Development of Comparative Genome Analysis Tool SHOE for Regulatory Regions of Higher Eukaryotes

Natalia Polouliakh
AIST Research Scientist
Sequence Analysis Team

Life exists on earth in a myriad of forms encompassing more than ten million species. Despite the fact that the more closely related two species are, the more genes they have in common (human and mouse, for example, share about 50% of their genes), it is a great mystery why our bodies (phenotypes) are so different. We believe that the secret is hidden in differences in the expression patterns of the specific genes, such as brain-specific genes and transcription factors, and these differences are responsible for the diversity observed amongst higher eukaryotes. It can be said that transcription factors, which account for about 10% of the protein-coding genes in humans, hold the key to this mystery of life.

Cross-species comparison of regulatory regions is very helpful in revealing the presence a functionally conserved or non-conserved gene groups. For the latter group of genes, however, recent evidence has shown that altered roles are more commonly due to altered regulatory regions than to the changes in their protein coding sequences.

We are developing an analytical tool, "SHOE-Sequence HOmology in higher Eukaryotes" (http://shoe.cbrc.jp) (Fig.1), for comparing the genomic control region between human and rodents, which diverged 75-80 million years ago. Transcriptional regulatory regions in higher eukaryotes include promoters adjacent to the transcriptional start sites and enhancers located thousands of nucleotides away, where activators or inhibitors bind (Fig. 2). In light of this, SHOE is designed to handle regions up to 100,000 bases in length. Comparing to other tools, SHOE is characterized by its abilities to combine the motif finding by position-specific scoring matrices (PSSM) with the expression analysis of available microarray data for the genes included in the dataset.

For each co-expressed human gene pair SHOE checks if highly conserved genomic regions exist across human, mouse and rat and then perform searches for shared transcription factor binding sites on those regions. In addition, SHOE can use the expression data of transcription factors themselves to further validate the functional significance of candidate transcription binding factor sites.

For expression analysis we developed a tool for extracting co-expressed genes from GEO derived Cell Montage (cellmontage.cbrc.jp) data and for motif finding, PSSMs from the TRANSFAC and JASPAR databases. In the future we plan to add a heuristic step for the discovery of novel motifs.

Reference

Polouliakh, N., Natsume, T., Harada, H., Fujibuchi, W., Horton, P.: "Comparative genomic analysis of transcription regulation elements involved in the human MAP-kinase G-protein coupling pathway", *Journal of Bioinformatics and Computational Biology*, **4** (2), pp. 469-482 (2006).

Suzuki, Y., Yamashita, R., Shirota, M., Sakakibara, Y., Chiba, J., Mizushima-Sugano, J., Nakai, K., Sugano, S.: "Sequence comparison of human and mouse genes reveals a homologous blocks structure in the promoter regions", *Genome Res.*, **14**, pp.1711-1718 (2004).
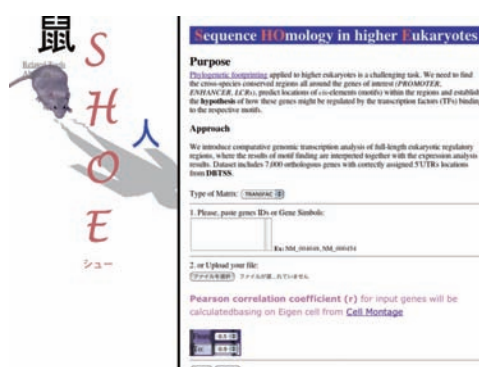
Fig.1 Home page for SHOE
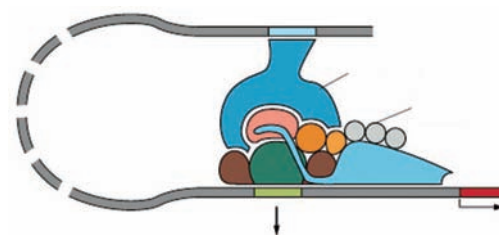
Fig.2 Eukaryote transcription regulation mechanism (adapted from http://www.csu.edu.au/)