SIGMA: *system for integrated genome map assembly*

Michael J. Cinkosky, James W. Fickett, William M. Barber, Michael A. Bridgers, and Charles D. Troup

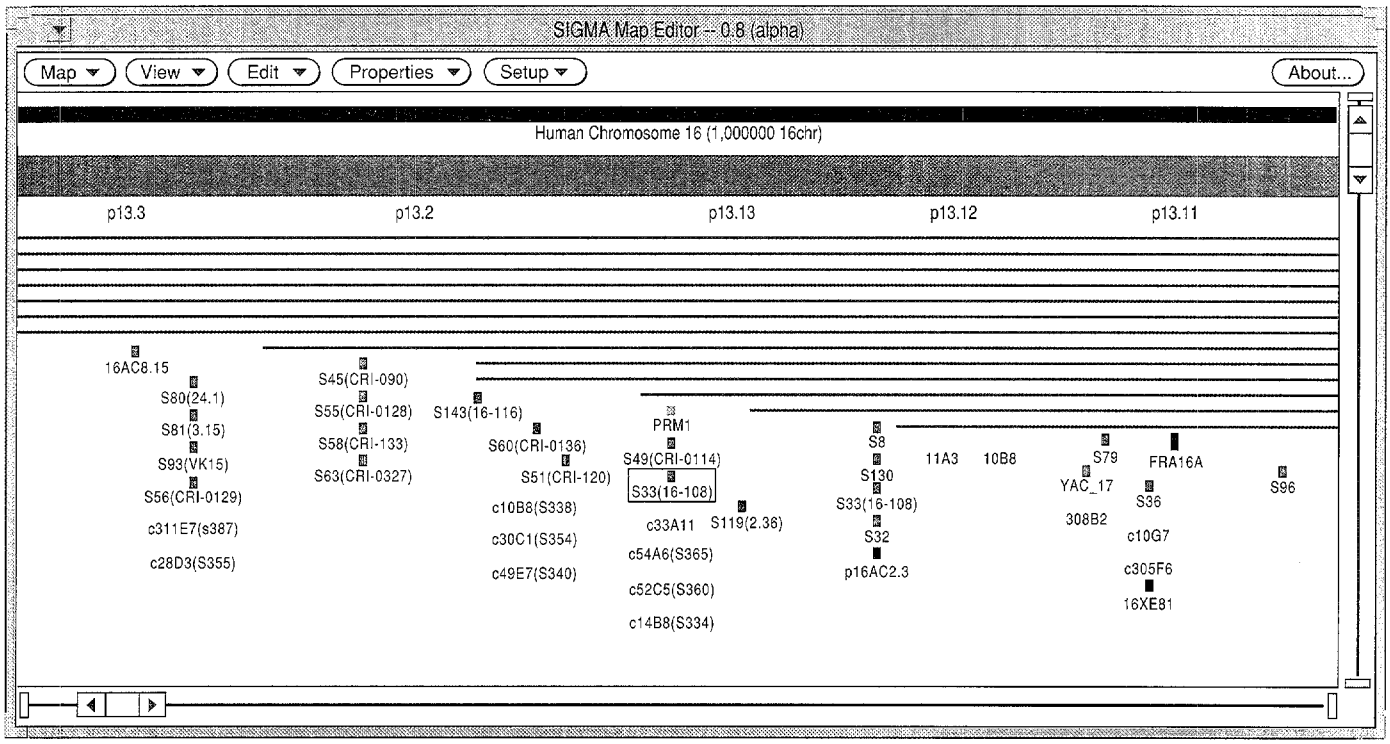
With high-quality road maps available at stores everywhere, it is easy to forget just how much effort went into the production of the first accurate geographical maps. Even maps only a few hundred years old contain glaring errors, such as the early maps of North America that show California as an island. However, when one considers how difficult it was to obtain accurate information on which to base those maps, one can understand why the maps were so inaccurate. The human genome is at present about as difficult to explore as that early wilderness was.

Although biologists have for some time been able to examine small regions in great detail, they are only now developing the experimental techniques that will allow the generation of reasonably detailed maps of each chromosome. Even now, data on the lengths of map elements and the distances between them are too fragmentary to use in building precise maps of entire chromosomes. In fact, with fragmentary data coming from many different types of experiments where even the units of measurement are incompatible, the present situation is remarkably similar to that of early cartographers who relied on the (doubtless contradictory) reports of numerous travelers returning from the area being mapped.

