CHAPTER 3

RECONSTRUCTING CONCORDANCE TREES AND TESTING THE COALESCENT MODEL FROM GENOME-WIDE DATA SETS

Cécile Ané

3.1 INTRODUCTION

In the absence of gene flow between ancestral species, a phylogenetic tree faithfully represents species relationships. Even so, gene trees can truly differ from the species tree due to incomplete lineage sorting (ILS), as explained in previous chapters. Other processes can also cause discordance among gene trees (Maddison 1997; Wendel and Doyle 1998). On the one hand, some processes cause difficult identification of true orthology or difficult inference of gene trees. On the other hand, processes involving gene flow—such as horizontal gene transfer, hybridization, hybrid speciation, or introgression—constitute an integral part of the species genealogy. In some cases, extensive gene flow may challenge the concept of a single species tree. For example, if the genetic material of an ancestral hybrid species is composed of exactly 50% from one parental species and 50% from the other parental species, then any single bifurcating tree cannot represent the actual history of these species. In unicellular and prokaryotic groups, organisms can engage in extensive lateral transfer, making it unclear that a single tree can faithfully represent the species genealogical relationships (Doolittle and Bapteste 2007; but see Galtier and Daubin 2008). The concept of concordance factors (CFs) was introduced by Baum (2007) to capture the variability of gene genealogies. Even though a single tree may not represent the history of a set of taxa, the genealogy at each site along the genome follows an actual tree. Bayesian concordance analysis (BCA) considers this plurality of trees, and uses CFs to measure the proportion of the genome that has a given clade (Ané et al. 2007). CFs might be used to describe an ancestral hybrid speciation event, for instance, by providing the proportion of genetic material that the ancestral species inherited from each of its two parents. One way to summarize information provided by CFs of many conflicting clades is to build a "dominant history" from the clades with the highest CFs. Baum (2007) argued that regardless of the processes that caused gene trees to disagree, clades supported by a plurality of the genome can be used as a representative of the dominant species history. In the ideal case, the species history is truly treelike, and ILS within the species tree is the only reason why gene trees disagree. In this case, it is desirable that the dominant history reconstructed from CFs matches the actual species tree. A much more complicated scenario is when horizontal gene transfers have a substantial role in the evolutionary

Estimating Species Trees: Practical and Theoretical Aspects, Edited by L. Lacey Knowles and Laura S. Kubatko

history of the organisms. Galtier and Daubin (2008) argued that both the vertical signal (speciation/extinction) and the horizontal signal (lateral genetic transfer) should be reconstructed. It is desirable in this case that the dominant history reconstructed from CFs matches the vertical phylogenetic signal, and that the horizontal phylogenetic signal be recovered from the CFs of clades that are not in the dominant history.

In this chapter, I explain how BCA can be used to reconstruct the dominant history—or the vertical phylogenetic signal—of a set of taxa, even if the processes that caused gene tree discordance are unknown. I also explain how to get new insights into which processes may have caused discordance among gene trees. All concepts are illustrated with real examples. Section 3.2 provides some background on the concordance approach implemented in the program BUCKy (Bayesian Untangling of Concordance Knots). Section 3.3 discusses the interpretation of CFs as genomic support, and how it differs from standard measures of statistical support that are usually obtained on gene trees. Section 3.4 demonstrates how credibility intervals around CFs can be used to get statistical support for particular branching patterns in the concordance tree. Section 3.5 illustrates the use of CFs to test the hypothesis that all discordance is due to ILS as modeled by the coalescent process, and Section 3.6 links concordance trees and species trees. Finally, Section 3.7 considers a challenge posed by genome-wide alignments: that of finding loci within which all sites share the same tree topology. A fast method is proposed to find homogeneous loci for later use by gene tree/species tree (GT/ST) inference methods.

3.2 BCA: BACKGROUND

BCA takes in a set of sequence alignments, such as a set of different genes, to reconstruct the dominant tree for the taxa under study. This dominant tree is built from the clades that are inferred to be true for a high proportion of genes in the genome. BCA assumes that all sites within a given alignment have tracked the same tree topology but recognizes that different alignments may have tracked different trees. A Bayesian framework is used to integrate out the different sources of uncertainty. If two different genes give very high support for two different trees for instance, BCA will recognize that the discordance between the two gene trees is likely due to a true difference between the underlying gene trees, not to a lack of phylogenetic information. In this regard, BCA is similar to BEST (Bayesian Estimation of Species Trees), the coalescent-based Bayesian method for species tree reconstruction (Liu 2008). Both methods use the same likelihood for a set of aligned sequences given a set of gene trees.

