ReView: A Viewer to Display the Relation of Protein-Coding Region and Repetitive Patterns

Yu-Tin Chen[1,2] g870811@ym.edu.tw
Yu-Chung Chang[1,2] d80106@ym.edu.tw
Yen-Hua Huang[2,3] binfo@ym.edu.tw
Ueng-Cheng Yang[1,2,3] yang@ym.edu.tw

[1] Institute of Biochemistry, National Yang-Ming University, Taipei, Taiwan 112
[2] Bioinformatics Center, National Yang-Ming University, Taipei, Taiwan 112
[3] Bioinformatics Program, National Yang-Ming University, Taipei, Taiwan 112

Keywords: pattern, repetitive sequence, genome, distribution, COG

1 Introduction

Data mining methods may discover some interesting correlation. However, its biological significance of a selected pattern will be more obvious if it is presented graphically. Human eyes are good at recognizing sequence patterns in graphical form. Visualization methods can display patterns at different scales. At one extreme, the relation of genes and patterns, such as transcription factor binding sites, can be observed locally. At the other extreme, the distribution of a pattern in a genome can be observed globally. This kind of flexibility is useful for exploring the biological significance of a sequence pattern.

2 Material and Methods

We have constructed a visualization tool to display a sequence pattern. There are two interfaces are available. The first interface (http://rnd.ym.edu.tw/review0/) computes the positions of a pattern that is less than 20 base pair in real time. In a second interface (http://rnd.ym.edu.tw/review/), the positions of a pattern that is 4-12 base long are retrieved from a database. The latter version was designed to couple with data mining tools. Both systems were implemented by perl programming language and the database version used mysql as management system. Currently, only the microbial genomes were linked. The protein coding information ("ptt" files) of microbial genomes were downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/Bacteria). The role of each coding region was assigned on the basis of COGs (Clusters of Orthologous Groups of proteins [2, 3]). The 19 functional categories were color coded according to the standard colors used by COGs.

3 Results and Discussions

ReView is developed to view the distribution of patterns relative to genes. Suppose, you find an interesting pattern AAAAAC using data mining tools. You can easily see that this pattern occurs asymmetrically around the replication origin in *B. burgdorferi* (the top panel in Fig. 1(a). There are 8 rows, which display the distribution of the given pattern in the genome, under the top panel. The first two rows simply plot the positions of the given pattern in the forward and reverse strands. The next 3 and the last 3 rows within this block presented the details of the forward and reverse strands, respectively. The detail information indicated that the pattern AAAAC distribute evenly in the coding region, but asymmetrically in the noncoding region. The relative position of genes and patterns were shown in the lower part of Fig. 1(a). A zoom in view of this region was shown in Fig. 1(b), in which each horizontal bar represented a coding region. The red and blue vertical lines mark the positions of the given pattern in forward and reverse strands, respectively. The COG functional class of each

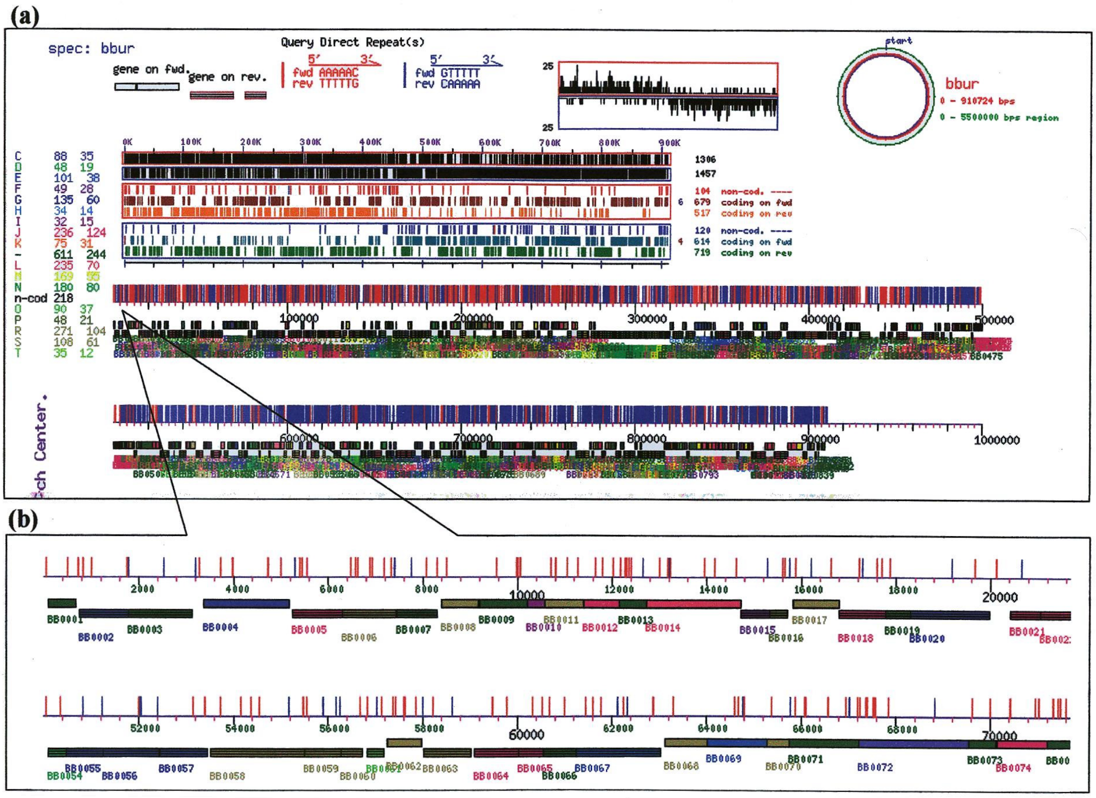

Figure 1: ReView displays the distribution of pattern AAAAAC and the protein coding genes in *B. burgdorferi*. (a) global view (b) zoom-in view.

coding region was represented by color. The meaning of color code was shown in the left of Fig. 1(a). After a user browses many patterns in ReView, he may find a few interesting pattern. If the user would like to browse the results on his personal computer offline, he can save results in Artemis[1] format. In Artemis, the user will be able to import different kinds of information and integrate them together even. The advantage of ReView is that you can browse the distribution of different patterns interactively, because ReView computes patterns in real time or retrieves patterns from a database.

4 Acknowlegements

YTC and YHH are supported by a grant from National Science Council, ROC (NSC 90-2318-B-010-003- M51). YCC is supported by another grant from National Science Council, ROC (NSC 90-2323-B-010-004). The computational resources are supported by Ministry of Education, ROC (Program for Promoting Academic Excellence of Universities, 89-B-FA22-2-4).

References

[1] Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B., Artemis: sequence visualization and annotation, *Bioinformatics*, 16:944–945, 2000.

[2] Tatusov, R.L., Koonin, E.V., and Lipman, D.J., A genomic perspective on protein families, *Science*, 278:631–637, 1997.

[3] Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V., The COG database: new developments in phylogenetic classification of proteins from complete genomes, *Nucleic Acids Res.*, 29:22–28, 2001.