# Algorithms for the Extraction of Synteny Blocks from Comparative Maps

Vicky Choi[1], Chunfang Zheng[2], Qian Zhu[2], and David Sankoff[2]

[1] Department of Computer Science, Virginia Tech, Blacksburg, VA 24061
vchoi@cs.vt.edu
[2] Departments of Biology, Biochemistry, and Mathematics and Statistics, University of Ottawa, Ottawa, Canada K1N 6N5
{czhen033,qzhu012,sankoff}@uottawa.ca

**Abstract.** In comparing genomic maps, we try to distinguish mapping errors and incorrectly resolved paralogies from genuine rearrangements of the genomes. This can be formulated as a Maximum Weight Independent Set (MWIS) search, where vertices are potential strips of markers syntenic on both genomes, and edges join conflicting strips, in order to extract the subset of compatible strips that accounts for the largest proportion of the data. This technique is computationally hard. We introduce biologically meaningful constraints on the strips, reducing the number of vertices for the MWIS analysis and provoking a decomposition of the graph into more tractable components. New improvements to existing MWIS algorithms greatly improve running time, especially when the strip conflicts define an interval graph structure. A validation of solutions through genome rearrangement analysis enables us to identify the most realistic solution. We apply this to the comparison of the rice and sorghum genomes.

## 1 Introduction

Comparing two genomic maps containing orthologous sets of markers induces a decomposition of the genomes into synteny blocks, segments of chromosomes containing orthologous markers in the same or reverse order in the two genomes. The blocks may be differently grouped into chromosomes, and differently ordered and oriented, in the two genomes being compared.

In the course of genomic evolution, as more and more rearrangements intervene since the common ancestor, the synteny blocks in common between the two genomes become more fragmented, i.e., shorter, and eventually contain only one marker, or none.

The construction of the synteny blocks based on traditional comparative maps is different in both spirit and technique from the analogous problem based on genome sequences, and is very vulnerable to errors and ambiguities in the position of the markers on a map, depending on the specific mapping technology. Another kind of problem involves ambiguous homology, leading to the risk of matching up inappropriate pairs of markers as orthologs in the two genomes. These problems tend to artifactually increase the number of synteny blocks induced by the comparison, disrupting true synteny blocks by artifactual blocks containing only one or two markers.

Thus, when many rearrangements have intervened since the common ancestor, or where the sampling density of markers on the chromosome is sparse, it may be unclear

R. Giancarlo and S. Hannenhalli (Eds.): WABI 2007, LNBI 4645, pp. 277–288, 2007.

whether any particular one of the increasing number of short synteny blocks is due to error or to rearrangement. These considerations suggest the principle that inferences that depend on the position of a single marker should not be given as much weight as inferences that are supported by more markers. We would thus like to construct a set of synteny blocks that are conflict-free, contain as much of the data as possible, and are credible from a genome rearrangement viewpoint.

In [9], we proposed the following strategy: first, construct a set of *pre-strips*, which are certain short common subsequences of one chromosome from each genome; second, extract from this set a subset of mutually compatible (non-intersecting) containing a maximum number of markers; third, add to this subset any markers that do not increase the rearrangement distance [7] between the genomes; fourth, assemble the synteny blocks from the markers in the solution.

This approach encountered a bottleneck at the second step, formulated in terms of a solution for the NP-hard maximum weight clique (MWC) problem in a graph representing pre-strip compatibilities. It was not feasible to run the whole data set using available algorithms. Thus we devised biologically-motivated constraints to reduce the data set and were then able to run moderate size instances.

In this paper, our main contributions are: first, based on a key combinatorial observation, the establishment of constraints on the set of pre-strips that are necessary to a solution, thus reducing the amount of data that must be input to MWIS without losing optimality (Section 3), and second, the design of a new algorithm for the maximum weight independent set (MWIS) problem[1], specifically motivated by the nature of pre-strip data (Section 4.1). Finally, taking advantage of the source of the incompatibilities in the chromosome-based data, we propose a natural decomposition of the graph which allows us to solve relatively large instances of the problem extremely efficiently – 1 to 2 seconds on a Pentium IV computer for instances that took days or that proved infeasible with the previous techniques. As a prerequisite to this material, in Section 2, we review the definition of strips and pre-strips, as well as a polynomial-time algorithm for generating all pre-strips. And after the theoretical development, we discuss the question of restoring additional markers to the solution in Section 6 and analyze the rice and sorghum comparative map in Section 7.

## 2 Problem and Terminology: Strips, Pre-strips, Pure Strips

Let $n$ be the number of markers in common in two genomes with $\chi_1$ and $\chi_2$ chromosomes. In one genome, number all these markers on any one of the chromosomes from left to right in increasing order starting with marker 1. Continue the numbering sequence on a second chromosome and so on, until finishing with the $n$-th marker on the $\chi_1$-st chromosome. Then each marker in the second genome receives the same label as its supposed ortholog in the first genome.

We recall the definition of *strips, pre-strips* and *pure strips* in [9]. Consider any $l \geq 2$ consecutive contiguous markers on a chromosome in one genome. If the same $l$ markers are consecutive on a chromosome in the other genome, with the same (or

[1] Equivalent to the MWC formulation in the complementary graph in [9].