Conservation rate of transcription factor binding sites in Saccharomyces genomes

Kovaleva G.(1,2), Bazykin G.(3), Brudno M.(4), Gelfand M.(1,2).

(1) Department of Bioengeneering and Bioinformatics, Moscow State University, Moscow, Russia (2) Institute for Information Transmission Problems, RAS, Moscow, Russia. (3) Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey, USA. (4) Department of Computer Science, Stanford University, Stanford, USA.

Motivation

Extracting the complete functional information encoded in a genome, including genic, regulatory and structural elements, is a central challenge in biological research. Prediction of non-protein-coding functional regions, such as regulatory elements, is especially difficult because they are usually short (6-15 bp for S.cerevisiae and many other eukaryotic genomes), often degenerate, and can reside on either strand of DNA at variable distances from the genes they control. Since functional sequences tend to be conserved through evolution, they can appear as 'phylogenetic footprints' in alignments of genome sequences of different species. Recently, two groups sequenced several Saccharomyces genomes. The main goal of these studies was to identify the regulatory sites in Saccharomyces spp. using multiple whole-genome alignment or and multiple alignments of gene upstream regions. Results were represented as two lists of predicted binding motifs. Our comparison of these lists shows a rather moderate intersection. This prompted us to analyze the conservation rate for known and predicted binding sites in Saccharomyces genomes in more detail.

Methods

Complete genome sequence of Saccharomyces cerevisiae was extracted from GeneBank. Fragments covering 750 nucleotides of upstream regions and 150 nucleotides of protein-coding genes were considered. Search for orthologs was done using fungiBLAST(http://www.ncbi.nlm.nih.gov/BLAST/Genome/FungiBlast.html). Protein-coding genes without indentifiable orthologs were ignored. Conversely, regions upstream of orthologous genes were used even if they did not produce a strong alignment. For identification of site patterns, Genome Explorer program was used. SignalX was used to construct positional nucleotide weight matrices. Multiple sequence alignments were done using ClustalX. Transmembrane domains were predicted using servers TMHMM (http://www.cbs.dtu.dk/services/TMHMM) and PSORT (http://psort.nibb.ac.jp). Multiple whole-genome alignments were done using Lagan.

Results

We investigated conservation rates of binding sites for transcriptional regulators of two metabolic pathways, biosynthesis of methioinine and leucine. Both pathways are regulated by the global regulator of amino acid biosynthesis Gnc4p, and pathway-specific regulators Met31/Met32 and Cbf1/Met4/Met28 regulatory complexes for the methionine biosynthesis and Leu3p for the leucine biosynthesis. We show that conservation rates of known and strong predicted binding sites for all studied regulators are similar. Still, even strongest sites are not necessary conserved in all examined genomes, contrary to the initial assumption. Using the

comparative analysis of regulation, we have identified the candidate for the alpha-isopropylmalate carrier, transporting an intermediate of leucine biosynthesis from the mitochondrial matrix to the cytosole. This prediction was based on conservation of binding site for Leu3p, protein-DNA binding data produced by large-scale ChIP on chip experiments and analysis of the protein similarity. Using multiple whole-genome alignments of seven Saccharomyces genomes, we studied conservation rates of binding sites longer than 5 bp for all transcription regulators of S.cerevisiae present in TRANSFAC database. As expected, the conservation rate of transcription factors binding sites was much higher than conservation rate of non-regulatory part of upstream region. However, high conservation rate was observed not only in the core region of the binding site, but also in its neighborhood extending to approximately 50 bp. One possible explanation for this observation is that the regions adjacent to known binding sites contain binding sites for other transcription factors or promoter sites. This study was partially supported by Howard Hughes Medical Institute and the Russian Academy of Sciences (programs “Molecular and Cellular Biology” and “Origin and Evolution of the Biosphere”).

Contact email: kovaleva@iitp.ru