Phylogenetic Comparison of Human and Mouse Promoter Regions and Visualization of a Variety of Conserved Elements

Hirokazu Chiba  
hchiba@ims.u-tokyo.ac.jp

Kenta Nakai  
knakai@ims.u-tokyo.ac.jp

Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

Keywords: comparative genomics, promoter region, local alignment

1 Introduction

To extract biologically functional information from promoter regions is a difficult problem. Searching for potential Transcription factor binding sites (TFBS) by using TRANSFAC database [2] is a widely used method, but it is not so effective because the majority of obtained candidates seem to be false positives. Eugene et al. developed CONREAL algorithm [1], which can reduce the false positives of the TFBS search by finding a series of TFBSs that are arranged in the same order in an orthologous pairs of sequences. In compensation for the specificity, however, the CONREAL algorithm neglects TFBSs that are changed in the order in the query sequences. Although there are pioneering studies phylogenetically comparing genomic sequences [5], or comparison of human and mouse promoter regions [4], our understanding of the structures of promoter regions is still at a very primitive stage. To see what kind of characteristics tends to be conserved, we compared human and mouse orthologous promoter sequences, and visualized various types of data.

2 Methods

We obtained human and mouse transcription start site (TSS) information from DBTSS [3]. We used sequences of 1200 bp length covering from 1000 bp upstream to 200 bp downstream of the TSS. Among them, 2,084 sequences could be paired as promoter regions of orthologous genes using the table of human and mouse homologous genes [7]. LALIGN [6] was used to align promoter sequences with the default settings. LALIGN outputs several local alignments allowing overlaps. Among them, we selected alignments with higher E-values so that they are not overlapped each other. CONREAL [1] was used to find conserved series of the TFBSs. GC content in the promoter regions was calculated with 100 bp of window size. In the 2,084 pairs of human and mouse promoter pairs, mouse sequences are shuffled in the order, and generated pairs are used as negative control set.

2 Results and Discussion

The results of LALIGN, CONREAL and calculated GC content were visualized together. Some examples are shown in Fig. 1. Comparison of promoter regions of orthologous genes are shown in (A). Comparison of negative controls is shown in (B). Filled regions in the Figures represent local alignments generated by LALIGN. Bars represent results of CONREAL. Line graphs show GC content. Small boxes on the horizontal lines represent the TFBSs. E-values of local alignments, similarity of TFBSs between human and mouse and scores of weight matrices for the putative TFBSs are shown with a gradation of shades in the figures.

Suzuki et al. [4] found a homologous block structure in the upstream regions of human and mouse orthologous genes and that the average length of the blocks is 510 bp, but it is a too simplified view, for varied length of alignments exist. The structures of the conserved elements are so complicated and more detailed analyses remain to be done. This study is a foothold to those analyses and, furthermore, to a more effective interpretation of the biological content of promoter regions.
Figure 1: Comparison of promoter regions of human and mouse. (A) Promoter regions of orthologous genes. (B) Negative controls

References


