Abstract

Graph Algorithms for Assembling Integrated Genome Maps

Eric Richard Harley

Doctor of Philosophy

Graduate Department of Computer Science

University of Toronto

2003

This dissertation is a computational study in the field of bioinformatics, a relatively new discipline at the intersection of computer science and biological science. As such, the dissertation contributes to each of these disciplines. The problems we address within biological science are in the area of physical mapping, i.e., mapping experimentally-produced fragments of a genome to their original locations in the genome. The thesis makes two primary contributions in this area: (*i*) a method to improve contig identification and verification, and (*ii*) a method which improves contig formation by uniformly integrating probe and overlap data. The problems addressed within computer science arise in the above application and are in the area of heuristic graph algorithms. The thesis makes three primary contributions in this area: (*i*) a modification of a well-known algorithm for listing maximal cliques, which markedly improves its efficiency on large, sparse graphs; (*ii*) a new clustering algorithm which is appropriate for "noisy" interval graphs; and (*iii*) a method which reveals piecewise linearity in a noisy interval graph.

The methods and algorithms are tested on human genome mapping data produced by the Human Genome Project and also on simulated data. The results from applying our structure graph algorithm to Sequenced Tag Site (STS) mapping data reveal a network-like pattern where linear segments often correlate with data from continuous pieces (contigs) of the genome. The underlying interval graph nature of the probe mapping data is thus revealed, thereby facilitating identification and extraction of contigs.

Results from the application of our “virtual probe” technique for integrating probe and overlap data indicate that this method can produce longer double-linkage contigs than STS probes alone, and that it provides a new way of assessing overlap data. Results of experiments on enumeration of cliques in large, sparse graphs representing genome mapping data show that our modification of the Bron-Kerbosch algorithm requires time which is approximately linear in the number of vertices, whereas other tested algorithms had worse than linear time complexity. Results from tests of the clustering algorithm on simulated noisy overlap data indicate that it effectively counters the increase in clique numbers caused by false negatives.