C., Hammond,M., Rocca-Serra,P., Cox,T. and Birney,E. (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.*, **14**, 160–169.
8. 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.
9. Jenuth,J.P. (2000) The NCBI. publicly available tools and resources on the Web. *Methods Mol. Biol.*, **132**, 301–312.
10. Birney,E., Andrews,T.D., Bevan,P., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cuff,J., Curwen,V., Cutts,T. *et al.* (2004) An overview of Ensembl. *Genome Res.*, **14**, 925–928.
11. Alcazar,O., Ho,R.C., Fujii,N. and Goodyear,L.J. (2004) cDNA cloning and functional characterization of a novel splice variant of c-Cbl-associated protein from mouse skeletal muscle. *Biochem. Biophys. Res. Commun.*, **317**, 285–293.
12. Winter,J., Lehmann,T., Krauss,S., Trockenbacher,A., Kijas,Z., Foerster,J., Suckow,V., Yaspo,M.L., Kulozik,A., Kalscheuer,V. *et al.* (2004) Regulation of the MID1 protein function is fine-tuned by a complex pattern of alternative splicing. *Hum. Genet.*, **114**, 541–552.
13. Huang,Y.H., Chen,Y.T., Lai,J.J., Yang,S.T. and Yang,U.C. (2002) PALS db: putative alternative splicing database. *Nucleic Acids Res.*, **30**, 186–190.
14. Thanaraj,T.A., Stamm,S., Clark,F., Riethoven,J.J., Le Texier,V. and Muilu,J. (2004) ASD: the Alternative Splicing Database. *Nucleic Acids Res.*, **32**, D64–D69.
15. Modrek,B. and Lee,C.J. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature Genet.*, **34**, 177–180.
16. Sorek,R. and Ast,G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome. Res.*, **13**, 1631–1637.
17. Paschen,W. and Djuricic,B. (1994) Extent of RNA editing of glutamate receptor subunit GluR5 in different brain regions of the rat. *Cell. Mol. Neurobiol.*, **14**, 259–270.
18. Herb,A., Higuchi,M., Sprengel,R. and Seeburg,P.H. (1996) Q/R site editing in kainate receptor GluR5 and GluR6 pre-mRNAs requires distant intronic sequences. *Proc. Natl Acad. Sci. USA*, **93**, 1875–1880.