A comparative phylogenetic approach for dating whole genome duplication events

Brad A. Chapman1,2,* John E. Bowers2, Stefan R. Schulze2 and Andrew H. Paterson2

1Department of Plant Biology, University of Georgia, Athens, GA 30602, USA and
2Plant Genome Mapping Laboratory, University of Georgia, Athens, GA 30602, USA

Received on November 9, 2003; accepted on November 20, 2003

ABSTRACT

Motivation: Whole genome duplications have played a major role in determining the structure of eukaryotic genomes. Current evidence revealing large blocks of duplicated chromatin yields new insights into the evolutionary history of species, but also presents a major challenge for researchers attempting to utilize comparative genomics techniques. Understanding the timing of duplication events relative to divergence among taxa is critical to accurate and comprehensive cross-species comparisons.

Results: We describe a large-scale approach to estimate the timing of duplication events in a phylogenetic context. The methodology has been previously utilized for analysis of Arabidopsis and Saccharomyces duplication events. This new implementation provides a more flexible and reusable framework for these analyses. Scripts written in the Python programming language drive a number of freely available bioinformatics programs, creating a no-cost tool for researchers. The usefulness of the approach is demonstrated through genome-scale analysis of Arabidopsis and Oryza (rice) duplications.

Availability: Software and documentation are freely available from http://plantgenome.agtec.uga.edu/bioinformatics/dating/

Contact: chapmanb@uga.edu

INTRODUCTION

The completed sequences of several eukaryotic organisms have dramatically increased the scope of molecular evolutionary studies. One particularly important early finding is that whole genome duplication has played a much larger role in evolution than had been previously anticipated. Whole genome duplications have been detected in organisms ranging from yeast (Wolfe and Shields, 1997) to human (McLysaght et al., 2002) to Arabidopsis (Arabidopsis Genome Initiative, 2000; Blanc et al., 2000; Paterson et al., 2000; Vision et al., 2000). The prevalence of these duplications in a broad range of eukaryotic organisms demonstrates the importance of whole genome duplications in the structuring and evolution of genome structure.

Understanding the timing of these duplications provides essential context for genome-level comparisons between organisms. Similarity between species in genome organization can be obscured by duplications and subsequent rearrangements—taking into account duplication events that have occurred after the divergence of species from one another increases our ability to detect ancestral similarities in gene arrangement (often called synteny or collinearity) (Bowers et al., 2003; Zhu et al., 2003).

Traditional methods of dating duplications have relied on synonymous (KS) and non-synonymous (KA) substitution rates. This technique has been used for estimating the timing of duplication events in yeast (Gu et al., 2002), human (Makova and Li, 2003), Arabidopsis (Vision et al., 2000; Zhang et al., 2002; Ziolkowski et al., 2003; Blanc et al., 2003) and others (Lynch and Conery, 2000). These values have the advantage of being rapid and easy to generate, but suffer from the assumption of a constant molecular clock and thus consistent rates of KS. Recently, KS values have been found to vary much more widely among individual genes than had been previously thought (Zhang et al., 2002), and thus may represent a very rough indicator of the timing of duplication if based on data from small groups of genes (Bowers et al., 2003).

An alternative dating technique has been recently described for analysis of duplications in yeast (Langkjaer et al., 2003) and Arabidopsis (Bowers et al., 2003). This technique involves creating and analyzing large numbers of phylogenetic trees, comprising of a pair of duplicated genes from one taxon, the best homolog from a second taxon, and the best homolog from an outgroup. The dating of duplications is thereby performed relative to the time of divergence between organisms.

This paper describes a flexible implementation of this phylogenetic dating procedure. Written in the Python programming language and driving freely available bioinformatics tools, this set of programs automates the steps from database generation to analysis of dating results. Results from its application to dating of duplications in the Arabidopsis and Oryza genomes are described.

*To whom correspondence should be addressed.
METHODS
Phylogenetic dating involves a three-step process which culminates in rooted phylogenetic trees (Fig. 1). The starting material for the analysis is a file of duplicate protein pairs for a genome in which one wants to investigate the timing of duplication. Such a dataset is obtained through detailed analysis of patterns of gene duplication in the genome of interest. Software and algorithms have been described elsewhere for detecting duplicate pairs (Conant and Wagner, 2002; Calabrese et al., 2003) and a number of ad hoc methods have also been used (Bowers et al., 2003), so delineation of duplicate pairs is not a goal of this software package and will not be discussed here.

