Genomic analysis of regulatory regions in organelle DNA

Ruy Jauregui Sandoval ruy@kuicr.kyoto-u.ac.jp.edu
Masumi Itoh itoh@kuicr.kyoto-u.ac.jp

Minoru Kanehisa
kanehisa@kuicr.kyoto-u.ac.jp

Bioinformatics Center, Institute for Chemical Research, Kyoto University,Uji, Kyoto 611-0011, Japan

Keywords: Transcriptional regulation, DNA curvature, mitochondria, chloroplast.

Introduction

The prediction and analysis of transcriptional regulatory regions in DNA are of central importance to understand the mechanisms governing gene expression. Computational approaches to identify promoters and regulatory sites in DNA have been mainly directed to the statistical detection of overrepresented "words" on the DNA sequence. Various algorithms and strategies based on weight matrices and Hidden Markov Models have been successfully applied to identify biologically relevant regulatory sites (1).

The availability of many completely sequenced genomes allowed the identification of clusters of orthologous sequences present in different organisms. Comparative analysis of these orthologous groups offers a novel possibility for the study of candidate regulatory elements, by the localization of conserved sequence signals in the upstream regions of these genes. The presence of a common signal in several organisms would imply a similar regulatory mechanism, which is based on the premise that regulatory elements have been conserved during evolution due to functional constrains.

It is known that some regulatory proteins strictly recognize a conserved sequence, but others bind to less conserved sites and in certain cases regulatory proteins are known to recognize bent DNA structures regardless of sequence conservation (2). Exhaustive knowledge on the structure of the DNA molecule has lead to the development of programs that predict the spatial structure of a given DNA sequence (i.e. DNA curvature). This technology enables us to identify the sites which may have a role in transcriptional regulation due to their conserved structural profile (similar bent sites). These regulatory elements usually couldn't be detected by cross-species orthologous gene comparison because of poor or nonexistent consensus sites. Both of these sequence-based and structure-based approaches have successfully identified biologically relevant signatures in several organisms (2,3).

Complete sequences of 433 mitochondrial, and 27 chloroplast genomes are publicly available, and the transcriptional regulation of their genes is not yet fully understood. This offers the possibility of making a multi-genome analysis of their regulatory regions through the use of clusters of orthologous proteins. Using this data set and implementing sequence-based and structure-based approaches, we performed an analysis to identify three kinds of conserved signals: a) sequence based signals, b) sequence based signals that present a conserved structure and c) structure based signals without sequence conservation.

Methods

Extraction of putative regulatory regions from orthologous genes

Clusters of orthologus proteins from genomic data were obtained by using the Smith-Waterman alignment algorithm, and identifying the bidirectional best hit for every protein pair between different genomes. Clusters were derived by single and complete linkage methods, resulting in 322 independent clusters for mitochondrial and chloroplast proteins. By using a similar procedure, 1123 clusters of ribosomal and transfer RNA were obtained.

Putative regulatory regions were assigned to orthologous genes based on the annotations of the corresponding Genbank files, where the length of the intergenic upstream region was at least 200 nt. Genes with intergenic distances under this value, and with similar functional annotations were assumed to be in operons, and were associated with the regulatory region of the first gene in the operon. The regulatory regions were extended up to 400 nt. upstream of the start codon, since most of the regulatory elements are present within this interval, and it facilitates the identification of structural signals.

Identification of sequence and structural signals

Overrepresented motifs present in the upstream regions were searched using the MEME algorithm (4), with an E-value threshold of 10^{-10}. Only clusters with at least 5 organisms represented were searched.

In order to identify significant curvature sites, the mean and standard deviation values of the complete genome's DNA curvature profile were calculated (5). Curvature values in the regulatory regions were searched for local maximums. If these maximums were above 3 standard deviations form the genome's mean, they were annotated as putative structural signals.

Results

Preliminary results indicate 3 mitochondrial RNA clusters with both a conserved sequence signature and a bent site. These signals are coincident with the control region in 98 animal mitochondrial genomes, this region is known to contain the origin of replication and many regulatory sites (6). Surprisingly, the structural signals are not coincident with the identified motifs.

Twenty chloroplast protein clusters also present a significant number of both structural and sequence signals, among them five clusters are ribosomal proteins, five are related to oxidative phosphorylation, and four are photosystem enzymes. Comparison of the photosystem genes' regulatory regions against the Plant cis-acting regulatory DNA elements (PLACE) database (7) revealed promoter regions and elements related to the circadian cycle. Further characterization and bibliographical validation of these regulatory elements will demonstrate their biological relevance.

References

[1] Qiu, P. Recent advances in computational promoter analysis in understanding the transcriptional regulatory network. *Biochem. Biophys. Res. Commun*. 309(3):495-501, 2003.

[2] Jauregui, R., Abreu-Goodger, C., Moreno-Hagelsieb, G., Collado-Vides, J. and Merino E. Conservation of DNA curvature signals in re gulatory regions of prokaryotic genes. *Nucleic Acids Res.* 31(23):6770-6777, 2003.

[3] Guo, H. and Moose, S.P. Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. *The Plant Cell*, 15: 1143-1158, 2003.

[4] Bailey, T.L., and Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proc. 2nd IC-ISMB*, 28-36, 1994.

[5] Shpigelman, E.S., Trifonov, E.N., and Bolshoy, A. CURVATURE: software for the analysis of curved DNA. *Comput. Appl. Biosci.* 9(4):435-440, 1993 .

[6] Fernandez-Silva, P., Enriquez, J.A., and Montoya, J. Replication and transcription of mammalian mitochindrial DNA. *Exp. Physiol*. 88(1):41-56, 2003.

[7] Higo, K., Ugawa, Y., Iwamoto, M., and Korenaga, T. Plant cis-acting regulatory DNA elements (PLACE) database: 1999, *Nucleic Acids Res*. 27(1):297-300. 1999.