Models for Similarity Distributions of Syntenic Homologs and Applications to Phylogenomics

David Sankoff, Chunfang Zheng, Yue Zhang, João Meidanis, Eric Lyons and Haibao Tang

Abstract—We outline an integrated approach to speciation and whole genome duplication (WGD) to resolve the occurrence of these events in phylogenetic analysis. We propose a more principled way of estimating the parameters of gene divergence and fractionation than the standard mixture of normals analysis. We formulate an algorithm for resolving data on local peaks in the distributions of duplicate gene similarities for a number of related genomes. We illustrate with a comprehensive analysis of WGD-origin duplicate gene data from the family Brassicaceae.

Index Terms—gene tree, species tree, whole genome duplication, algorithms, mixture of distributions, Brassicaceae.

1 INTRODUCTION

In this paper, we study the set of \( \binom{N}{2} + N \) distributions of similarity of homologous gene pairs within and across \( N \) species where WGD has affected one or more of these species. This typically involves many thousands of genes. In Section 2.1, we first sketch out a model of gene similarity distribution under random sequence divergence, speciation and fractionation, providing the basis for a principled treatment of the statistical inference of divergence and fractionation rates and for speciation and WGD times.

In the rest of Section 2, we work out the detailed combinatorial and probability analyses of the case of two WGD in a genome (Section 2.2), the case of three WGD (Section 2.3), and the case of one whole genome triplication (WGT) followed by a WGD (Section 2.4). Although we have not yet developed a general software package implementing our methodology, we can nonetheless calculate particular cases using R routines. In Section 2.5, we illustrate with calculations for the \( \text{Populus trichocarpa} \) (poplar) genome under the WGT + WGD model.

Where Section 2, following [1], deals only with paralogous gene pairs \( (N = 1) \), in Section 3 to 6, we broaden the focus to include the orthologous gene pairs generated by speciation \( (N \geq 2) \). As an example of \( N = 2 \) in Section 3, we analyze the case of a WGD followed by a WGT followed by a speciation. The latter analysis allows us to dissect the similarity distribution between orthologous gene pairs in \( \text{Brassica rapa} \) and \( \text{Brassica oleracea} \).

For \( N \geq 3 \), we enter the domain of phylogenetics. A combinatorial explosion of the number possible gene trees for \( N \geq 3 \), already evident for the \( N = 2 \) case in Section 3, precludes the kind of exhaustive case-by-case analysis presented in Section 2 until some automated procedure can be developed. We can nonetheless proceed by simply identifying local modes or "peaks" in all the similarity distributions, and translating these into phylogenetically related paralogous and orthologous event times. In Section 4, we propose an algorithm, improving on the one in [1], for inferring a rooted tree from these event times, where the WGD are located on specific lineages in a consistent way. This means that the peaks identified in a comparison of any two genomes correspond to all the WGD in the lineage of their common ancestor, plus a more recent peak corresponding to the speciation event from which the two genomes diverge, with no more recent WGD-generated peaks.

In Section 5, we pursue the study of the order Brassicales with the application of the new algorithm. Our data contains twelve genomes spanning several genera of the family Brassicaceae as well as one genome from the sister-family Cleomaceae.

2 DISTRIBUTIONS OF GENE SIMILARITY

2.1 The building blocks

To analyze distributions of homolog similarity, it is a common practice in genomics to resort to numerical procedures embodied in software such as EMMIX [2] for resolving mixtures of normal distributions. These methods, however, powerful and flexible as they may be, are not tailored to the problem of detecting speciation and WGD in a set of related similarity distributions. For any mixture of normals, EMMIX will identify these components as long as there is enough data. But not every mixture of normals credibly reflects some sequence of genomic events. More important, among the \( \binom{N}{2} + N \) gene similarity distributions within and across \( N \) species, there are many constraints that are not handled by software packages, such as requiring the time estimate to be the same for an event in all the distributions that are affected by it.

Here we model gene pair divergence in terms of a probability \( p \) reflecting similarity — the proportion of nucleotide positions that are occupied by the same base in the two genes, orthologs or paralogs, although the same principles
of distribution of duplicate gene pair similarities is an estimate of the size of a WGD. We denote by \( \rho \) the fractionation parameter.