Organism database creation
A preliminary step in running the dating analysis involves establishing sequence databases for a set of organisms. Our approach requires only partial sequence data and makes use of the taxonomy structure databases at the National Center for Biotechnology Information, so that organism-specific databases can be retrieved for any level of taxonomical organization.

Duplication events are evaluated relative to the divergence from a common ancestor of the organism in which we are analyzing duplication patterns (duplicate organism) and the organism represented in the database (comparison organism). Therefore, comparison databases should be chosen with varying evolutionary distances from the duplicate organism.

The organism databases can be subjected to clustering or other treatments to remove redundant or low-quality sequences.

Detection of homologs
With organism-specific databases in hand, the next step involves selection of a best detectable homolog for each duplicate protein pair. This homolog is defined as the BLAST (Altschul et al., 1997) (tblastn) hit with the longest match region above a configurable significance (E-value) threshold (the default is an E-value of $10^{-5}$).

In this way, we attempt to obtain the longest detectable region of sequence similarity. At this stage there are no restrictions on the lengths of alignments accepted as the best hit—length restrictions are applied during preparations for phylogenetic analysis. To try and obtain a useful sequence beyond the boundaries detected by BLAST, additional translated sequence to either side of the original hit location is included if available. Subsequent local alignment comparisons will remove non-informative added sequence.

Generation of phylogenetic trees
The starting materials for phylogeny construction are: (a) the two protein sequences from a duplicate gene pair; (b) the best homolog from a comparison organism representing a particular taxonomic node; and (c) the best homolog from an outgroup organism known to be a very distant relative and thus used as the root in the final phylogeny. These four protein sequences are subjected to Smith–Waterman pairwise alignments [‘water’ in EMBOSS (Rice et al., 2000)] to find the detectable regions conserved between all pairs.

With the regions of sequence similarity overlap for all four proteins, multiple alignments are performed using T-Coffee (Notredame et al., 2000). Other multiple sequence alignment programs such as ClustalW (Chenna et al., 2003) are also supported. Alignment quality is assessed using T-Coffee’s ability to produce reliability scores.

Multiple alignments are used as input into phylogenetic analyses to produce rooted trees comparing the duplicate pairs and best homolog. The PHYLIP set of programs (Felsenstein, 2003, http://evolution.genetics.washington.edu/phylip.html) is used for bootstrapped maximum likelihood
Bayesian approaches as implemented by MrBayes (Ronquist and Huelsenbeck, 2003) are also supported. The results of phylogenetic analysis lead to only two possible rooted tree topologies. In one topology (Fig. 1, external tree), the members of the duplicated pair are more similar to one another than either is to the best homolog. In this case, the best homolog is referred to as external to the duplicate pair. In the alternative topology (Fig. 1, internal tree), the homolog is more similar to one member of the duplicate pair then is the other member of the duplicate pair. For this result, the best homolog is considered internal to the duplicate pair.

**ALGORITHM**

**Interpretation of individual trees**

The analyses described above produce the best possible tree of two duplicates and a homolog, rooted using an outgroup sequence. Interpretation of this tree provides the mechanism for dating the duplication event being examined. The goal of this analysis is to assign a relative order to two different events:

- The genome duplication being examined in our organism of interest.
- The divergence time, from a common ancestor, of two organisms—the study organism and the comparison organism.

Making the assumption that the length of evolutionary time since separation of genes from a common ancestral gene (whether by speciation or duplication) is reflected by their degree of phylogenetic similarity, we can thus evaluate which of the two scenarios pictured in Figure 2 correlates with our two possible phylogenetic tree topologies (Fig. 1). If speciation occurred prior to duplication, then we expect a tree with the duplicates that are closest (external) more often. In contrast, if the duplication event preceded speciation, then we expect the organism homolog to be closer to one of the duplicate pairs (internal) more often.

**Interpretation of collected results**

Large numbers of duplicate pairs are evaluated to minimize the impact of evolutionary events differing from our assumptions (tandem duplications and deletions, gene conversion). Practically, we also expect some failures to identify true homologs as a consequence of using sequence similarity searches with partial sequence databases (expressed sequence tag, genomic fragment) of necessarily limited size. The requirement of a divergent organism as an outgroup root results in highly conserved regions being most commonly represented in the analyses.

External trees are prone to include a disproportionately large share of these artifactual results, hence inferences rely heavily upon comparative ratios of internal trees to total trees evaluated. These ratios are examined across all comparisons to look for organisms which provide different amounts of evidence for one particular evolutionary history of duplication and speciation (Fig. 2). Methods such as correlated ANOVA analysis between duplication blocks (Bowers et al., 2003) can be applied to test the significance of differences in frequencies of internal trees between organisms.

