Parts of the Problem of Polyploids in Rearrangement Phylogeny

Chunfang Zheng, Qian Zhu, and David Sankoff

Departments of Biology, Biochemistry, and Mathematics and Statistics,
University of Ottawa, Ottawa, Canada K1N 6N5
{czhen033,qzhu012,sankoff}@uottawa.ca

Abstract. Genome doubling simultaneously doubles all genetic markers. Genome rearrangement phylogenetics requires that all genomes analyzed have the same set of orthologs, so that it is not possible to include doubled and unduplicated genomes in the same phylogeny. A framework for solving this difficulty requires separating out various possible local configurations of doubled and unduplicated genomes in a given phylogeny, each of which requires a different strategy for integrating genomic distance, halving and rearrangement median algorithms. In this paper we focus on the two cases where doubling precedes a speciation event and where it occurs independently in both lineages initiated by a speciation event. We apply these to a new data set containing markers that are ancient duplicates in two yeast genomes.

1 Introduction

Basic rearrangement phylogeny methods require that the genomic content be the same in all the organisms being compared, so that every marker (whether gene, anchor, probe binding site or chromosomal segment) in one genome be identified with a single orthologous counterpart in each of the others, though adjustments can be made for a limited amount of marker deletion, insertion and duplication.

Many genomes have been shown to result from an ancestral doubling of the genome, so that every chromosome, and hence every marker, in the entire genome is duplicated simultaneously. Subsequently, the doubled genome evolves through mutation at the DNA sequence level and by chromosomal rearrangement, through intra- and interchromosomal movement of genetic material. This movement can scramble the order of markers, so that the chromosomal neighbourhood of a marker need bear no resemblance to that of its duplicate.

The present-day genome, which we refer to here as a doubling descendant, can be decomposed into a set of duplicate or near-duplicate markers dispersed among the chromosomes. There is no direct way of partitioning the markers into two sets according to which ones were together in the same half of the original doubled genome. Genomic distance or rearrangement phylogeny algorithms are not applicable to doubling descendants, since there is a two-to-one relationship between markers in the doubling descendant and related species whose divergence predates the doubling event, whereas these algorithms require a one-to-one correspondence.

G. Tesler and D. Durand (Eds.): RECOMB-CG 2007, LNBI 4751, pp. 162–176, 2007.

We have undertaken a program [11,9] of studying rearrangement phylogeny where doubling descendants are considered along with related unduplicated genomes. We believe there is no other computationally-oriented literature on this particular problem. To focus on the problem of marker ambiguity in doubling descendants, and to disentangle it from the difficulties of constructing phylogenies, we pose our computational problems only within the framework of the "small" phylogenetic problem, i.e., identifying the ancestral genomes for a given phylogeny that jointly minimize the sum of the rearrangement distances along its branches.

In Section 2, we outline a model for generating an arbitrary pattern of doubled descendants observed at the tips of a given phylogeny. Based on this model, we then present an simple algorithm for inferring the doubling status of the ancestral genomes in terms of an economical set of doubling events along the branches of the phylogeny. Once we have the ancestral doubling statuses, we can approach the actual rearrangement problem.

First, in Section 3, we identify three kinds of component of this problem for which algorithms already exist, one a calculation of the genomic distance between two given genomes with clearly identified orthologs, i.e., the minimum number of rearrangements necessary to transform one genome into another; the second a "halving" algorithm for inferring the genome of a doubled genome based on internal evidence from its modern descendant only, and the third a "medianizing" process for inferring an ancestral genome from its three neighbouring genomes in a binary branching tree.

In Section 4, we discuss our recent papers [11,9] on incorporating algorithms for the three components into an overall procedure for inferring ancestral genomes in the case of one doubling descendant and two related unduplicated genomes. The contribution of the present paper starts in Section 5 where we analyze two ways of relating genomes from two doubling descendants, one where they result from a single genome doubling event followed by a speciation, and the other where speciation precedes two genome doublings, one in each lineage. In Section 7, we apply these two methods to a large data set on yeast.

1.1 Terminology and Scope

In biology, the concept of genome doubling is usually expressed as tetraploidization or autotetraploidization, and the both the doubled genome and its doubling descendant are called tetraploid, even though, generally, the descendants soon undergo a process called (re-)diploidization and function as normal diploids, still carrying a full complement of duplicate markers that evolve independently of each other. Though unambiguous in biological context, implicit in this terminology are many assumptions that are not pertinent to our study. In the yeast data we study here, for example, *Saccharomyces cerevisiae* exists during most of its life cycle as a haploid, only sometimes as a diploid, while *Candida glabrata* exists uniquely as a haploid.