Two WGD

Consider a genome that has undergone two successive WGDs. We denote by \( t_1 \)-pairs and \( t_2 \)-pairs those duplicate gene pairs created at \( t_1 \) and \( t_2 \) respectively, with expected similarities \( p_1 \) and \( p_2 \). For fixed \( \rho \), \( u \) and \( v \) are functions of \( t_1 \) and \( t_2 \) only, representing the probabilities \( e^{-\rho(t_1-t_2)} \) and \( e^{-\rho(t_2-0)} = e^{-\rho t_2} \), respectively, for a pair of genes present at the start of the time interval, that neither gene is lost by the end of the interval. Note that in this and later models, we assume, for simplicity, that a fractionation regime from one WGD is supplanted by that set into operation by the next WGD. That is, fractionation involving older pairs is no longer operative.

In Figure 1, let

\[
A = \mathbb{E}(t_1 \text{ pairs}) = 4uv^2 + 4u(1-v) + u(1-v)^2 = u(1+v)^2
\]

\[
B = \mathbb{E}(t_2 \text{ pairs}) = 2uv^2 + 2u(1-v) + (1-u)v = v(1+u)
\]

\[
C = \mathbb{E}(\text{unpaired genes}) = (1-u)(1-v)
\]

\[
P(A) = \text{proportion of } t_1 \text{ pairs} = \frac{A}{A+B+C}
\]

\[
P(B) = \text{proportion of } t_2 \text{ pairs} = \frac{B}{A+B+C}
\]

\[
P(C) = \text{proportion unpaired} = \frac{C}{A+B+C}
\]

For a fixed gene length \( G \) and \( \lambda \), let \( N_p(s) \) be the density at point \( s \) of a normal distribution with mean \( p \) and variance \( \frac{p(1-p)}{G} \). The probability that gene pair will be observed to be with similarity \( s \in [0, 1] \) is

\[
Q(s) = P(A)N_{p_1}(s) + P(B)N_{p_2}(s)
\]

and the probability of an unpaired gene is

\[
Q^* = P(C).
\]

The likelihood of a data set with gene pairs at \( s_1, \ldots, s_t \) and \( k \) unpaired genes is

\[
L = \prod_{i=1}^{t} Q(s_i) Q^k
\]
The log likelihood $L = \log \mathcal{L}$ is

$$L = \sum_{i=1}^{l} \log Q(s_i) + k \log Q^* \tag{12}$$

$$= \sum_{i=1}^{l} [\log(P(A)N_{p_1}(s_i) + P(B)N_{p_2}(s_i))] + k \log Q^*$$

There is no closed form for the maximum likelihood of a mixture of normals, so in practice we use numerical means such as Newton-Raphson or an EM algorithm to derive the MLE.

### 2.3 Three WGD

Consider now three successive WGD affecting a genome (for example the $\tau, \sigma$ and $\rho$ WGD that occurred in the common ancestor of the cereals [4]). The scenarios producing various numbers of gene pairs of various ages are depicted in Figure 2, where $u, v$ and $w$ are the retention probabilities for pairs produced at $t_1, t_2$ and $t_3$.

$E(t_1 \text{ pairs}) = (1 - 3u^2 + 2w)w^2 + (2 + 6u^2 + 4w)uw + (1 + u^2 + 2w)u$

$E(t_2 \text{ pairs}) = (1 + w^2 + 2w)u + 1 + w^2 + 2w)v

$E(t_3 \text{ pairs}) = -2uv^2w^2 + ((2u^2 - w)u + uv + w)$

$E(\text{unpaired}) = (1 - u)(1 - v)(1 - w) \tag{13}$

From this analysis, we can predict the number of pairs remaining from the each of the three events, and perform MLE calculations to determine the parameters.