3.2.1 Sharing of Information across Gene Trees

BCA and BEST differ in the prior distribution they assign to a set of gene trees. BEST uses the distribution obtained from the coalescent process along a species trees so that two estimated gene trees influence each other even if these gene trees differ. In order to allow for any process of gene tree discordance, BCA uses a Dirichlet prior distribution on gene tree topologies (Ané et al. 2007). This prior distribution captures the expectation that many genes agree with the species tree—and therefore with each other. One parameter, α, is required to specify the strength of our expectation that different genes share the same topology. This parameter is similar to an a priori level of discordance because the Dirichlet prior assigns a probability $(1+\alpha/T)/(\alpha+1) \sim 1/(\alpha+1)$ that two randomly selected genes share the same tree, where T is the total number of possible gene tree topologies. At one extreme, the choice $\alpha = 0$ corresponds to a 1.0 prior probability that two randomly chosen

genes share the same tree. It corresponds to the prior assumption that all genes in the genome have the same topology. With this choice, information from all alignments is combined as in a concatenation approach to infer a single tree. At the other extreme, the choice of α = infinity corresponds to the smallest prior probability that two genes share the same topology ($1/T$), just as if gene trees were independent. This extreme choice corresponds to a consensus approach where gene trees are estimated separately and do not influence each other. With intermediate levels of α, BCA clusters genes into a number of sets. Genes in different clusters are inferred to have different genes, and their gene tree estimates do not influence each other. Genes that are placed in the same cluster are inferred to share the same tree topology. Sequences from all genes in the same cluster are thus combined to obtain a more accurate estimate of their common tree topology. Uncertainty in gene clustering is accounted for and integrated out.

3.2.2 How to Choose the A Priori Level of Discordance α

BCA was implemented in BUCKy (Larget 2008). A default prior level of discordance $\alpha = 1$ was chosen, which corresponds to a prior probability of $0.5 + 1/(2T) \sim 0.5$ that two randomly chosen genes share the same topology. If this prior probability seems too large or too low a priori in a particular system, the α value can be adjusted by the user accordingly. An interactive Web site* is available to users who want to adjust α to match their prior expectation. This Web site takes in an α value, a number of taxa, and a number of genes. It then plots the prior distribution for the number of distinct tree topologies represented in the set of sampled genes. As Galtier and Daubin (2008) demonstrate, the a priori probability that two genes share the same topology should decrease with the number of taxa, even under the coalescent process along a treelike species history. Therefore, the adequate prior level α is expected to increase with increased taxon sampling. For this reason, it was desirable that the interactive Web site be able to visualize the effect of taxon number on the prior distribution of gene trees.

3.2.3 The Choice of an Infinite α in BCA

The prior level α = infinity amounts to assuming independent gene trees, and provides an interesting case. It is a conservative choice, in that estimation error in one gene tree does not affect the estimation of other gene trees. Undetected paralogy might cause such errors. The use of inadequate evolutionary models can also cause systematic errors, such that an incorrect gene tree may receive very high support. The conservative choice α = infinity will prevent errors in one gene to affect other genes. However, this conservativeness comes at a cost because phylogenetic information is not shared across genes. Mathematically, it is easy to show that the CFs of clades estimated with α = infinity will just be the average posterior probabilities of clades, averaged over all genes in the sample. Therefore, the concordance tree estimated with α = infinity will match the consensus tree built from the individually estimated gene trees. BCA will further provide credibility intervals around CFs, which the consensus approach does not. However, discordance is expected to be overestimated with an infinite α because it is confounded with uncertainty in gene tree reconstruction. When a finite α is chosen, genes with low phylogenetic information can be clustered with other genes. The pooled phylogenetic information across genes in the same cluster allows tree uncertainty to decrease, and no longer be confounded with true discordance among gene trees.

* http://bigfork.botany.wisc.edu/concordance/

3.2.4 A Nonparametric Prior Distribution on Gene Trees

The Dirichlet prior distribution used in BCA is nonparametric in the sense that there is no limitation to the number of clusters of genes, each cluster representing a group of genes that have the same topology. In the current implementation of BCA, the topologies of different clusters are given a uniform prior over all topologies and are assumed to be independent. This prior assumption can account for any kind of horizontal gene transfer or hybridization or both, or even for any kind of undetected paralogy in an outlier gene. However, the assumption that truly different gene trees are independent does not reflect the expectation that different gene trees might still share many clades in common. Future work will use a different prior distribution, in which gene trees from different clusters will be able to influence each other.

