

ORGANIZATIONAL ANALYSIS OF EXTENDED HUMAN GENOMIC CONTIG REGIONS

Eric C. Rouchka and David J. States
Institute for Biomedical Computing
Washington University
700 S. Euclid Avenue
Saint Louis, MO 63110-1012
ecr@ibc.wustl.edu states@ibc.wustl.edu

The collaborative efforts of the Human Genome Project to completely sequence the human genome by 2003 using clone-based strategies has led to massive amounts of finished human sequence data. As more data becomes available, methods to assemble and analyze full chromosomes on a sequence level are desired.

We attempt to collate a definitive set of nonredundant extended segments of human genomic sequence by extending individual human entries in GenBank greater than 10 kilobases (kb) in length by connecting adjacent overlapping regions previously sequenced. We seek to identify overlapping regions of at least 70 base pairs (bp) and an identity of greater than 98% using wu2blastn (Gish, 1994-1997). Through May 31, 2000, a total of 3,081 contigs have been assembled, covering 627,658,467 bases of finished human genomic sequence. This total represents nearly 19.5 percent of the human genome. A web interface for the human genomic contigs is available at <http://stl.wustl.edu/contigs/HUMAN/>

With the recent announcement of the completion of chromosome 22 (Dunham, et. al., 1999) and chromosome 21 (Hattori, et. al., 2000), it has become possible to start studying the organization of whole chromosomes at the sequence level. Previous density gradient centrifugation experiments (Bernardi, 1993) have suggested that regions of chromosomes are organized into heterogeneous regions called isochores. We propose methods to study segmentation of isochores at the sequence level.

In addition, studying the maintenance of isochore regions is an interesting question. We are looking into several hypotheses concerning how isochores are maintained, including maintenance by repetitive element structure, compositional biases for substitution, and the recent addition of tissue specific genes. Initial analysis using RepeatMasker defined repeat families (Smit and Green, unpublished) on chromosome 22 indicates that repetitive elements maintain a relatively constant G+C content around 44%, and thus do not maintain the G+C% of the surrounding region. Preliminary data indicate that substitutional biases may exist, with a higher concentration of C or G to A or T mutations in higher C+G regions, and higher A or T to C or G mutations in lower C+G regions of chromosome 22.

Bernardi, G. (1993) "The isochore organization of the human genome and its evolutionary history -- a review." *Gene*, **135**:57-66.

Dunham, I., et. al., (1999) "The DNA sequence of human chromosome 22." *Nature*, **402**(6761):489-495.

Gish, W., (1994-1997). unpublished.

Hattori, M., et. al. (2000) "The DNA sequence of human chromosome 21." *Nature*, **405**(6784):311-319.

Smit, A.F.A., Green, P., unpublished data. <http://repeatmasker.genome.washington.edu>