### 2.4 Whole genome triplication followed by WGD

The core eudicots contain more species than all the other groups of flowering plants combined. A whole genome triplication occurred in the eudicot lineage just before the emergence of the core eudicots, and a large number of these have undergone further WGD. The model analyzed in Figure 3 is appropriate for this case. Here

$E(t_1 \text{ pairs}) = (u'' + 3u'''v)^2 + (2u'' + 6u'''v + b + 3u''')v$

$E(t_2 \text{ pairs}) = -3u'''v^3 + 3u'''v^2 + (1 + 2u''' - u'')v$

$E(\text{unpaired}) = (1 - u''' - u')(1 - v). \tag{14}$

### 2.5 The case of Populus trichocarpa

Though the work we have presented consists of combinatorial models, and the inference procedures are not implemented in a user-friendly package, we did analyze one data set using functions on the R platform according to the model in the previous section. We extracted data on the poplar genome [5] from the CoGe platform [6], [7], and calculated syntenic gene pairs, producing the distribution of similarities in Figure 4. Focusing on syntenic, or colinear, sets of duplicate genes tends to ensure that these pairs are all produced during the same WGD or speciation event. This excludes tandem pairs and isolated pairs out of syntenic context.

In estimating the parameters using our preliminary code, we were confined to sampling 500 out of the 13,000 pairs.
Running the sampling repeatedly, the results were quite reproducible, represented by the blue density curve in the diagram.

In Figure 4, the broadening of the calculated peak representing a recent WGD (at \( p = 0.89 \)) is partially due to the inclusion of the highly similar pairs at the far right of the diagram, possibly alleles of the same gene. The small volume of the earlier peak is accurately reflected in very small estimates of \( u''' \) and \( u' \), less than 0.1, compared to about 0.5 for \( v \).

**3 THE EFFECT OF SPECIATION**

Up to now we have considered only WGD events, including triplications. In comparing two species, there are peaks at times corresponding to all their shared WGD, followed by a single peak dating from their speciation event, but no further peaks. Additional WGD after speciation increase the number of orthologs and paralogs in a uniform way across the board, but do not change the number of peaks.

**3.1 Of neep and kraut**

The broadening of effects of event age and of fractionation on the peaks in similarity distributions as time elapses are well illustrated by the two peaks in Figure 4. Eventually, all events become indistinguishable from noise caused by random gene resemblances, widespread domain sharing, tandem and near-tandem duplications, gene-order rearrangements, gene conversion and other processes.

We keep this in mind as we compare *Brassica rapa* and *Brassica oleracea*. These species are thought to descend from a common WGT “\( \gamma \)” at the base of the eudicots (the same as the wide peak in Figure 4), two WGD “\( \beta \)” and “\( \alpha \)” (shared with *Arabidopsis* and other members of the order Brassicales) and another WGT specific to the family Brassicaceae [8]. It is known, however, that before the two most recent events,
there was a rapid speed-up of gene divergence processes, so that the earlier events are difficult to discern in comparative data. Data from CoGe on *B. rapa* and *B. oleracea* shows a clear speciation peak at 97% similarity and a clear WGT peak at 89% similarity, as in Figure 5. There is a large shoulder on the curve, likely due to the large overlap of the similarity distributions due to α, β and γ. Other evidence such as the distribution of $K_v$ scores (not shown) and the comparisons within and between the *Arabisopsis* genomes (detailed below), suggest a hidden peak at 81%. This motivated us to set up a three-event model, with a WGD at 81%, a WGT at 89% and a speciation at 97%.

The model is depicted in Figure 6. The probability that an α paralog pair (generated at time $t_1$ before the present) survives until time $t_2$ is $u$. The probability that all three pairs generated at the *Brassica* triplication time $t_2$ before the present survive until time $t_3$ is $v$ and the probability that only one pair survives in $v'$. The probability that an ortholog pair generated at the speciation time $t_3$ survives to the present is $z$.

For each kind of pair (i.e., according to its time of origin) adding up the number of these in each of the 39 configurations shown, multiplied by the number of versions (in parentheses), multiplied by the probability of the configuration as written, we obtain

$$\mathbb{E}(t_1 \text{ pairs}) = u + 21uv^2z^2 - 16uw^2z^3 + 4uv^2z^4 + 68uwvz^4 + 4uv^2 + 2uz + 2uv$$
$$-72uv'vz^4 + 72uv'vz^3 + 6vz + 3uv$$
$$-68uw'vz^4 + 144uw^2 - 144uz^3v$$
$$+14uv'v' - 7uz^2v' + 6uvz + 3uz^2v$$
$$-12uv^2 + 72uz^2v^2 - 72uz^4v^2 + 72uz^3$$
$$-72uz^4 + z^2v'$$,

