Genome analysis

ECRbase: database of evolutionary conserved regions, promoters, and transcription factor binding sites in vertebrate genomes

Gabriela Loots and Ivan Ovcharenko1,*

CMLS and 1Computation Directorates, Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA, 94550, USA

Received on September 29, 2006; revised and accepted on October 19, 2006
Advance Access publication November 7, 2006
Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Evolutionary conservation of DNA sequences provides a tool for the identification of functional elements in genomes. We have created a database of evolutionary conserved regions (ECRs) in vertebrate genomes, entitled ECRbase, which is constructed from a collection of whole-genome alignments produced by the ECR Browser. ECRbase features a database of syntenic blocks that recapitulate the evolution of rearrangements in vertebrates and a comprehensive collection of promoters in all vertebrate genomes generated using multiple sources of gene annotation. The database also contains a collection of annotated transcription factor binding sites generated using multiple sources of gene annotation. ECRbase currently includes human, rhesus macaque, dog, opossum, rat, mouse, chicken, frog, zebrafish and fugu genomes. It is freely accessible at http://ecrbase.dcode.org.

Contact: ovcharenko1@llnl.gov

Supplementary information: Supplementary Data are available at Bioinformatics online.

1 INTRODUCTION

Cross-species sequence comparison is a powerful method for identifying functional regions in a genome (Loots et al., 2000). In recent years, evolutionary conservation has guided the discovery of novel genes (Pennacchio et al., 2001) and regulatory elements (Woolfe et al., 2005). While sequences coding for proteins are strongly conserved across species, they encompass a small portion of a vertebrate genome. Some fraction of noncoding sequences is also conserved in the phylogeny of vertebrates, and increasing lines of evidence highlight the functional role of these evolutionary conserved regions (ECRs) in different aspects of vertebrate biology.

If ECRs are functionally important in vertebrate genomes, these regions should become a critical hunting ground for transcriptional regulatory signals that determine when, where and in what quantities genes are expressed. In addition, genetic variation in these elements may be responsible for individual variability of gene expression, which can therefore define susceptibility to disease (Stranger et al., 2005).

Contemporary genomics research is moving towards high-throughput and systematic whole-genome analysis that requires investigators to access comprehensive genomic data. Generating large datasets of genome alignments, ECR and transcription factor binding site (TFBS) data on a genome scale requires extensive computational resources that are not always readily available.

To facilitate genome-wide experimentation for investigators interested in pursuing global genomic analyses, we have created a portal to pre-computed, post-processed whole-genome comparative data that allows the extraction of ECRs, and promoter sequences as well as the TFBS associated with them, for all available vertebrate genomes.

2 RESULTS

ECRbase includes ECRs identified in pairwise alignments of publicly available vertebrate genomes. The database is created on a platform that allows for constant growth to accommodate the dynamic nature of genome research where newly emerging genomes and improved releases of current genomes are constantly made available to the public (see Supplementary material for technical details on implemented data extraction and analysis methods).

Currently, it includes data generated from 10 vertebrates: human, rhesus monkey, dog, opossum, rat, mouse, chicken, frog, fugu and zebrafish. In general, the number of ECRs in pairwise genome alignments reflects the evolutionary distance separating these genomes. For example, we observe 2.3 million (M) human/rhesus macaque ECRs while only 73 thousand (k) human/fugu ECRs. An exception to this trend is observed when species with dramatically different generation times are compared. For example, while humans and dogs are phylogenetically more distantly related than humans and rodents, human/dog comparisons reveal a greater degree of sequence conservation due to the fact that rodents have a shorter generation time and therefore have had more opportunities to diverge (Kirkness et al., 2003; Waterston et al., 2002).

Correspondingly, the ECR coverage for the non-repetitive part of the human genome decreases 65-fold as we move from the most closely to the most distantly related genome—from 53.3% in human/rhesus macaque to 0.8% in the human/fugu comparison (Fig. 1). In contrast to the human genome, the variation in the number of ECRs and the genome coverage is relatively small for vertebrates occupying distant and distinct niches in the evolutionary tree. Consistent with this observation, the number of ECRs in the fugu genome slightly
ECRbase—a database of evolutionary conservation

Fig. 1. Coverage of the human genome by ECRs from different species comparisons. ECR pie-charts indicate the percentage of elements that bin to different functional categories (coding, UTR, putatively coding)—those that overlap only with an mRNA exon, or noncoding) accompany each interspecies comparison. Annotation of coding exons and UTRs is created using available RefSeq and knownGene annotation from the UCSC Genome browser (Hinrichs et al., 2006; Pruitt et al., 2005).

varies from 67 to 74 k in comparison to six other vertebrate genomes (Supplementary Table S1).