**IMPLEMENTATION**

The analyses described above were implemented as a set of scripts and modules in the Python programming language. This code drives the bioinformatics tools listed in Figure 1 and is modularized for use in clustered and single-computer environments. A flexible Python-based configuration file allows the code to be utilized for a wide variety of duplicate organisms and comparison databases.

**RESULTS**

The implementation described above was used for dating of duplications in the Arabidopsis and Oryza (rice) genomes. Duplications have played an important role in shaping current plant genomes (Schmidt, 2002; Mitchell-Olds and Clauss, 2002). Understanding duplication events will have important consequences for detecting long-range synteny and applying information from sequenced plant genomes to important crop species with limited sequence data.

**Arabidopsis**

Since the discovery of large blocks of duplicated genes in the Arabidopsis genome (Arabidopsis Genome Initiative, 2000;
Comparative phylogenetic approach for dating whole genome duplications

Table 1. *Arabidopsis* duplication dating results for three duplication events: $\alpha$, $\beta$ and $\gamma$, in order of increasing age (Bowers et al., 2003)

<table>
<thead>
<tr>
<th>Event</th>
<th>Pinaceae</th>
<th>Oryza</th>
<th>Solanaceae</th>
<th>Medicago</th>
<th>Malvaceae</th>
<th>Citrus</th>
<th>Brassica</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TaxId:3318</td>
<td>TaxId:4527</td>
<td>TaxId:4070</td>
<td>TaxId:3877</td>
<td>TaxId:3629</td>
<td>TaxId:2706</td>
<td>TaxId:3705</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.034 (669)</td>
<td>0.055 (965)</td>
<td>0.070 (719)</td>
<td>0.090 (636)</td>
<td>0.070 (567)</td>
<td>0.073 (449)</td>
<td>0.485 (526)</td>
</tr>
<tr>
<td>$\alpha$ short</td>
<td>0.037 (54)</td>
<td>0.029 (68)</td>
<td>0.058 (52)</td>
<td>0.064 (47)</td>
<td>0.049 (41)</td>
<td>0.059 (34)</td>
<td>0.488 (41)</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.091 (243)</td>
<td>0.214 (290)</td>
<td>0.313 (211)</td>
<td>0.328 (192)</td>
<td>0.328 (177)</td>
<td>0.292 (154)</td>
<td>0.688 (218)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.324 (71)</td>
<td>0.532 (94)</td>
<td>0.655 (87)</td>
<td>0.675 (80)</td>
<td>0.545 (66)</td>
<td>0.596 (57)</td>
<td>0.808 (78)</td>
</tr>
</tbody>
</table>

Ratios indicate the number of internal trees to total trees examined; numbers in parentheses are the total number of trees analyzed for each event and organism. Grayed numbers indicate inferred organisms post-dating the duplication event based on eyeball comparisons of internal tree ratios.

Our laboratory has previously performed dating of *Arabidopsis* duplication events using a phylogenetic approach (Bowers et al., 2003) similar to the one described in this paper. The results presented here (Table 1) use the new software tool described, incorporating more rigorous requirements for tree generation and interpretation. The analysis was performed using multiple sequence alignments of at least 35 amino acids. 100 bootstrap replicates were examined using protein parsimony, with PHYLIP’s ‘protpars’ default implementation for scoring amino acid changes. Only trees with at least 80 (out of 100) bootstrap confidence support in the informative branch were used in the analysis.

The results of this dating analysis coincide with previous results (Bowers et al., 2003), with a few notable exceptions. First, *Citrus* is included in the analysis, which provides another dating point along with the Malvaceae family to specify the most recent ($\alpha$) duplication event. The $\alpha$ event appears to have occurred after the divergence of these taxa, but before the divergence of Brassica and *Arabidopsis*. Dating results from short duplicate blocks generated by the $\alpha$ event also confirm these results.

Second, this analysis places the subsequent $\beta$ event at a more ancient point than previously predicted (Bowers et al., 2003). Specifically, we find a relatively larger number of internal trees in the *Oryza* analysis, indicating that the $\beta$ event may have occurred near or prior to the divergence of the monocots from the dicots (represented by *Arabidopsis*). In addition to the difference in rigor of analysis, these results also include nearly twice as much rice sequence data as our previous work (Bowers et al., 2003) due to the current emphasis on genome sequencing of *Oryza*. This highlights the importance of considering database size in interpreting dating results, and also the tendency toward external trees in incomplete datasets.

