Good Maps are Straight*

Eric Harley
University of Toronto
Dept. of Computer Science
Toronto, Ont, Canada M5S 3G4
eharley@db.toronto.edu

Anthony J. Bonner
University of Toronto
Dept. of Computer Science
Toronto, Ont, Canada M5S 3G4
bonner@db.toronto.edu

Nathan Goodman
Whitehead Institute/MIT
Center for Genome Research
Cambridge, MA 02139, USA
nat@genome.wi.mit.edu

Abstract

This paper proposes a simplified approach to the assembly of large physical genome maps. The approach focuses on two key problems: (*i*) the integration of diverse forms of data from numerous sources, and (*ii*) the detection and removal of errors and anomalies in the data. The approach simplifies map assembly by dividing it into three phases—*overlap*, *linkage* and *ordering*. In the first phase, all forms of overlap data are integrated into a simple abstract structure, called *clusters*, where each cluster is a set of mutually-overlapping DNA segments. This phase filters out many questionable overlaps in the mapping data. In the second phase, clusters are linked together into a *weighted intersection graph*. False links between widely separated regions of the genome show up as crooked, branching structures in the graph. Removing these false links produces graphs that are straight, reflecting the linear structure of chromosomes. From these straight graphs, the third phase constructs a physical map. Graph algorithms and graph visualization play key roles in implementing the approach. At present, the approach is at an early stage of development: it has been tested on real and simulated mapping data, and the results look promising. This paper describes the first two phases of the approach in detail, and reports on our progress to date.

Introduction

A major goal of the Human Genome Project is to construct detailed physical maps of the human genome and the genomes of other organisms. A physical map gives the position on a genome of numerous small fragments of DNA, called *clones*. Most genomic experiments do not determine clone positions directly. Instead, they determine spatial relationships between them, such as whether two clones overlap. Inferring clone position from the experimental data is called *map assembly*. Assembling complete and detailed maps of large genomes requires the integration of many forms of experimental data from many sources. The resulting maps are said to *integrate* the data.

Algorithms for assembling integrated physical maps face several challenges. First, they must handle the increasingly large volumes of mapping data now being generated. Second, they must be flexible enough to accommodate new forms of data as they become available. Third, they must deal effectively with the complexities of the data, *e.g.*, with anomalies, inconsistencies and imprecision, and with subtle relationships between the different forms of data. Unfortunately, this last point complicates the map-assembly process enormously, limiting the effectiveness and flexibility of many map-assembly algorithms. To address this problem, we propose to simplify map assembly by dividing it into three phases—*overlap*, *linkage* and *ordering*.

The first phase integrates all forms of overlap data into a simple abstract structure, called *clusters*. A cluster is any set of mutually overlapping clones. Because they mutually overlap, the clones in a cluster must have at least one point in common; so the cluster can be thought of as a point on the genome. In addition, clustering filters out many false overlaps, because each overlap in a cluster is corroborated by many others. In this paper, all clusters have at least three clones. The section on overlap describes an algorithm for generating clusters.

After clusters are generated, the second phase links them together into a *weighted intersection graph*. Intersection graphs are a well-studied class of graphs in which each node is a set, and an edge means that two sets intersect. In our case, each node is a cluster, and an edge of weight M means that two clusters have M clones in common. Intuitively, an edge is evidence that two clusters are close together on the genome, and the weight is the strength of that evidence. In this paper, we only consider edges of weight at least 2. This filters out many false links between widely separated regions of the genome (caused by an experimental anomaly

*Published in *Proceedings, Fourth International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pages 88-97. AAAI Press. Conference held June 12–15, 1996, Washington University, St. Louis, Missouri. Available on-line at ftp://db.toronto.edu/pub/bonner/papers/genome.mapping/