Unlike early explorers, however, biologists today can bring the power of computers to bear on the problem. To this end, we are producing a special-purpose tool for building accurate genome maps called SIGMA (System for Integrated Genome Map Assembly). SIGMA applies several modern ideas including object-oriented databases, optimization theory, genetic algorithms, and interactive computer graphics.

Building maps in SIGMA involves two basic activities: collecting information and drawing working maps (representations of the structure of the genome that are in reasonable agreement with experimental data). At the heart of the SIGMA system is an object-oriented database that stores all the data used in the map-building process, including all of the (potentially inconsistent) data on which the maps are based.

Maps in SIGMA can be constructed either automatically (by routines discussed below) or by users. The primary interface to SIGMA is the interactive graphical map editor shown in the figure on the next page. With this editor, users can see the positions assigned to map elements and change the positions to build or improve maps. The editor works like computer-aided drafting and design tools to let users easily view and edit the map without requiring them to understand the structure of the database in which the map is stored. Furthermore, because the software was



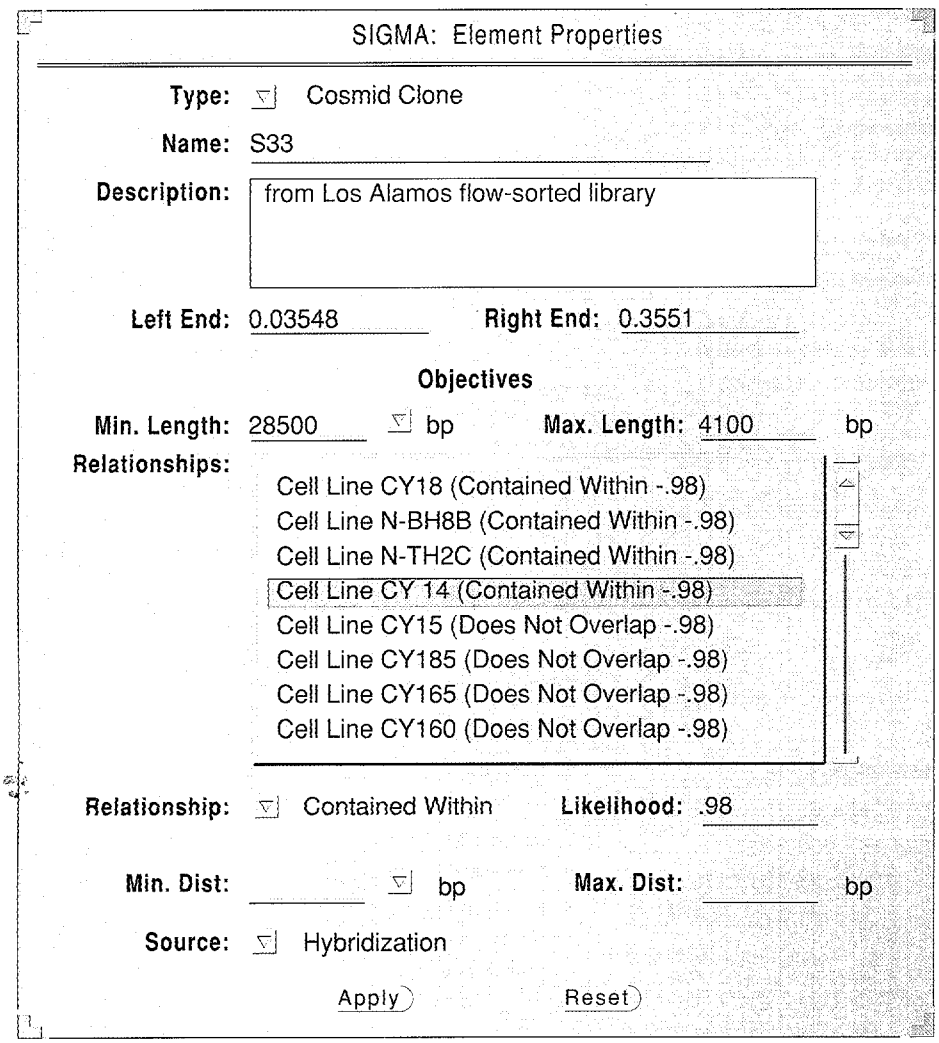
A SIGMA window is shown above as it might appear on a user's computer screen. The window contains the SIGMA map canvas, here showing a portion of a map of human chromosome 16. The display includes several different types of map elements: chromosome bands (thick bars at the top of the canvas), chromosome fragments from hybrid-cell lines (thin blue lines), anonymous DNA markers from the Genome Data Base (red bars), cosmid clones from the Los Alamos mapping effort (orange bars), YACs (blue bars), genes (green bars), and fragile sites (black bars). (The clones and fragile sites are not drawn to scale in this view because they would be too small to see.)

designed explicitly for genome maps, users have a wide choice of styles in which maps can be displayed, depending on the particular question of interest.