### Rice

Having gained some insights into the evolutionary history of dicotyledonous plants (dicots) from the analysis of *Arabidopsis*, a logical next step is to explore the history of the other major branch of the plant family tree, the monocots, as exemplified by the emerging sequence of the *Oryza* (rice) genome. Two pictures of rice duplications have already emerged. One analysis describes rice duplications as being generated from aneuploidy events, involving only a subset of the rice chromosomes (Vandepoele et al., 2003). Our own early analysis suggests a duplication event involving most of the genome (Paterson et al., 2003).

We performed an analysis of the rice duplication data (Paterson et al., 2003) using the described phylogenetic dating implementation. Predicted rice proteins were ordered using bacterial artificial chromosomes (BACs) physically mapped along rice chromosomes (Paterson et al., 2003). Intra-genome BLAST comparisons were then conducted and

Table 2. Rice duplication dating results

<table>
<thead>
<tr>
<th>Block</th>
<th>Pinaceae</th>
<th>Sorghum</th>
<th>Hordeum</th>
<th><em>O.minuta</em></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TaxId:3318</td>
<td>TaxId:4557</td>
<td>TaxId:4512</td>
<td>TaxId:63629</td>
</tr>
<tr>
<td>1</td>
<td>0.025 (121)</td>
<td>0.318 (88)</td>
<td>0.309 (97)</td>
<td>0.234 (47)</td>
</tr>
<tr>
<td>2</td>
<td>0.015 (65)</td>
<td>0.286 (42)</td>
<td>0.419 (31)</td>
<td>0.414 (29)</td>
</tr>
<tr>
<td>3</td>
<td>0.000 (61)</td>
<td>0.381 (42)</td>
<td>0.333 (54)</td>
<td>0.429 (21)</td>
</tr>
<tr>
<td>4</td>
<td>0.038 (53)</td>
<td>0.333 (33)</td>
<td>0.300 (40)</td>
<td>0.471 (17)</td>
</tr>
<tr>
<td>5</td>
<td>0.059 (34)</td>
<td>0.375 (24)</td>
<td>0.267 (30)</td>
<td>0.714 (14)</td>
</tr>
<tr>
<td>8</td>
<td>0.091 (33)</td>
<td>0.433 (30)</td>
<td>0.481 (27)</td>
<td>0.333 (9)</td>
</tr>
<tr>
<td>Total</td>
<td>0.033 (427)</td>
<td>0.314 (309)</td>
<td>0.302 (338)</td>
<td>0.368 (155)</td>
</tr>
</tbody>
</table>

Values, shading and taxonomy ids are as explained in Table 1. Results are shown for all duplicated blocks with 10 or more comparisons done, and for the overall event. Ten duplicated blocks were used in the analysis. The ratios indicate that the examined duplication event occurred after the divergence of Pinaceae from the rice lineage but prior to the divergence of *Sorghum* and *Hordeum*.

Blanc et al., 2000; Vision et al., 2000; Paterson et al., 2000; Kowalski et al., 1994, much work has focused on understanding the nature of these duplicated blocks. The genome is made up of duplicated and rearranged blocks which are thought to have been generated by a whole genome duplication event that occurred relatively recently in evolutionary history, before the divergence of the legumes and the mallows from *Arabidopsis* (Ermolaeva et al., 2003; Bowers et al., 2003). There is also evidence of older duplication events. Different authors have suggested that from one to three or more additional partial duplication events may have occurred (Ziolkowski et al., 2003; Blanc et al., 2003; Bowers et al., 2003; Raes et al., 2003; Simillion et al., 2002; Vision et al., 2000).

Having gained some insights into the evolutionary history of dicotyledonous plants (dicots) from the analysis of *Arabidopsis*, a logical next step is to explore the history of the other major branch of the plant family tree, the monocots, as exemplified by the emerging sequence of the *Oryza* (rice) genome. Two pictures of rice duplications have already emerged. One analysis describes rice duplications as being generated from aneuploidy events, involving only a subset of the rice chromosomes (Vandepoele et al., 2003). Our own early analysis suggests a duplication event involving most of the genome (Paterson et al., 2003).
visualized to select detectable regions of duplication (blocks) (Bowers et al., 2003; Paterson et al., 2003).

Conditions to prepare duplicates and homologs for phylogenetic analysis were as enumerated earlier for Arabidopsis. For phylogenetic analysis, 25 bootstrap replicates were examined with maximum likelihood, using the Jones–Taylor–Thornton amino acid substitution probabilities. Rate variability between sites was allowed, with a gamma model using one invariant site and two other sites with an \( \alpha \) value of 0.9. Only trees with at least 20 (out of 25) bootstrap confidence support were used in the analysis.