In general, the decrease in the number of ECRs observed as the evolutionary distance increases is different for coding and noncoding regions. For example, while >80% of ECRs shared by mammals are noncoding in nature, >75% of ECRs shared between humans and either fish or amphibians are coding (Fig. 1). It has been previously reported that noncoding elements that are deeply conserved throughout the evolution of vertebrates have particular DNA signatures (Ovcharenko et al., 2004; Prabhakar et al., 2006) and are tightly linked to developmental and transcription factor genes (Woolfe et al., 2005). To account for variation in divergence rates, the analysis of noncoding ECRs that flank genes from different functional categories requires the ability to dynamically select the species to be compared in loci evolving at different rates. Therefore, the availability of multiple genome comparisons provided by the ECRbase comes with an additional value: it allows users to customize searches by selecting the most informative species for different loci in a genome.

ECRbase also provides users with detailed synteny structure interconnecting each pair of genomes. The identified synteny blocks are based on nucleotide alignments, not on protein similarity, and thus are capable of demarcating synteny breakpoints in long intergenic regions devoid of protein coding genes. This feature is particularly useful when analyzing synteny in gene deserts, which are vast intergenic regions >500 kb in length that comprehensively cover >25% of the human genome and whose function is not yet fully understood.

While transcription is known to depend on promoter function—a paradigm that has long been established (Thomas and Chiang, 2006)—increasing lines of evidence highlight the importance of highly conserved cis-acting regulatory elements that are positioned at great distances from the genes they regulate (Ghanem et al., 2003; Nobrega et al., 2003). To generate a resource that is all inclusive, ECRbase is not restricted to the analysis of promoter sequences, instead it comprises all ECRs in any available genome. ECR annotation includes the length and percent identity values that allow users to select the most conserved ECRs in a locus of interest. Also, automatically pre-computed lists of the most conserved, core ECRs (Ovcharenko et al., 2004) are made available and can be used as candidate regulatory elements in highly conserved loci.

Sequence analysis of noncoding ECRs and promoter elements is essential for searching for gene regulatory elements. Since the understanding of gene regulatory mechanisms requires the identification of transcription factors binding and acting on transcriptional regulatory elements, ECRbase provides detailed annotation of TFBS across all ECRs and promoter elements stored in the database. TFBS are identified using available libraries of transcription factor binding motifs represented as position weight matrices (PWM) from the most recent version of the TRANSFAC database (Matys et al., 2006) in combination with the previously described tSearch TFBS mapping algorithm (Ovcharenko et al., 2005). Transcription factors favor binding to short DNA motifs that usually range from 6 to 12 bps in length. Because of the highly degenerate nature of TFBS, it has been shown that computational annotation of TFBS can result in a large number of false positive predictions. To partially overcome this problem we are using a previously published method to decrease the number of false positive predictions by increasing the thresholds of TFBS mapping, such that the number of TFBS predictions is limited to ≤5 TFBS per 10 kb of random sequence (Ovcharenko et al., 2005). Although the application of these thresholds reduces the number of false positive predictions by an order of magnitude, still its application to entire genome datasets results in the identification of 4.8 and 73.5 M TFBS in human promoters and human–mouse ECRs, respectively. Therefore, statistical post-processing may be required to select TFBS that have a high likelihood of being functional. One post-processing strategy is to focus on associations of TFBS that are enriched in regions flanking functional or co-expressed genes (Qin et al., 2003). The ECRbase provides ECR information for both sequences being compared, therefore, the overlap of TFBS cohorts in orthologous ECRs could allow for the identification of actively conserved TFBS using phylogeny as a filter.

DATABASE NAVIGATION

All compiled data are publicly available and downloadable with a mouse click or through automated data fetching utilities. The files are formatted as compressed, tabulated text files.

ACKNOWLEDGEMENTS

G.G.L. and I.O were supported by LLNL LDRD-04-ERD-052 grant; and I.O. was in part supported by LLNL LDRD-06-ERD-004 grant. The work was performed under the auspices of the United States Department of Energy by the University of California, Lawrence Livermore National Laboratory Contract W-7405-Eng-48.

Conflict of Interest: none declared.

REFERENCES
