Genomics & Informatics Vol. 3(2) 52-54, June 2005

BioCovi: A Visualization Service for Comparative Genomics Analysis

Jungsul Lee[1], Daeui Park[2]* and Jong Bhak[3]*

[1]Department of BioSystems, KAIST, Daejeon 305-701, Korea, [2]Object Interaction Technologies Inc., Daejeon 305-701, Korea, [3]NGIC, KRIBB, Daejeon 305-806, Korea

Abstract

Visualization of the homology information is an important method to analyze the evolutionary and functional meanings of genes. With a database containing model genomes of *Homo sapiens*, *Mus muculus*, and *Rattus norvegicus*, we constructed a web-based comparative analysis tool, BioCovi, to visualize the homology information of mammalian sequences on a very large scale. The user interface has several features: it marks regions whose identity is greater than that specified, it shows or hides gaps from the result of global sequence alignment, and it inverts the graph when total identity is higher than the threshold specified.

Key words: comparative genomics, visualization
Availability: http://bioinfra.ngic.re.kr/cgi-bin/index.cgi?page=bioCovi or http://bioviewer.org

Summary

Since 1995, over 250 genomes have been completely sequenced (Bernal *et al.*, 2001; Shendure *et al.*, 2004). Although the number of known genome sequences increased rapidly, we still need to mine information about gene function, gene regulation, and the origin of genes. To extract information in genomes, we need a comparative analysis for large scale genomes. In the past, the comparative analysis of eukaryotic genomes, such as mouse and human, resulted in discovering novel conserved gene regions, novel non-coding regions, and gene regulatory sequences (Pennacchio and Rubin, 2001; Rubin *et al.*, 2000). In the comparative analysis of completely sequenced genomes, visualization is helpful as the sequences are very long and complicated in terms of their organization.

Currently, there are several popular visualization tools for sequence alignments. Dotplot (Sonnhammer and Durbin, 1995) can display sequence homology in two genomes, repeat sequences, and palindromes in the genomes in UNIX X server environment. There are also several visualization tools using ASCII characters. APOLLO (Lewis *et al.*, 2002) visualizes gene annotation using GFF (General Feature Format) file. PipMaker (Schwartz *et al.*, 2000) visualizes conserved regions in two genomes. However, it is difficult for these tools to represent information of conserved coding-region and non-coding region such as promoters in genomes. Consequently tools such as Mulan (Ovcharenko *et al.*, 2005) and VISTA (Frazer *et al.*, 2004) were developed. These tools implemented the visualization of annotated information such as gene position, gene structure, and repeat region on a whole genomic comparative analysis perspective.

Our program visualizes the annotation of a genome and sequence homology among genomes using BLAT (Kent, 2002), AVID (Bray *et al.*, 2003), and VISTA, with some added features. It efficiently visualizes the information obtained from comparative genomic analyses. The main features added are: 1) gap option: gaps can be shown or hidden. In global sequence alignments, gaps mean either an insertion or a deletion. This option makes the visualization clearer by identifying conserved regions of reference and the query sequences; and 2) inverting graphs: when the mutual alignment identity of sequences is high, it is more effective to see where the sequences are different as in the case where SNP (single nucleotide polymorphism) locations are important. Thus we invert a graph when the total identity is higher than a certain threshold specified, as people usually use the upper part in a sequence graph. By integrating AVID, BLAT, and VISTA programs and making a graphical visualization, we can obtain information from coding regions very easily and can achieve a more elaborate comparative genomics analysis.

