Revealing Hidden Interval Graph Structure in STS-Content Data

Eric Harley[1], Anthony Bonner[1] and Nathan Goodman[2]

Abstract

Motivation: Contig mapping is an important and difficult problem in genomic technology. The data used to construct such maps contain numerous errors. No fully automated means to analyze such data yet exists.
Results: This paper introduces a graph algorithm which creates a simplified view of physical mapping data. The shape of the resulting *structure* graph provides a quality check — coherent data produce a straight line, while anomalous data produce branches and loops. Besides being a useful rule of thumb, the method can help formalize steps taken in physical mapping. In cases where the data contain false positives creating a complex structure graph, it is sometimes possible to disentangle the various paths into subsets of the data covering contiguous regions of the genome, i.e., *contigs*. These straight subgraphs can then be analyzed in standard ways to construct a physical map. A theoretical basis for the method is presented along with examples of its application to current STS data from human genome centers. The technique has general applicability to any data that have an underlying (albeit noisy) interval model.
Availability: Not applicable.
Contact: Eric Harley `eharley@cs.toronto.edu`

1 Introduction

Two epic efforts to construct contig maps of the human genome were begun in the early 1990's, one led by Daniel Cohen at the Centre d'Etudes du Polymorphisme Humain and Généthon (CEPH/Généthon), and the second led by Eric Lander at the Whitehead Institute for Biomedical Research/MIT Center for Genome Research (WI/MIT). Both groups adopted a general strategy of combining *bottom-up* methods for constructing contigs across modest regions, with *top-down* methods for ordering the contigs across entire chromosomes. Both groups used STS-content mapping as a key bottom-up method, though the CEPH/Généthon team (which started first) initially used fingerprint mapping for this purpose. For top-down ordering of contigs,

[1]Department of Computer Science, University of Toronto, Toronto, Ont, Canada M5S 1A4
[2]The Jackson Laboratory, Bar Harbor ME 04609-1500