D32–D35 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 29 November 2006*
doi:10.1093/nar/gkl977

Developments in CORG: a gene-centric comparative genomics resource

C. Dieterich*, M. W. Franz and M. Vingron[1]

Department of Evolutionary Biology, Max Planck Institute for Developmental Biology, Spemannstrasse 35-37, 72076 Tübingen, Germany and [1]Department of Computational Biology, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany

Received September 14, 2006; Revised and Accepted October 26, 2006

ABSTRACT

The CORG resource (Comparative Regulatory Genomics, http://corg.eb.tuebingen.mpg.de) provides extensive cross-species comparisons of promoter regions in particular and whole gene loci in general. Pairwise as well as multiple alignments of 10 vertebrate species form the key component of CORG. We implemented a rapid alignment approach based on weight matrix motif anchors to ensure efficient computation and biologically informative alignments. All CORG workbench components have been enhanced towards more flexibility and interactivity. Reference sequence based data presentation and analysis was put into the well-known and modular Generic Genome Browser framework. Herein, various plugins facilitate online data analysis and integration with static conservation data. Main emphasis was put on the design of a new JAVA WebStart application for comparative data display. Flexible data import and export options for standard formats complete the provided services.

INTRODUCTION

Web-based comparative genomics resources support researchers in finding significant patterns in biological sequences and likely functions thereof. The CORG workbench and database are dedicated to the analysis of non-coding DNA of homologous gene loci. It has served this purpose ever since its start in 2003 (1). Unlike other resources [e.g. ref. (2,3)] we do not compare whole genomes but DNA sequence of whole homologous gene loci and flanking regions. Local sequence conservation is an indicator of functional elements that either act in *cis* (e.g. promoter elements) or *trans* (e.g. RNA genes). Small transcriptional units such as micro-RNAs or enhancers have been reported to reside within larger protein-coding genes [e.g. ref. (4)].

To this end, CORG has been extended in content, form and function since the last report (1) and this paper details the improvements. Two sequence regions are in the center of CORG's new capabilities: (i) upstream regions that are likely to encompass one or more promoters of a gene and (ii) whole gene loci covering the complete gene structure plus 5 kb of flanking sequence. In the first group we expect to find functional promoter elements by conservation combined with motif searches and experimentally defined transcription start sites. The non-coding portion of the second group is largely unexplored territory where cross-species conservation patterns for homologous gene groups may provide crucial hints on function.

Furthermore, we have improved on the strategy of alignment computation and visualization. CORG became also more flexible with respect to data manipulation as it provides interactive analysis tools and data exchange facilities.

RESULTS

The current release of CORG contains local pairwise and multiple alignments for 10 vertebrate species: Man, rhesus monkey, rat, mouse, dog, cow, chicken, frog and two fish. This species selection covers an enormous evolutionary distance and can be used to address questions as to the turnover of functional elements.

Alignment of homologous gene loci

Local pairwise alignments are still the key component of the CORG database. Multiple cross-species comparisons demand new ways for computing meaningful local sequence similarities. Upstream regions are aligned by SITEBLAST (5); a modified version of BLASTZ that employs weight matrix scans to find alignment anchors. The rational is that biologically meaningful anchors can be extended into alignments that capture further conserved functional elements. The JASPAR library (6) of family profiles provides seed motifs. Program parameters are set to pick up all motif instances that have power of 0.2 or less. The scoring scheme follows the HKY model that takes sequence specific GC content

*To whom correspondence should be addressed. Tel: +49 7071 601 405; Fax: +49 7071 601 498; Email: christoph.dieterich@tuebingen.mpg.de

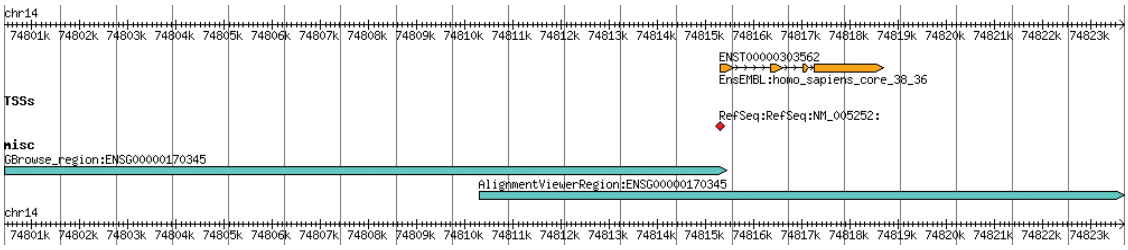

Figure 1. Genome location and structure of the human *c-fos* gene. The upper panel of the CORG workbench shows the selected gene's structure and orientation on the genome assembly. RefSeq transcription start sites are represented by red diamond symbols. The extent of associated CORG regions is given by turquoise filled arrows.