One problem in integrating genome maps is that conversions between the various units employed vary from one region of the chromosome to another and are even non-linear. In SIGMA, the different scales are integrated by dividing the map into regions of arbitrary size in which users can specify linear conversions between various units. For instance, in one part of the chromosome a centimorgan (the unit of genetic distance) may be set equal to a million base pairs, while in another part a centimorgan may correspond to half a million base pairs. Users can freely change the units in which the map is displayed. In the figure above the chosen linear scale is spatial distance along a metaphase chromosome as observed under a microscope. Therefore SIGMA shows element lengths and inter-element distances given in base pairs, say, according to the conversion between base pairs and spatial distance assigned for the part of the chromosome in which the elements lie.

SIGMA handles the problem of fragmentary data by treating the map-assembly process as an optimization problem. In optimization theory, one is presented with a number of (possibly inconsistent) statements that should be true about a solution to a particular problem. These statements, perhaps in conjunction with estimates of their certainty, are called "objectives". The goal is the generation of one or more solutions that satisfy the objectives as well as possible.

For genome maps, an objective is typically either a statement about a single element in the map (such as, "This YAC is about 400,000 base pairs long"), or a statement about the positional relationship between two elements (such as, "These two clones probably overlap by about 10,000 base pairs"). Even a map of only modest complexity can be based on literally millions of such objectives, far more than a human can sensibly handle. SIGMA, on the other hand, easily tracks this quantity of information and can help users find maps that meet the objectives as closely as possible. The figure opposite shows the user's view of how SIGMA manages objectives.



To demonstrate how SIGMA handles map objectives, one element, clone S33, has been selected in the map canvas; consequently its properties appear in the Element Properties Window (left). That window displays, in addition to the type, name, and description of the element, the graphical coordinates of the element in the canvas and some of the objectives involving the element. The first two objectives shown give the minimum and maximum lengths of clone S33 consistent with experiment. The objectives that follow state relationships inferred from experiments in which clone S33 was hybridized with a panel of hybrid-cell lines, each containing only a portion of chromosome 16. For each hybrid-cell line that the clone hybridized with, an objective has been created indicating that the clone lies within that chromosome fragment. For each hybrid-cell line that the clone did not hybridize with, an objective has been created indicating that the clone and that chromosome fragment do not overlap. All those objectives have been assigned a 0.98 probability of being correct, based on the uncertainty of the experiments. Finally, the last two distances in the window are the maximum and minimum values of the distance between the left endpoint of the clone and the left endpoint of the highlighted hybrid-cell line. (If the two elements overlapped, the length of the overlap would be given; if they did not touch, the distance between them would be given.)

SIGMA includes special optimization routines to automate map assembly. (The routines currently use only objectives concerning clone lengths, clone-overlap probabilities, and lengths of overlaps, which are the data used in constructing contig maps.) The optimization is performed by algorithms inspired by natural genetics, called "genetic algorithms". (See the discussion of genetic algorithms in the main text.) Whether a map was made by the optimization routines or by hand, SIGMA can automatically evaluate how well it fits the objectives. Thus the user can edit the map interactively, seeing how each change affects the map's agreement with the data.

As the map grows and new data become available, the collection of map objectives grows. Old objectives are never discarded unless a user explicitly deletes them. Because the objectives can be passed along to other users as part of a map, subsequent users of the map have access to all the information on which it is based, allowing them to make their own judgements about the correctness of the conclusions. This ability is very important when one laboratory's data appear to conflict with prior results from another group. Instead of being limited to the final product of the earlier work, the second team can look "inside" the map, examining the assumptions on which the map is based to find the specific causes of discrepancies.

Finally, SIGMA was designed from the beginning to be used with Electronic Data Publishing (see the sidebar "Electronic Data Publishing in GenBank" immediately following). Not only can users easily share data with other SIGMA users, but they can prepare submissions to the public mapping databases with just a few keystrokes. ■