(16)

$$\mathbb{E}(t_2 \text{ pairs}) = 24uv'z_4 + 6uzv'v + 3uzv - 48uz^3v$$
$$-24uv'z^3 - 4uzv^2 + 48uvz^4$$
$$+24uz^3v^2 - 24uz^4v^2 + vz - 24uz^4$$
$$+2z - 24uv'vz^4 + 24uz^3 + 24uv'uz^3(17)$$

and

$$\mathbb{E}(\text{unpaired}) = (1 - u)(1 - v' - v)(1 - z).$$

(18)

Among the $t_1$ pairs in this analysis half will be paralogous and will not show up in the predicted distribution of ortholog similarity. The same is true of the $t_2$ pairs, but the $t_3$ pairs are all orthologous. Based on these facts, we can calculate the density of ortholog pairs similarity for maximum likelihood or other estimates of $u, v, v'$ and $z$. Figure 5 compares this density, based on $u = 0.2, v = 0.15, v' = 0.15$ and $z = 0.5$, with the distribution of similarities in the data. It can be seen that the fit is reasonable, but not perfect, especially at the earliest times. The pairs at this time probably reflect the γ triplication. The combinatorial task of adding yet another WGT event to the configurations in Figure 6 is beyond our current scope of our manual methods.

Interestingly, attempts to fit three normal these same data with EMMIX returns only two normals, with the peaks at 0.87 and 0.81 being replaced by a single intermediate density with large area, with mean at 0.82, fitting data from very early dates (similarity from 0.6-0.7), but corresponding very poorly to the data from the rest of the pre-speciation period (especially from 0.75-0.9).

4 A PHYLLOGENOMIC ALGORITHM BASED ON SPECIATION AND WGD

It is commonplace that the divergence of many of the $\binom{N}{2}$ pairs of species in an $N$-species phylogeny may date from a single speciation event, and this should be represented by a common node in the inferred phylogenetic tree. In the same way, $M \leq N$ species may be descendants of a given WGD, and this relationship should be represented by the position of this event on a branch of a directed tree, such that all $M$ of these species, and only these, are in the subtree descending from this event.

Because the computational problems with WGD-event detection illustrated in Section 3.1 preclude its application to large numbers of comparisons at this time, we will assume that we can infer $p$, and hence the age of an event, simply by identifying the mode, or “peak” of the similarity distribution, without recourse to other estimation procedures. This unfortunately foregoes any attempt at present to pick out events visible as “shoulders” of other events on a similarity distribution derived from a pair of species, like we did for the *B. rapa* – *B. oleracea* comparison, though it allows for these events to emerge from the comparison of one or more other pairs of species.
Fig. 6. Components of the set of surviving orthologous pairs in two species diverging at time \(t_1\) after a (shared) WGD at \(t_2\), and a shared WGT at \(t_3\). Number of cases of same component with different labelling in parentheses. "x + y" indicates \(x\) pairs dating from the common WGD, \(x\) pairs created by the shared WGT, and \(y\) pairs from the speciation event. Parallel dotted and solid lines distinguish two species.
4.1 The algorithm

There are two principles underlying our method for reconstructing a phylogeny from a set of inter- and intra-genome syntenic homolog similarity distributions:

- each intra-genome distribution of similarities can only have peaks due to the WGD in its direct lineage, and
- each inter-genome distribution may contain several peaks due to WGDs, but only one peak due to speciation, the most recent one, i.e., at the date of the most recent common ancestor of the two species.

To incorporate these principles into a phylogenetic inference procedure, we adapt the neighbor-joining (NJ) algorithm [9], to produce a directed tree rather than the undirected tree for which it was first designed. To attempt to satisfy the two principles above, we need to specify not only on which tree branch a WGD is located, but whether it is the the subtree emanating from one endpoint or the other of the branch that inherits the WGD, which implies specifying a direction to all branches, at least those leading from a WGD.

We preserve NJ’s transformation of a distance matrix allow for differing evolutionary rates on different branches of the phylogeny, at each step of the agglomerative procedure. We add two elements to the procedure, namely an “age” calculation for each possible new node that could be produced, and a simple dynamic programming step to align and evaluate how similar are the ancestral WGD of two nodes being aligned. This step potentially adds a penalty to the distance between two nodes if they do not share the same WGD, especially a recent WGD. With these added elements, we no longer have the NJ algorithm, strictly speaking, since the new elements cannot be incorporated into the distance matrix at the outset, but must be calculated during execution.