3.3 GENOMIC SUPPORT VERSUS STATISTICAL SUPPORT

We illustrate here the concept of a CF as a measure of *genomic support*, in contrast to the usual measures of *statistical support* that are commonly used to annotate phylogenetic trees. Bootstrap support and posterior probabilities of clades are calculated on a 0–1 or 0–100% scale, just like CFs. However, these measures mean very different things. Bootstrap values and posterior probabilities aim to measure how *confident we are* that a particular clade truly is in the one tree that is assumed to drive the evolution of the gene(s). In contrast, CFs measure *how much of the genome* (or how many of the sampled genes) truly have a particular clade in their tree. It is possible for a clade in the species tree to be true in only 60% of the genes and for two other conflicting clades to be true for 20% of the genes each. For the clade in the species tree, we would like to be able to make two statements: a statement about genomic support (60% of genes truly have this clade) and a statement about statistical support (1.0 posterior probability that this clade is in the species tree). BCA aims to provide both genomic support and statistical support at once.

We illustrate the different kinds of support with a set of 30,066 alignments from human, chimpanzee, gorilla, orangutan, and rhesus, which were assembled and analyzed by Ebersberger et al. (2007). In their paper, Ebersberger et al. (2007) focused on the genomic support for the human-chimpanzee clade. They estimated that 77% of our genome shares immediate genetic ancestry with the chimpanzee genome. They also provided a statement of statistical support based on the number of alignments that were included in their analysis: the percentage of our genome sharing immediate ancestry with the chimpanzee was estimated with high precision, as its standard error was 0.4% and its 95% confidence interval was (76.2%, 77.8%). Because the lower limit of this confidence interval is above 50%, we are confident that the human–chimpanzee sister relationship is truly in the species tree. Actually, the 99.99% confidence interval for the human-chimp CF would still be well above 50%. Therefore, the study by Ebersberger et al. (2007) provides both the genomic support (77% of the genome has human-chimp sister to each other) and the statistical support (almost 100% confidence that human-chimp are sisters in the species tree). The 77% CF and the ~100% confidence level are by no means contradictory.

Because Ebersberger et al. (2007) ignored uncertainty in the reconstruction of individual gene trees, we reanalyzed their data with BCA. BCA does not assume any clock, and therefore all 30,066 alignments were included in our analysis. Only a subset of 11,945 highly informative clocklike alignments was used by Ebersberger et al. (2007). Like in their original study, each locus was analyzed under the HKY (Hasegawa, Kishino, Yano;

Hasegawa et al.
categories. Becau
analyzed in MrBa
This individual l
except for the pri
0.02). Twenty-six

The compl
BCA. Three prio
values were chos
is the most extre
assumed to be in
dance on five tax
genes share the s
a very low level
randomly sample
(1 and 0.1), seque
more accurate e
independent runs
(Larget 2008) ve
central processir
second step com

Figure 3.1
The human-chim
than half of the g
these two clades
posterior probab
11.2% of loci i
human-gorilla cl
estimates are ve
was obtained wi
with >95% poste
argued, their cor
tree discordance
and it is remark
19,000 lower-qu

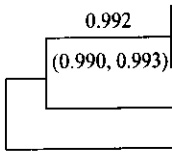


Figure 3.1 Baye
Ebersberger et al.
branches) measur
for concordance f
(*Right*) Concorda
and human-gorill
expected when in

Hasegawa et al. 1985) substitution model with rate variation across sites and four rate categories. Because there are only five taxa and for computational speed, each locus was analyzed in MrBayes (Ronquist and Huelsenbeck 2003) with 110,000 generations per run. This individual locus analysis constituted the first step of BCA. Default values were used except for the prior distribution of branch lengths (exponential distribution with prior mean 0.02). Twenty-six alignments were excluded because of corrupted data files.

The complete samples provided by MrBayes were then used in the second step of BCA. Three prior levels of discordance were used: $\alpha = 0.1$, 1, and infinity. These three values were chosen to encompass a very wide array of prior discordance levels. Infinity is the most extreme value, corresponding to a consensus approach where gene trees are assumed to be independent. The default level $\alpha = 1$ represents a moderate level of discordance on five taxa because it corresponds to a 0.53 probability that two randomly sampled genes share the same topology a priori. At the other extreme, the value of $\alpha = 0.1$ provides a very low level of discordance. Indeed, it corresponds to a 0.997 probability that two randomly sampled genes share the same topology a priori. With the two finite levels of α (1 and 0.1), sequences from compatible gene trees will influence each other so as to provide more accurate estimations of their common tree topologies. For this second step, four independent runs, three chains, and 110,000 generations were used in the program BUCKy (Larget 2008) version 1.3.0. The first step of BCA took about 4.1 days using three 3.0 GHz central processing units (CPUs) for a total of about 12.5 days of CPU time, while the second step completed in 5.5 h on a single CPU.