Comparing genomes using BioCovi: For the comparison of up to 10 genomes, one can load FASTA sequence format files. The BioCovi runs the above mentioned algorithms internally and shows the text based results in sequence pictures on the web server. The web server is

*Corresponding author: E-mail jong@kribb.re.kr,
daeui@oitek.com,
Tel +82-42-879-8500, Fax +82-42-879-8519
Accepted 7 March 2005

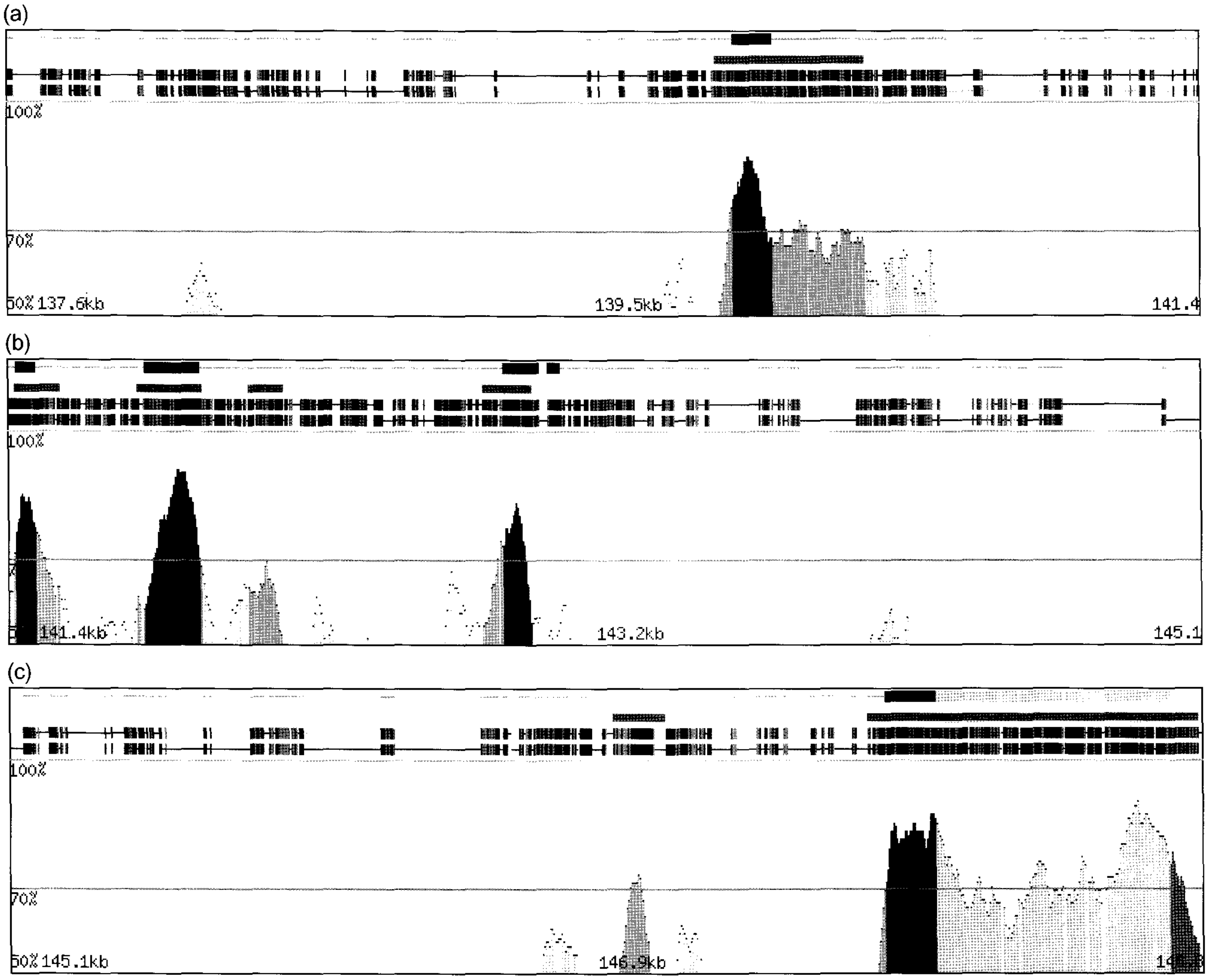

Fig. 1. BioCovi results for the NFkB gene.

The result of comparative analysis on NFkB gene in human and mouse genomes. Red bars mean conserved regions. Blue bars at the top-most line in each graph mean exons. It is evident that exon regions are conserved in the two genomes as well as UTR region of yellow.

a part of BioInfra (http://bioinfra. org or http://bioinfra.ngic.re.kr), a very large bioinformatics infrastructure containing various steps of genome analysis. The major advantage of BioCovi is that it can reveal exon and intron information using known reference sequence.