The input to Algorithm 1 includes a standard distance matrix, say of divergence times calculated from the $p$ at a speciation peak in the similarity distribution of each pair of genomes, plus a list of WGD times ancestral to each genome, obtained from the similarity distribution of syntenic paralogs.

5 The Brassicaceae

To illustrate our discussion, we draw on twelve published genomes in the Brassicaceae family, two in the genus Brassica: B. rapa (turnip, Chinese cabbage) [10] and B. oleracea (cabbage, cauliflower) [11], two in Raphanus: Raphanus sativus (radish) [12] and R. raphanistrum (wild radish) [13], two in the genus Arabidopsis: A. lyrata (rock cress) [14] and A. thaliana (thale cress, mouse-ear cress) [15] and one each in the genera Sisymbrium: S. irio (London rocket) [16], Schrenkiaella: S. parvula (dwarf spikerush) [17], Eutrema: E. salugineum (saltwater cress) [18]. Capsella: C. rubella [19] (red shepherds purse), Leavenworthia: L. alabamica (Alabama gladecress) [16] and Aethionema: A. arabicum [16]. In addition, we included one species from the sister family Cleomeae, genus Cleome: Tarenaya hassleriana (spider flower) [20].

Before reporting on the analysis of this full data set, we illustrate with the details of a subset of six of the species

\begin{algorithm}
\caption{Reconstruct the Modified Neighbour Joining Tree with WGDs}
\begin{algorithmic}[1]
\State \textbf{Input}: speciation times, WGD times, list of $n$ leaf genomes
\State \textbf{Output}: tree with WGD events
\State \textbf{while} $n \geq 3$ \textbf{do}
\State \hspace{1em} get the distance matrix as in the standard Neighbour Joining algorithm
\State \hspace{1em} for each pair of unconnected nodes \textbf{do}
\State \hspace{2em} get a candidate new node (parent) by joining them (two child nodes)
\State \hspace{2em} calculate the distances from this candidate to the other nodes as in the standard Neighbour Joining algorithm
\State \hspace{1em} get the minimum cost candidate using
\State \hspace{2em} \textbf{Algorithm 2}/* update the tree with this minimum cost candidate*/
\State \hspace{1em} set the WGDs
\State \hspace{1em} replace two old nodes by a new node
\State \hspace{1em} get $n−1$ unconnected nodes and distances among these $n−1$ nodes for the next iteration
\State \hspace{1em} set the root node by balancing the left-side distance and right-side distance
\State \hspace{1em} \textbf{return} the resulting tree
\end{algorithmic}
\end{algorithm}

\begin{algorithm}
\caption{Count Cost}
\begin{algorithmic}[1]
\State \textbf{Input}: $n$ unconnected nodes, $n(n−1)/2$ candidate nodes, distances from candidate nodes to unconnected nodes
\State \textbf{Output}: minimum cost candidate, placement of WGDs
\State \textbf{for} each candidate node \textbf{do}
\State \hspace{1em} for each genome (child node) of this candidate node \textbf{do}
\State \hspace{2em} parental age ← age of this child node + branch length between this child and its parent (candidate new node)
\State \hspace{2em} separate the WGDs into two groups:
\State \hspace{3em} group1: WGDs below parental age
\State \hspace{3em} group2: WGDs above parental age
\State \hspace{2em} keep the group1 WGDs in the child node
\State \hspace{2em} apply dynamic programming on the group2 WGDs of both child nodes to place them above joint node
\State \hspace{2em} age ← minimum of two parental ages
\State \hspace{2em} cost ← age
\State \hspace{2em} if the final penalty of the dynamic programming is more than a threshold then
\State \hspace{3em} add some extra cost to the candidate
\State \hspace{1em} \textbf{return} candidate with minimum cost, WGD placement
\end{algorithmic}
\end{algorithm}
(i.e., only \( \binom{6}{3} = 15 \) of the \( \binom{13}{2} = 78 \) of the pairwise comparisons and 6 self-comparisons), namely B. rapa, B. oleracea, R. sativa, A. lyrata, A. thaliana and S. irio. Figure 7 shows the phylogenetic relationship among the six species (cf. [16]).

