

A Database and Browser for Genome Analysis and cDNA Assembly

Yuan Liu*, Yuhong Wang*, Guochun Xie, Yu Lin# and Richard Blevins#

Department of Bioinformatics, Merck & Co., Inc. WP42-300, West Point, PA 19486.

#RY80-A1, Rahway, NJ 07065 yuan_liu@merck.com richard_blevins@merck.com

The flood of human genomic sequence demands new technologies for its characterization. To extract information from genomic sequence data and to facilitate gene discovery, we have developed a system for data mining genomic sequences. The system is integrated, flexible and upgradable and provides scientists with the most up-to-date information. All data is stored in an Oracle relational database (Xie et al., 2000). A set of interactive visualization tools has been developed using the Java programming language. The software environment can be either deployed as a stand-alone application or be used as a direct window into the Oracle database. This data mining tool-set is designed to enable laboratory bench scientists to identify and assemble novel human cDNAs *in silico* (paralogs or orthologs of known genes) from genomic sequences. For instance, since the Merck Gene Index (MGI) is indexed through 3' EST sequences (Eckman, et al., 1998), mapping MGI onto human chromosomal sequences brings a wealth of information about each EST cluster chromosome location and possible splicing variants. One may also study promoter regions of a given gene, map human cDNA onto chromosomes, visualize gene structure, and identify disease gene candidates. Using this GenoCloner, a number of novel human genes were identified and cloned.

References

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.

Eckman BA, Aaronson JS, Borkowski JA, Bailey WJ, Elliston KO, Williamson AR, Blevin, RA (1998) The Merck Gene Index browser: an extensible data integration system for gene finding, gene characterization and EST data mining. *Bioinformatics* 14(1):2-13.

Padgett, Richard A., Grabowski, Paula J., Konarska, Maria M., Seiler, Sharon, and Sharp, Phillip A. (1986) Splicing of Messenger RNA Precursors. *Ann. Rev. Biochem.* 55:1119-1150.

Wang, Yuhong., Liu, Yuan., Lin, Yu., Blevins, Richard., and Xie, Guochun. (2000), A Tool for Storing and Visualizing Sequences. Submit to

Xie, Guochun., DeMarco, Reynold., Blevins, Richard., and Wang, Yuhong. (2000). Storing Biological Sequence Databases in Relational Form. *Bioinformatics*. Vol.16 no.1:1-2.