Figure 3.1 summarizes the results with $\alpha = 1$, showing both measures of support. The human-chimp clade and the human-chimp-gorilla clade received support from more than half of the genome with 1.0 posterior probability with all three prior choices. Therefore, these two clades are dominant and the species tree shown in Figure 3.2 receives a 1.0 posterior probability. Still, significant discordance among loci is detected: 12.2% and 11.2% of loci in the genome are estimated to have tracked the chimp-gorilla and the human-gorilla clades, using $\alpha = 1$. The same results were obtained with $\alpha = 0.1$. These estimates are very similar to those in Ebersberger et al. (2007). Note that their estimate was obtained with one-third of all loci only, from fragments whose tree was supported with >95% posterior probability and that were consistent with a molecular clock. As they argued, their consensus approach would mistakenly interpret gene tree uncertainty as gene tree discordance if low-informative loci were used. Here, BCA was run on all 30,040 loci, and it is remarkable that the estimated concordance has not dropped, even though about 19,000 lower-quality alignments were included for this analysis. This consistency can be

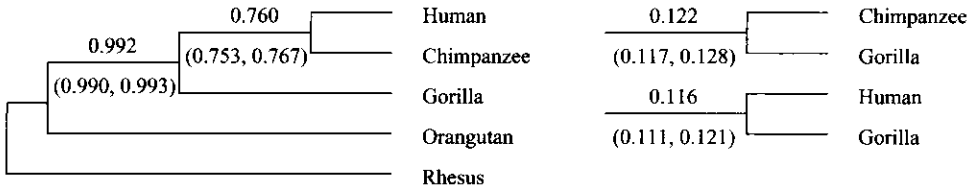


Figure 3.1 Bayesian concordance analysis of the great ape data (30,040 alignments from Ebersberger et al. 2007) with $\alpha = 1$. Genome-wide concordance factors of clades (above branches) measure genomic support. Statistical support is provided by 95% credibility intervals for concordance factors (below branches). (*Left*) Concordance tree, 1.0 posterior probability. (*Right*) Concordance factors of conflicting clades. The concordance factors of the chimp-gorilla and human-gorilla clades are not significantly different (their credibility intervals overlap) as expected when incomplete lineage sorting is the only cause of gene tree discordance.

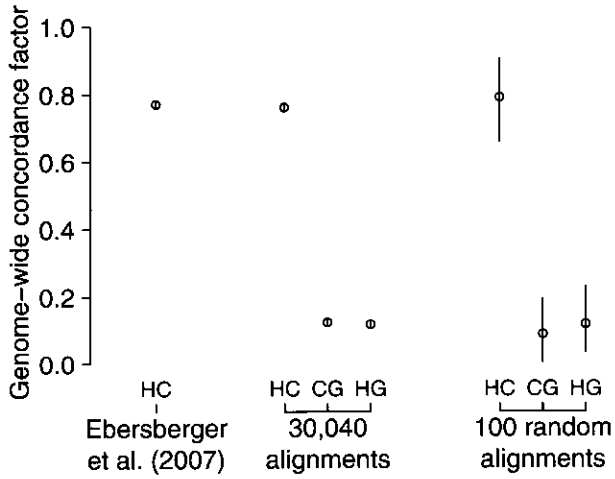


Figure 3.2 Estimated concordance factor (proportion of the genome) for the human-chimpanzee (HC), chimp-gorilla (CG) and human-gorilla (HG) clades, with their 95% confidence/credibility intervals, from Ebersberger et al. (2007) (*left*) and from concordance analysis with $\alpha = 1$ (*center* and *right*). The concordance analysis infers that about 76–77% of the human genome is sister to chimpanzee, with a 1.0 posterior probability that the HC clade is in the species tree: the credibility interval for this clade's concordance factor does not overlap with credibility intervals for the concordance factor of other conflicting clades.

explained by a sufficient number of highly informative loci to influence the number of clusters, the common topology for each cluster of genes, and the proportion of genes in each cluster. The uncertainty from low-quality loci is not confounded with discordance because these loci are combined with highly informative loci in clusters with robustly estimated trees.