We extracted genomic data from these species using the database in CoGe [6], [7]. We then used the SynMap routine (with default parameters) on this platform to compare the gene orders of each of the \( \binom{6}{3} = 15 \) pairs of genomes. This procedure identifies orthologs produced by speciation by detecting collinear arrays of several duplicate pairs in two species with approximately the same divergence: “syntenic blocks”. Similarly, we did a self-comparison of five of the six genomes; the sixth one, the Sisymbrium genome, did not have enough closely spaced duplicate pairs for SynMap to produce paralogous syntenic blocks. The distributions of similarities calculated are shown in Figure 8. The peaks found in each genome are tabulated in Table 1.

From Figure 8 and Table 1, we note that the earliest duplication, detected at 79-80% in the Arabidopsis self-comparisons, shows no peaks in the other self-comparisons – there is a shoulder or heavy tail in the appropriate place in the Brassica self-comparisons, but this is swamped by the later triplication. The triplication itself is visible in all three Brassica self-comparisons and in the comparison of B. oleracea and B. rapa, but not in the weaker signals involving Raphanus.

More interesting is that the peaks at 90% reflecting the Sisymbrium speciation, known to occur before the Brassica triplication, suggest that speciation is more recent, since the triplication peak is at 89%. This apparent conflict is clearly ascribable to a slower rate of evolution (lower \( \lambda \)), since the divergence of Arabidopsis from Sisymbrium also seems to occur more recently (88%) than the divergence of Arabidopsis from the Sisymbrium sister genus Brassica (86%). Note that the small differences between peak similarities are not insignificant, given the many thousands of gene pairs involved in these comparisons.

Applying our algorithm to the data derived from these six genomes reflects this rate anomaly, with S. irio branching after the triplication instead of preceding it, unless the dynamic programming penalty on discordant WGD evidence is increased, in which case S. irio branches with the Arabidopsis-Capsella group, an equally poor result.

Figure 9 depicts a phylogeny of our full set of Brassicaceae genomes (all of the sequenced ones we have been able to access through CoGe) plus one genome from the sister family Cleomaceae. Though most of this tree is uncontroversial, the taxonomic positions of S. irio, E. selsugineum and S. parvula has been changed several times in recent years [22]. The particular configuration shown in the figure is drawn from [16].

We computed all 78 pairwise comparisons and 12 of the self-comparisons and picked out the visible peaks in each. These were used as input to our algorithm. With any of range of reasonable values for the dynamic programming
TABLE 1
Peak similarity level, by genome. np: no peak, but one could be found by mixtures of distribution methods. -: no peak expected. Note peak 3 occurring before peak 4 due to slow evolutionary rate (λ) of Sisymbrium.

| peak number | description                           | BR  | BO  | RS  | SI  | AL  | AT
|-------------|---------------------------------------|-----|-----|-----|-----|-----|-----
| 1           | alpha duplication [21]                | np  | np  | np  | np  | 80  | 80,79
| 2           | divergence of genus Arabidopsis       | 86  | 86  | 86,87 | 88 | 88-86 | 88-86
| 3           | whole genome triplication             | 89  | 89  | 87  | np  | np  | np
| 4           | divergence of genus Sisymbrium        | 90  | 90  | 90  | -   | -   | -
| 5           | divergence of genus Raphanus          | 93  | 93  | 93  | np  | -   | -
| 6           | speciation of Arabidopsis T & L      | -   | -   | -   | 95  | 95  | -
| 7           | speciation of B. rapa & B. oleracea   | 97  | 97  | -   | -   | -   | -

Fig. 9. Brassicales phylogeny derived from the literature

Fig. 10. Phylogeny calculated by proposed algorithm

penalty, the output tree was as shown in Figure 10.

The only difference with Figure 9 is that the taxonomically volatile group S. irio, E. salsugineum and S. parvula appear as a monophyletic clade, sister to the Brassica-Raphanus group, which is not implausible, and is certainly preferable than branching S. irio within the latter, after the Brassica triplication. More important, without the dynamic programming constraint taking account of the placement of the WGD, L. alabamica branches before the Arabidopsis-Brassica split, whereas in Figure 10 it is appropriately grouped as a sister taxon to the Arabidopsis-Capsella clade.