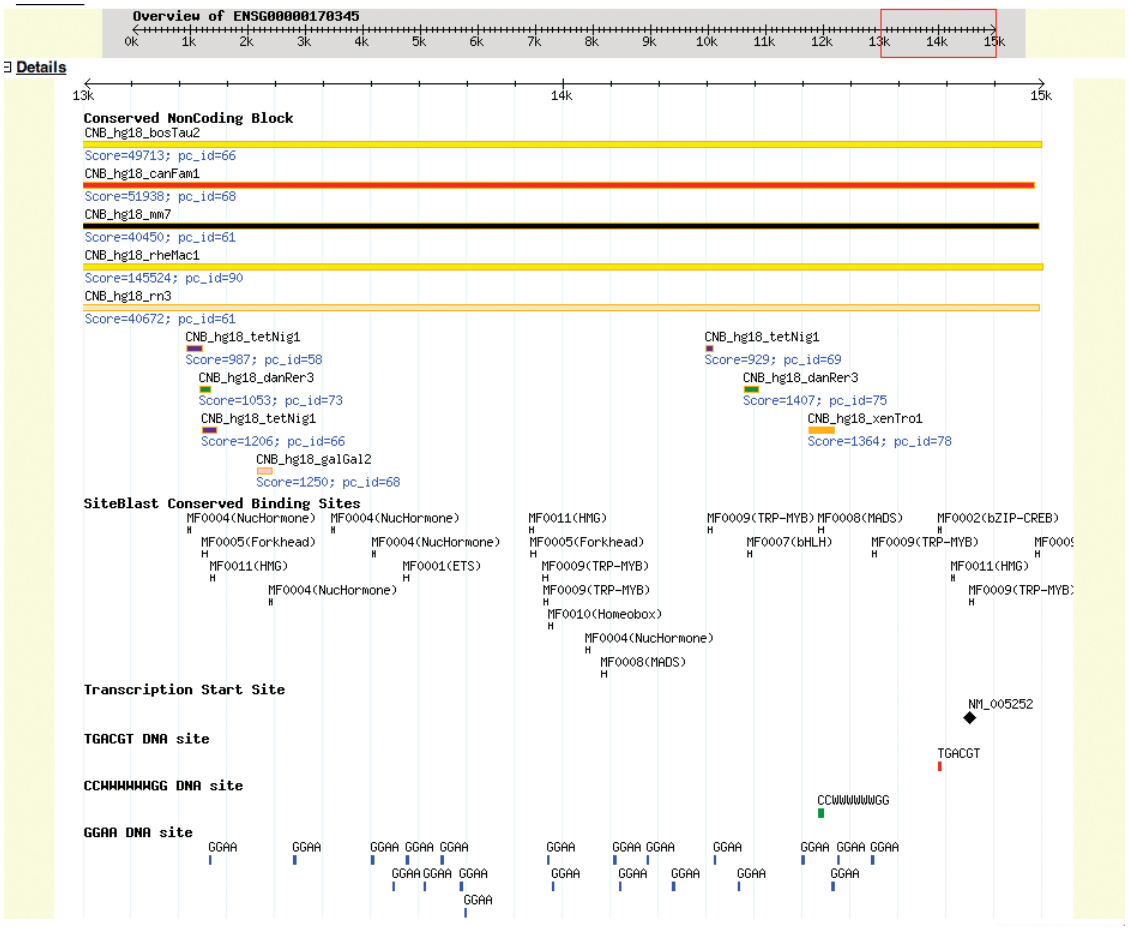

Figure 2. Genome Browser view on 2 kb segment of *c-fos* upstream region. Conserved blocks are displayed on the top track, followed by conserved binding sites as given by pairwise alignment anchor points. All items are clickable and provide additional details on demand. A diamond shaped symbol represents the annotated transcription start site. The three bottom tracks display analysis results from the 'Annotate DNA sites' plugin for queries of consensus sequences of CREB boxes, Serum response elements and ETS binding sites.

into account and is set to an expected distance of 47 PAM (7). No positional constraints are enforced on the alignments, but alignment scores must exceed 10 times the average match score. The second set of alignments is computed for whole gene loci applying the same scoring scheme, but ensuring collinear block structure and a traditional alignment seed approach (12 matching positions out of 19).

Multiple alignments are subsequently computed from pairwise ones. We enumerated all distinct and maximally long paths through the graph of overlapping alignments. Corresponding sequences are optimally realigned with the POA software (8).

CORG workbench

The website is divided into four sections: Search page, Batch retrieval page, DAS tutorial and Help page. An interactive user session usually starts with searching for a particular gene identifier. Matching identifiers are displayed along with a concise description to guarantee the right choice.