To apply a consensus approach to the full set of 30,040 loci and obtain credibility intervals for CFs, the highest a priori level of discordance can be used in BCA (α infinite: no sharing of information across genes). With this value, gene trees are assumed to be independent so that information is not shared across alignments, just like in a consensus approach. Given the high number of low-informative alignments in the data set, the consensus prior level α = infinity is expected to overestimate discordance. Not surprisingly, a much higher proportion of the genome is inferred to have tracked a tree different from the species tree with this choice of α. The CF of clades with high genomic support is underestimated: 0.545 (0.542, 0.549) for human-chimp and 0.848 (0.846, 0.851) for the human-chimp-gorilla clade. As expected, the CF of clades with low genomic support is overestimated: 0.200 (0.197, 0.203) for chimp-gorilla and 0.197 (0.194, 0.201) for human-gorilla. However, this discordance-biased analysis still infers a higher CF for the human-chimp and for the human-chimp-gorilla clades than for any other clades. Therefore, the species tree is still inferred with a 1.0 posterior probability with α infinite.

Statistical support such as bootstrap values, posterior probabilities, and standard errors for CF all reflect the amount of sampling error. Therefore, they heavily depend on the amount of data: the larger the sample size, the lower the sampling error. For instance, if the same gene or same set of genes is replicated several times, bootstrap values and posterior probabilities for clades in the estimated tree will increase up to 100% or 1.0. Standard errors and confidence intervals will shrink to a width of zero as more and more (identical) data sets are used. In contrast, genomic support is not expected to change with the amount of data: estimated CFs are expected to remain stable as more and more genes

are sampled. To illustrate this contrasting behavior of the two kinds of support, we randomly sampled 100 alignments from the great apes data set and analyzed them with BCA. Figure 3.2 shows that the resulting estimated CFs are similar to those obtained using the full set of 30,066 alignments. However, their precision is a lot lower than from the full data set. Statistical support was indeed expected to decrease with a reduced amount of data. Even though CFs are estimated with a lot of uncertainty, the correct concordance tree is still inferred with a 1.0 posterior probability from these 100 alignments.

3.4 COMPARING CFs OF CONTRADICTING CLADES FOR RECONSTRUCTING THE DOMINANT HISTORY

The primary concordance tree summarizes phylogenetic relationships that are true for large proportions of the genome. Among a set of contradicting clades, one would like to find the clade that is supported by more of the genome than any of the other clades. This dominant clade, which is more representative of the species relationships than any of the conflicting clades, is more worthy of being represented in the concordance tree. Because CFs are inferred with some estimation error, however, it may not always be possible to determine with certainty that a higher proportion of genes supports one clade than another conflicting clade. A fast comparison of CFs can be made on the basis of their credibility intervals, so as to determine if one CF is significantly higher than another one, or if instead there is insufficient data to make the comparison. Very simply, if the two CFs have non-overlapping 99% credibility intervals, then we can infer with high credibility (over 98%) that one of the two clades indeed has a higher CF than the other clade. In order to obtain an exact posterior probability that one particular clade has a higher CF than another clade, it would simply suffice to count the proportion of Markov chain Monte Carlo (MCMC) samples that satisfy this relationship, from the output of the program BUCKy. Comparing credibility intervals of individual CFs is a much easier and faster alternative.

In the great ape data set, most of the discordance is located within the human-chimp-gorilla clade. It is easy to determine that the human-chimp clade is dominant here because its CF is above 50% with 1.0 posterior probability. In such a case, any conflicting clade must receive support from less than 50% of the genome. When the level of discordance is higher however, the CF of the dominant clade may be closer to or lower than 50%. We provide such an example below, where we compare credibility intervals of CFs in order to determine which clade is dominant with statistical significance.

Rodriguez et al. (2009) investigated the evolution of wild tomatoes and wild potatoes (*Solanum*) using multiple markers which were randomly selected from a set of 2869 single-copy orthologs (conserved orthologous sets II, or COSII; Wu et al. 2006). Figure 3.3 shows the concordance analysis of 12 of these loci across nine wild potato species and two close outgroups (*Solanum etuberosum* and *Solanum palustre*). In this wild potato phylogeny, most clades have CFs above 0.50 with over 95% posterior probability. However, the CF of the *Solanum brevicaule-Solanum verrucosum* clade includes 0.50 (0.474, 0.924): it is not unlikely that less than 50% of the genome have this clade. In order to determine if this clade is dominant, we examined the CFs of the two conflicting clades (Fig. 3.4). *Solanum raphanifolium-S. brevicaule* has a CF within (0.029, 0.425), while *S. raphanifolium-S. verrucosum* has a CF within (0, 0.14) with 95% credibility. Because of the lack of overlap (Fig. 3.4), we can conclude that the *S. brevicaule-S. verrucosum* clade is dominant, and we can place this clade in the concordance tree with high posterior probability. A similar comparison between the CFs of the three possible placements of the main wild potato clades reveals a lack of resolution (Fig. 3.5). Given the amount of discordance