6 CONCLUSION

We have introduced a concerted approach to plant phylogenomics that gives a central role to the whole genome duplication (WGD) and triplication evidence from similarity distributions of syntenic (collinear) homologs. The first component of this methodology is a combined combinatorial and probabilistic analysis of orthologous and paralogous gene pairs in pairwise genome comparisons and self-comparisons, biologically more interpretable than statistical mixture-of-distributions analyses. The second half uses the results from this analysis in a phylogenetic algorithm inspired by neighbour-joining, but that produces rooted trees to accommodate the inherent directionality of WGD. This algorithm incorporates a dynamic programming subroutine to align the ancestral WGD events of two observed or inferred ancestral genomes. For both the first part and the second part, we have implemented proof-of-principle software and applied these to genomes from the order Brassicales, in particular to the comparison of B. rapa and B. oleracea for the detailed analysis of a similarity distribution, and to 12 members of the family Brassicaceae, plus one outgroup, for the phylogenomics.

Our theoretical considerations pertain to the simple model assumed at the beginning of this section. In practice, various other processes affect the distribution of similarities so that the number of gene homologs between and within genomes may be severely reduced from those expected from the model. We have seen that the divergence rate λ may vary somewhat for individual lineages, and ρ is certainly even more variable. Genomic processes such as chromosomal rearrangements disrupt gene order and degrade the recovery of synteny blocks and duplicate gene pairs. These issues should all be addressed in future work. DNA substitution models with more parameters and rate variation among sites could also be incorporated. For example, one commonly used distance metric $K_s$ (substitutions per synonymous sites) is typically calculated using more specific codon substitution models. The $K_s$ distance scales...
linearly with time and \( \log n \).

**Acknowledgments**

This work was partially funded by a NSERC Discovery grant, NSF grant IOS-1339156 and Fapesp grant 2016/01511-7. DS holds the Canada Research Chair in Mathematical Genomics.

**References**


David Sankoff  David Sankoff received the PhD degree in Mathematics from McGill University, and has been a member of the Centre de recherches mathematiques in Montreal for many years. He currently holds the Canada Research Chair in Mathematical Genomics in the Mathematics and Statistics Department at the University of Ottawa, and is cross appointed to the Biology and the Computer Science Departments. His research interests include comparative genomics, particularly probability models, statistics, and algorithms for genome rearrangements, with a focus on the genomes of flowering plants.

Chunfang Zheng  Chunfang Zheng received the master’s and PhD degree in Biology from the University of Ottawa, where she has been a research associate in Dr. Sankoff’s lab. She has published extensively on algorithms for genome rearrangements and participated in the evolutionary analysis for many flowering plant genome sequencing projects.

Yue Zhang  Yue Zhang received the master’s degree in Statistics from Carleton University in Ottawa and is currently working toward the PhD degree in Dr. Sankoff’s lab at the University of Ottawa, focusing on the statistical analysis of subgenome evolution after paleopolyploidy.

João Meidanis  João Meidanis received the PhD from University of Wisconsin-Madison, with a thesis on DNA fragment assembly and algorithms for gene functional prediction. He is a faculty member of the University of Campinas where he founded the Computational Biology group. He has coordinated the bioinformatic analysis for several genome projects and is the director of the consulting enterprise Scylla Bioinformatics.

Eric Lyons  Eric Lyons received the PhD in Plant Biology from the University of California, Berkeley. Previously, he spent five years working in industry at biotech, pharmaceutical, and software companies. Dr. Lyons was responsible for developing the CoGe platform for comparative genomics data management and analysis. He has published extensively on these topics and has been invited to teach numerous workshops on the analysis of genomic data to plant, vertebrate, invertebrate, microbe, and health researchers. He is currently Assistant Professor in the School of Plant Sciences, CyVerse.

Haibao Tang  Haibao Tang earned his PhD in Plant Biology from the University of Georgia. He is currently a professor at Fujian Agricultural and Forestry University and Senior Scientist at the University of Arizona, working on genome assembly, annotation, and analyses of plant genomes. He has devised novel computational methods to clarify the evolution of flowering plants, which he has applied to papaya, sorghum, Medicago, Brassica, tomato, Amborella and numerous other genomes. He has two pending patents for isolating genes responsible for important agronomic traits in the cereal crops.