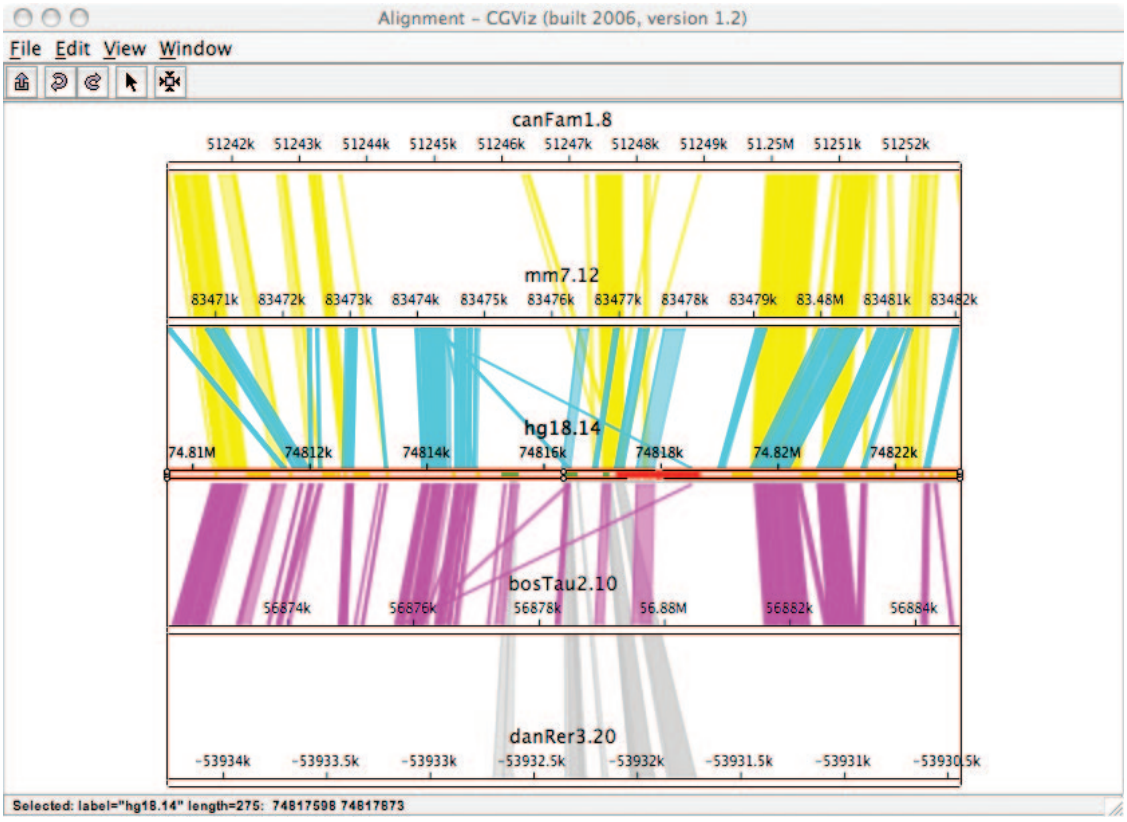

Figure 3. New Comparative alignment viewer implemented as JAVA WebStart application. The screenshot shows the gene locus of the *c-fos* gene on chromosome 14 and corresponding pairwise alignments. Gene structure annotation (exons in green, repetitive elements in yellow) is imported from GFF-formatted files. Pairwise or multiple alignments are imported from MAF-formatted files.

The following page features three views on the database content. The top panel (Figure 1) shows the genomic position and structure of the selected gene (transcript variants along with transcription start sites). Two additional tracks represent the range of the two CORG regions, upstream and whole gene. The middle panel constitutes a local view of the selected gene and annotation. Herein, we use the Generic Genome Browser framework (9) to display, edit, ex- and import annotations of CORG regions. The modular plugin architecture of this browser allowed us to add more function such as filter, finders and annotators. Our filter plugin can reduce the number of displayed conserved blocks with respect to their score, length or percentage identity. Two annotator plugins serve to pinpoint motif matches. Either one or more IUPAC consensus sequences serve as input or alternatively a weight matrix search can be performed on the displayed segment. Figure 2 shows a 2000 bp section of the upstream region of the human *c-fos* gene. *c-fos* is a component of the AP-1 transcription factor complex and therefore requires tight regulation (10). It is known that the promoter can be activated by either protein kinase A, Ras or JAK-STAT signaling (11). Endpoints of these signaling cascades sit at locations −60, −300 and −340 relative to the transcription start site (11). Figure 2 demonstrates the use of plugins along with CORG annotation to highlight the aforementioned promoter elements. Both, static database content and newly generated annotation may be exported by the 'Download Sequence File' plugin. Furthermore, we continue to develop plugins for emerging data sources in gene regulation such as binding assays, reporter gene assays or high-resolution expression profiling experiments.