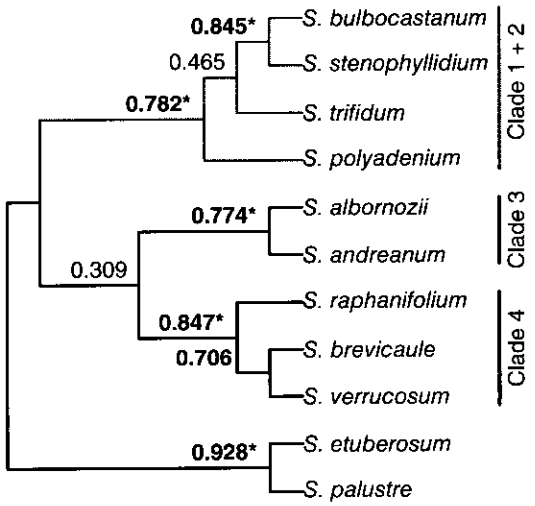


Figure 3.3 Concordance analysis ($\alpha = 1$) from 12 loci on wild potatoes (*Solanum*; Rodriguez et al. 2009). Above branches: genome-wide concordance factor estimates. Bold values indicate clades that are dominant with high credibility. Asterisks indicate clades with concordance factor above 50%, with over 0.95 posterior probability.

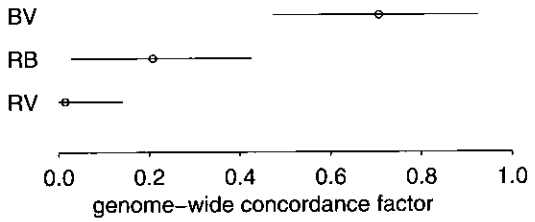


Figure 3.4 Comparing the concordance factors of three conflicting clades: *Solanum brevicaule-Solanum verrucosum* (BV), *Solanum raphanifolium-Solanum brevicaule* (RB), and *S. raphanifolium-S. verrucosum* (RV). Estimated genome-wide concordance factors (o) and their 95% credibility intervals (—).

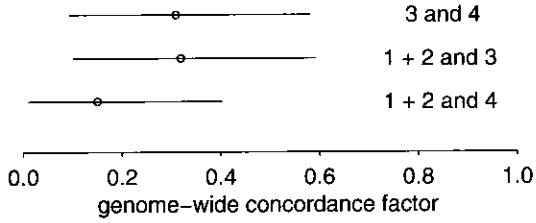


Figure 3.5 Estimated genome-wide concordance factors (o) and their 95% credibility intervals (—) for the three resolutions of the main potato clades shown in Figure 3.3. All three credibility intervals overlap, showing uncertainty about which resolution has the highest concordance factor.

among genes regarding the placement of these three groups, there is insufficient data at this time to determine which of the three resolutions is supported by the largest proportion of the genome. A parametric coalescent-based approach such as STEM (Species Tree Estimation using Maximum likelihood, Kubatko et al. 2009) or BEST (Liu et al. 2008) might have greater power to determine the species tree resolution at this node, at the cost of assuming that all discordance is due to ILS.

3.5 TESTING THE HYPOTHESIS THAT ALL DISCORDANCE IS DUE TO ILS

The most parsimonious, null model for explaining gene tree discordance is that of ILS only. Mathematically, the coalescent model for ILS predicts exact relationships between the CFs of clades that do not belong in the species tree, also called minor clades. On a four-taxon unrooted tree or a three-taxon rooted tree for instance, there are three possible gene trees, or three possible clades. Under the coalescent model, the clade in the species tree is true for the largest proportion of genes, while a minority of genes follows each of the other two clades. The coalescent model predicts that these two minor clades have equal CFs. If t is the length of the internal branch in coalescent units (number of generations/effective population size) then the CF for this branch in the species tree is $CF = 1 - 2/3\exp(-t)$, while the CFs of the two competing splits are equal, and equal to $CF = 1/3\exp(-t)$. The exact values of these CFs depend on the branch length t in the species tree, but the equality between the CFs of the two minor clades can be tested even if t is unknown. More generally, the ILS hypothesis along a branch predicts that the two minor resolutions of that branch have equal CFs.

With its nonparametric assumptions, BCA can be used to test the hypothesis that ILS is the sole discordance mechanism. For any given lineage, we can reject the ILS hypothesis if we can determine that the two minor resolutions of that branch have CFs that differ with statistical significance. With BCA, a simple way to assess statistical significance is to compare the credibility intervals for the CFs of the two minor resolutions. If the credibility intervals do not overlap, then we can be confident that the two CFs differ. On the other hand, if the credibility intervals of the conflicting clades' CFs overlap, the data do not provide evidence that these CFs differ, and no evidence against the ILS hypothesis.