These results lead to two interesting insights (Table 2). First, the rice duplication event appears to be more ancient than its divergence from both Sorghum and Hordeum (barley). Because these represent very divergent members of the Poaceae (cereals), it appears likely that this duplication event predated the divergence of most or all of the Poaceae from a common ancestor.

Second, the results of a dataset for Oryza minuta, closely related to the organism in which we are analyzing duplicates (Oryza sativa), indicate the complications associated with the use of smaller databases. The presence of only 5286 available sequences leads to a subsequently larger amount of variability between blocks. Although the summed results agree with dating conclusions obtained from more distantly related species, we see two major deviations from our expectations—a low 0.234 ratio for block 1 and a high 0.714 ratio for block 5. These numbers are likely due to a combination of smaller numbers of comparisons done in each block and increased failure to select true homologs due to the limited representation of the database.

### Table 3. In-depth examination of the timing of the Arabidopsis \( \beta \) duplication event relative to monocot divergence times

<table>
<thead>
<tr>
<th>Block</th>
<th>Pinaceae</th>
<th>Sorghum</th>
<th>Hordeum</th>
<th>Oryza</th>
<th>Solanaceae</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TaxId:3318</td>
<td>TaxId:4557</td>
<td>TaxId:4512</td>
<td>TaxId:4527</td>
<td>TaxId:4070</td>
</tr>
<tr>
<td>1</td>
<td>0.087 (23)</td>
<td>0.062 (16)</td>
<td>0.000 (11)</td>
<td>0.048 (21)</td>
<td>0.200 (20)</td>
</tr>
<tr>
<td>2</td>
<td>0.000 (16)</td>
<td>0.000 (12)</td>
<td>0.000 (13)</td>
<td>0.154 (13)</td>
<td>0.125 (8)</td>
</tr>
<tr>
<td>4</td>
<td>0.118 (34)</td>
<td>0.259 (27)</td>
<td>0.243 (37)</td>
<td>0.367 (49)</td>
<td>0.444 (36)</td>
</tr>
<tr>
<td>10</td>
<td>0.000 (13)</td>
<td>0.071 (14)</td>
<td>0.071 (14)</td>
<td>0.179 (28)</td>
<td>0.455 (11)</td>
</tr>
<tr>
<td>Total</td>
<td>0.091 (243)</td>
<td>0.137 (161)</td>
<td>0.151 (199)</td>
<td>0.214 (290)</td>
<td>0.313 (211)</td>
</tr>
</tbody>
</table>

Correlated ANOVA analysis between duplicate blocks (Bowers et al., 2003) indicates a duplication that occurred prior to the divergence of the Solanaceae and after the divergence of Pinaceae. However, interpretation of the timing relative to multiple monocot databases provides ambiguous placement of the event. Letters (A,B) indicate statistically different groups (\( \alpha = 0.05 \)) based on Tukey’s analysis with 250 random gaussian bootstrapped blocks based on \( \mu \) and \( \sigma \) estimates from the Arabidopsis block data.

Departments will lead to better utilization of genomic comparisons across dicots will be critical for attempting to detect homologous regions across this divide (Vandepoele et al., 2002).

### DISCUSSION

Whole genome duplication is a major evolutionary force affecting the structure of many eukaryotic genomes. Understanding and characterizing these duplications is a complicated task due to rearrangements and deletions that obscure duplicated blocks and differential rates of sequence evolution that confound dating efforts.

A phylogenetic approach to dating duplication events has numerous advantages over alternative approaches to determine whether a genome-wide duplication event happened before or after divergence of different taxonomic lineages. The gene trees generated by the analysis are readily classifiable into two categories, and these categories correspond with two different models of speciation and duplication. The phylogenetic dating implementation described here provides a flexible tool for characterizing whole genome duplications.

Arabidopsis and Oryza duplications were examined using this implementation. The results showed the utility of the tool and raised critical questions about the timing of duplication events relative to the monocot–dicot divergence. Understanding the numerous duplication events which have occurred throughout the history of plant genome evolution will lead to better utilization of genomic comparisons across...
species, facilitating the use of model species in understanding economically important crops.

Overall, this implementation provides a practical tool for researchers interested in elucidating the timing of whole genome duplication events.

ACKNOWLEDGEMENTS

We thank J. Estill for valuable comments on the manuscript, and J. Kissinger and E. Kraemer for suggestions on the techniques. B.C. is supported by a Howard Hughes predoctoral fellowship.

REFERENCES