Results

We carried out a comparative analysis on the NFkB gene in human and mouse genomes. The gene resides in the chromosome four at the 103,779,572 location in human and in the chromosome three at the 136,131,339 location in mouse. The following graphs Fig. 1 are taken from the BioCovi. The length of the human NFkB gene is 116,000 base pairs and that of the mouse NFkB gene is 107,000 base pairs.

The top-most line, the annotation field, in the graphs a, b, and c shows exons, introns, and UTRs. Exons, introns and UTRs are blue, cyan, and yellow. Red bars of the second line, are conserved region. The third and fourth lines, alignment fields, represent the result of global alignment. Bar graphs in the lower part of graphs a, b and c are identity fields. The identity is the ratio of matched nucleotides in a window, which has the point at the center.

Users can see the global alignment results intuitively with BioCovi. It can be recognized easily that UTR and exons are conserved in the two. On the other hand, there are many gaps in the intron regions. The result of comparative analysis with BioCovi is very intuitive and informative.

Acknowledgements

We thank our colleagues at Biomatics Lab, especially Mr. Chung MoonSoul for donating 25 million dollars for BioSystems research and Semin Lee. This project was supported by R01-2004-000-10172-0 grant of KOSEF and Biogreen21 program of RDA. We thank Maryana Bhak for editing the manuscript.

References

Bernal, A., Ear, U., and Kyrpides, N. (2001). Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res.* 29, 126-127.

Bray, N., Dubchak, I., and Pachter, L. (2003). AVID: A global alignment program. *Genome Res.* 13, 97-102.

Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32, W273-W279.

Kent, W.J. (2002). BLAT-the BLAST-like alignment tool. *Genome Res.* 12, 656-664.

Lewis, S.E., Searle, S.M., Harris, N., Gibson, M., Lyer, V., Richter, J., Wiel, C., Bayraktaroglir, L., Birney, E., Crosby, M.A., Kaminker, J.S., Matthews, B.B., Prochnik, S.E., Smithy, C.D., Tupy, J.L., Rubin, G.M., Misra, S., Mungall, C.J., and Clamp, M.E. (2002). Apollo: a sequence annotation editor. *Genome Biol.* 3, RESEARCH0082.

Ovcharenko, I., Loots, G.G., Giardine, B.M., Hou, M., Ma, J., Hardison, R.C., Stubbs, L., and Miller, W. (2005). Mulan: multiple-sequence local alignment and visualization for studying function and evolution. *Genome Res.* 15, 184-194.

Pennacchio, L.A. and Rubin, E.M. (2001). Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.* 2, 100-109.

Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W., Cherry, J.M., Henikoff, S., Skupski, M.P., Misra, S., Ashburner, M., Birney, E., Boguski, M.S., Brody, T., Brokstein, P., Celniker, S.E., Chervitz, S.A., Coates, D., Cravchik, A., Gabrielian, A., Galle, R.F., Gelbart, W.M., George, R.A., Goldstein, L.S., Gong, F., Guan, P., Harris, N.L., Hay, B.A., Hoskins, R.A., Li, J., Li, Z., Hynes, R.O., Jones, S.J., Kuehl, P.M., Lemaitre, B., Littleton, J.T., Morrison, D.K., Mungall, C., O'Farrell, P.H., Pickeral, O.K., Shue, C., Vosshall, L.B., Zhang, J., Zhao, Q., Zheng, X.H., and Lewis, S. (2000). Comparative genomics of the eukaryotes. *Science* 287, 2204-2215.

Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. (2000). PipMaker-a web server for aligning two genomic DNA sequences. *Genome Res.* 10, 577-586.

Shendure, J., Mitra, R.D., Varma, C., and Church, G.M. (2004). Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.* 5, 335-344.

Sonnhammer, E.L. and Durbin, R. (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 167, GC1-GC10.