We illustrate this test on a set of 18 COSII markers sequenced and analyzed by Rodriguez et al. (2009). We report here their concordance analysis on six wild tomato species, four tomato outgroup species (two species from section *Juglandifolia* and two species from section *Lycopersicoides*), and two further outgroups (*Solanum dulcamara* and *Datura inoxia*). Significant discordance was found at the placement of the close outgroups of the core wild tomatoes. The two species of section *Juglandifolia* formed a clade with a high CF and high statistical support, as did the two species from section *Lycopersicoides*, as well as the two further outgroups (*S. dulcamara* and *D. inoxia*) and the core tomato group. However, there was significant discordance between the three possible placements of the four groups mentioned above: the sister group to the core tomatoes was section *Lycopersicoides* for about half of the loci, and section *Juglandifolia* for the other half of the loci. The third placement, with sections *Lycopersicoides* and *Juglandifolia* sister to each other, received no support from any locus. Figure 3.6 shows the estimated genome-wide CFs for each of these three resolutions, along with their 95% credibility intervals. This analysis used a prior level of discordance $\alpha = 1$, and the conclusion was identical with a higher $\alpha = 10$.

This pattern of discordance is incompatible with the assumption of ILS as solely responsible for gene tree discordance. Because the credibility interval for the most minor history (the two tomato outgroups sister to each other) does not overlap with the credibility interval of any of the other two CFs, we can reject the null ILS hypothesis. Instead, this pattern of discordance is compatible with the hypothesis that the ancestral lineage leading to the core tomato group is a hybrid between the ancestral lineages of sections *Lycopersicoides* and *Juglandifolia*. Under this hypothesis, the CFs of the two major clades are expected to be about equal. Because the data included only 18 loci, the CFs of these

toes (*Solanum*; Rodriguez
ates. Bold values indicate
es with concordance factor

lades: *Solanum brevicaule-*
ule (RB), and *S.*
nce factors (o) and their

95% credibility intervals
re 3.3. All three credibility
highest concordance factor.

here is insufficient data at
d by the largest proportion
as STEM (Species Tree
r BEST (Liu et al. 2008)
n at this node, at the cost

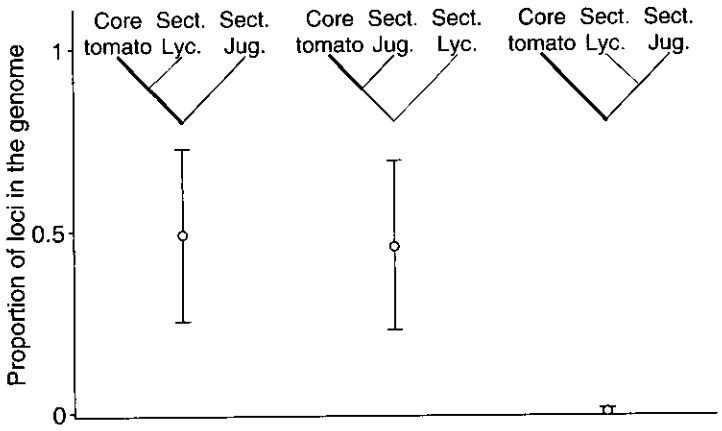


Figure 3.6 Estimated genome-wide concordance factors (o) and their 95% credibility intervals (—) for the three placements of the tomato outgroups relative to the core tomato group. A further outgroup is not shown. This pattern of discordance is incompatible with the incomplete lineage sorting hypothesis, which would predict that the two minor resolutions have equal concordance factors.

two major placements have quite wide credibility intervals. The pattern observed here could also be compatible with section *Juglandifolia* being sister to the core tomatoes and gene flow between the ancestral lineage of core tomatoes and the ancestral lineage of section *Lycopersicoides*. Collection of sequence data across more loci would be needed to determine which evolutionary process has shaped the gene tree discordance at the base of the core tomato group.

From the great ape data, an estimated 12.2% and 11.2% of the genome have tracked the chimp-gorilla and the human-gorilla clades (Fig. 3.1), using $\alpha = 1$ and $\alpha = 0.1$ as well. The overlapping and narrow credibility intervals for these two CFs provide evidence that the chimp-gorilla and the human-gorilla clades have very similar, if not equal CFs. This is in complete agreement with the hypothesis that ILS is the only source of gene tree discordance along the ancestral human-chimp-gorilla lineages.

