WEBMAP: radiation hybrid mapping on the WWW

William Newell1, Philippe Sanseau2, John Riley2 and Andrew Lyall2

1Oxford Molecular Ltd, The Medawar Centre, Oxford Science Park, Oxford OX4 4GA, UK, (Current address: Hoechst-ARIAD Genomics Center, 26 Landsdowne Street, Cambridge MA 02139, USA) and 2Glaxo Wellcome Medicines Research Centre, Gunnel's Wood Road, Stevenage SG1 2NY, UK

Received on June 1, 1998; revised on August 20, 1998; accepted on August 24, 1998

Abstract

Summary: A Java interface to radiation hybrid (RH) mapping software is described which enables users to build and interactively refine RH maps over the web.


Contact: William.Newell@ariad.com

A large amount of radiation mapping data is now publicly available at the RHdb (Rodriguez-Tome and Lijnzaard, 1997). The majority of this data comprises the mapping effort of an international consortium (Schuler et al., 1996), which set out to map one representative EST sequence from each of the EST clusters in UniGene (Schuler et al., 1996). The result was an assignment of each sequence to an interval between two markers on the Genethon genetic map (Dib et al., 1996). Here, we describe a Web-based mapping tool, which uses the raw data held at the EBI and mapping software developed for RH data (Newell et al., 1998) to generate EST maps. It allows maps to be rapidly constructed de novo from the raw data.

The system is distributed between a client and a server, which communicate via CGI/HTTP. Raw data input and retrieval are handled by Java client applets, and computationally intensive map construction accomplished by the server. Raw data can be entered or retrieved in three different ways. Raw data generated by the user can be copied from a data file and pasted into an applet window. Alternatively, the name of a known marker can be submitted and the nearest markers retrieved from WEBMAP’s data files. The user can also enter an experimental RH vector using a grid of checkboxes, and the nearest markers in RHdb again retrieved. The measure of ‘nearness’ is the Hamming distance, i.e. the number of hybrids with the same typing for both markers, which is fast to compute, but is not directly proportional to physical distance.

Having obtained a set of raw data, they are sent to the server as text using CGI/HTTP. The server mapping software is written in C (Newell et al., 1998), and calculates the distances between all pairs of markers using a simple model of fragment generation and retention. The distance units are the mean number of X-ray induced breaks between marker pairs (Ray, R), assumed to be proportional to physical distance. 1R on the G4 panel corresponds to ~25 Mb, and to ~5 Mb on the G3 panel. Optimal coordinates are obtained from this distance matrix using the techniques of distance geometry (DG).

The computed map is sent back to the Java client applet as text through the same CGI/HTTP connection, and is parsed into Java objects. These are then displayed in applet windows showing the locations of the markers using the DG technique, and their designated interval on the genetic map in GeneMap. The display indicates whether the data form a single well-linked linkage group, e.g. when markers are selected close to a known marker in the database. Outliers indicate the presence of a probable RH typing error. The map can be refined by selecting groups of markers with the mouse, by drawing around regions of the map. This enables separate clusters to be mapped independently of the other data, giving a more accurate local map, or exclusion of error-prone markers that skew the conformation of an otherwise good map. This process of excluding subsets, equivalently selecting good linkage groups, continues until the whole set is accurately mapped.

Other data sources have been incorporated into the system, by constructing a series of UNIX DBM databases linking synonyms in different databases. Links are made between RHdb, GeneMap, UniGene, OMIM, and Genethon. An example is the marker SHGC-12892 in GeneMap which has symbol RH9074 in RHdb, which is L25941 in EMBL, which is assigned to UniGene cluster Hs.1764 (Lamin B receptor) which has gene symbol LBR corresponding with OMIM entry number 600024. The DBM databases are used to generate flat files for each chromosome and panel. Each line in these files corresponds to a different marker, and has the format ‘RHdb, panel, chromosome, EMBL, UniGene, OMIM, RH data vector’. Of 66 334 entries in RHdb, 32 045 (48.3%) are represented in UniGene, and 3629 (5.5%) in OMIM. Absence of a corresponding entry is indicated by the symbol ‘#’. The names on the graphical display can be selected from any of the different data sources represented, and are sensitive to mouse presses. When clicked, a supplementary Web browser...
window is shown, containing the data for that marker at the remote site.

This interactive Java interface allows construction of locally accurate maps, often a starting point for positional cloning projects. It builds maps directly from the currently available raw data for ∼70,000 ESTs typed on the G3 and G4 panels. The positions of the markers calculated by DGMAP are compared with their binned locations on the genetic map (Dib et al., 1996) as reported in the transcript map (Schuler et al., 1996). Since the RH maps are generally considered to be ‘physical’, and X-ray-induced breaks are believed to be homogeneously distributed, such map comparison can indicate recombination hot- and cold-spots which may be correlated with disease (Parundare and Patel, 1997). The resolution of different names in different databases should be a useful addition to the techniques of ‘positional candidate’ cloning, since any of the ESTs can be immediately identified by their names in UniGene or OMIM (OMIM, 1997). This information is otherwise laborious to accumulate.

References