MyCORG—the CGViz AlignmentViewer

The bottom part of this webpage constitutes a paradigm shift away from the web browser towards a complete software application. We have abandoned the previous JAVA applet for the comparative display of alignments and implemented a JAVA WebStart application (requires Java 1.5, http://java.sun.com/products/javawebstart/) based on the CGViz framework (http://www-ab.informatik.uni-tuebingen.de/software/cgviz/welcome.html). This application is installed locally and requires standard input formats: MAF for alignments and GFF for sequence feature annotation (http://genome.ucsc.edu/FAQ/FAQformat). Both data record formats are provided by the CORG database, but data may be also obtained from other resources as well. This procedure guarantees maximal flexibility to the user. Figure 3 is an example of the Alignment viewer showing the whole gene locus of *c-fos* with CORG alignments and annotation. Indeed, our application offers seamless navigation through cross-species alignments.

Large-scale data retrieval

Complete CORG data sets can be easily obtained from the Batch retrieval section. The user can select from a variety

of GFF annotation and MAF alignment dumps. We do have a DAS service in operation. Please visit the DAS tutorial section for details (http://corg.eb.tuebingen.mpg.de/cgi-bin/DAS_tutorial.pl).

Future directions

The CORG resource will primarily remain web-based as this guarantees convenient access to CORG while having minimal requirements on the client computer. Client-side software components will continue to supplement the web resource. To keep CORG up-to-date with developments in sequencing and functional annotation, we strive to release a new CORG version every six months. Large-scale experimental data sets will be integrated into the CORG framework.

CONCLUSIONS

The CORG database has been updated to offer a great diversity of 10 vertebrate species. Pairwise as well as multiple guide the user in her or his data interpretation. This move became possible with improvements in alignment computation (SITEBLAST) using biological motifs as anchors. Whole gene loci alignments with locus-specific scoring schemes cover introns, UTR and downstream regions in the search for functional elements. Greatest flexibility is guaranteed by keeping the CORG workbench open and adjustable through the Generic Genome Browser framework. Finally, JAVA WebStart technology enables the seamless integration and visualization of CORG data on each local desktop.

ACKNOWLEDGEMENTS

We would like to thank Daniel Huson for sharing the source code of the CGViz framework. We also appreciate the help of Daniel Richter who provided us with valuable hints as to writing CGViz applications. Funding through the EU BioSapiens Network of Excellence is gratefully acknowledged. Funding to pay the Open Access publication charges for this article was provided by the Department of Computational Biology, Max Planck Institute for Molecular Genetics.

Conflict of interest statement. None declared.

REFERENCES

1. Dieterich,C., Wang,H., Rateitschak,K., Luz,H. and Vingron,M. (2003) CORG: a database for COmparative Regulatory Genomics. *Nucleic Acids Res*., **31**, 55–57.
2. Birney,E., Andrews,D., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cox,T., Cunningham,F., Curwen,V., Cutts,T. *et al*. (2006) Ensembl 2006. *Nucleic Acids Res*., **34**, 556–561.
3. Hinrichs,A.S., Karolchik,D., Baertsch,R., Barber,G.P., Bejerano,G., Clawson,H., Diekhans,M., Furey,T.S., Harte,R.A., Hsu,F. *et al*. (2006) The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res*., **34**, 590–598.
4. Lazorchak,A.S., Schlissel,M.S. and Zhuang,Y. (2006) E2A and IRF-4/Pip promote chromatin modification and transcription of the immunoglobulin kappa locus in pre-B cells. *Mol. Cell Biol*., **26**, 810–821.
5. Michael,M., Dieterich,C. and Vingron,M. (2005) SITEBLAST—rapid and sensitive local alignment of genomic sequences employing motif anchors. *Bioinformatics*, **21**, 2093–2094.
6. Vlieghe,D., Sandelin,A., Bleser,P.J.D., Vleminckx,K., Wasserman,W.W., van Roy,F. and Lenhard,B. (2006) A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res*., **34**, 95–97.
7. States,D., Gish,W. and Altschul,S. (1991) Improved sensitivity of nucleic acids database searches using application-specific scoring matrices. *Meth. Enzymol*., **3**, 66–70.
8. Lee,C., Grasso,C. and Sharlow,M.F. (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics*, **18**, 452–464.
9. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome. Res*., **12**, 1599–1610.
10. Rivera,V.M. and Greenberg,M.E. (1990) Growth factor-induced gene expression: the ups and downs of c-fos regulation. *New Biol*., **2**, 751–758.
11. Janknecht,R., Cahill,M.A. and Nordheim,A. (1995) Signal integration at the c-fos promoter. *Carcinogenesis*, **16**, 443–450.