An alternative, more computationally intensive test of the ILS hypothesis would be to determine if the pattern of gene tree proportions found nonparametrically by BCA can be predicted from a species tree under the coalescent. The program COAL (Degnan and Salter 2005) calculates the gene tree proportions predicted by the coalescent on a user-defined species tree. At this time, however, there is no program that takes the proportions of a set of unrooted gene topologies as input, and estimates the most likely species tree under the coalescent hypothesis. Nevertheless, I estimated the species tree branch lengths based on the simple relationship $p = 2/3e^{-t}$, where t is the internal branch length (in coalescent units) of a four-taxon asymmetric rooted tree and p is the proportion of gene trees that are truly discordant with the species topology. Proportions p were estimated nonparametrically from BCA for various four-taxon sets, and branch lengths t were then calculated with $t = -\log(3p/2)$. The prior choice ($\alpha = 1$ and $\alpha = 0.1$) and alternative choices of four-taxon sets had almost no impact of the estimated branch lengths. Figure 3.7 shows the estimated species tree with branch lengths in coalescent units. Coalescent gene tree probabilities were then obtained with COAL, which were used to calculate the CFs predicted by the coalescent model. The data were consistent with this model, as all the predicted CFs were within the credibility intervals obtained from BCA ($\alpha = 1$ and $\alpha = 0.1$).

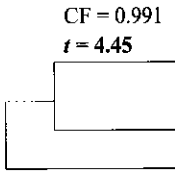
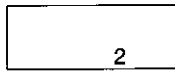


Figure 3.7 Great
conflicting splits (
within the credibi



Consistent
greedy consensu

Figure 3.8 Two
(*right*) is in the to

3.6 SPECIES TREE

Under the coale
and the species
that case, if CF
species tree with
method to recor
above 50% as v
dicted by a clad
that such a gree
the species tree
Rosenberg 2006
gies and branch
tree, even if fed

The two
((((H,C),G),O),
tral to the HCG
to one AGT (fig
the proportion o
tion of genes ha
vs. 11.1% for tr

For these
are shown in Ta
struct a tree fro
the tree, as indi
rectly reconstru
the split HCG|C

Degnan et
to consistently r

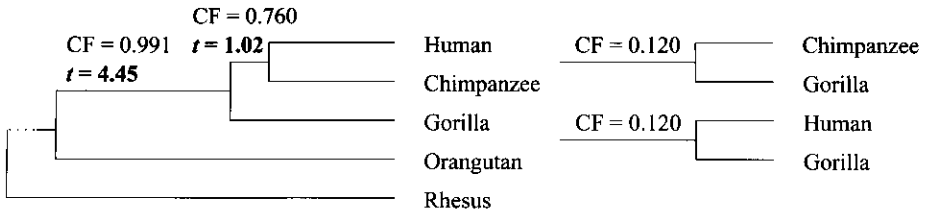


Figure 3.7 Great ape species tree (*left*) with estimated branch lengths t in coalescent units, and conflicting splits (*right*). The concordance factors predicted by the coalescent model (CF) are within the credibility intervals shown in Figure 3.1.

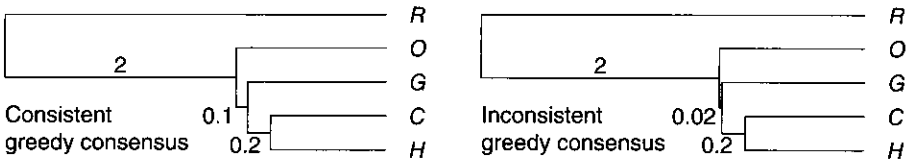


Figure 3.8 Two hypothetical species trees, each giving rise to one anomalous gene tree. Tree 2 (*right*) is in the too-greedy zone (fig. 3.3 in Degnan et al. 2009), while tree 1 (*left*) is not.

3.6 SPECIES TREE RECONSTRUCTION FROM CFs

Under the coalescent, there is a perfect relationship between the frequencies of gene trees and the species tree with branch lengths in coalescent units (Degnan and Salter 2005). In that case, if CFs are known with precision, one should also be able to reconstruct the species tree with high certainty. BUCKy currently uses a "greedy" majority-rule consensus method to reconstruct a primary concordance tree. This tree includes all clades with CF above 50% as well as all clades whose CF is below 50% so long as they are not contradicted by a clade with higher CF already in the tree. Degnan et al. (2009) recently showed that such a greedy method might yield an incorrect species tree when internal branches of the species tree are short and when there are anomalous gene trees (AGTs; Degnan and Rosenberg 2006). They identified a "too-greedy" zone, that is, a set of species tree topologies and branch lengths where the greedy consensus method returns an inconsistent species tree, even if fed correct CFs. I illustrate this below.

The two hypothetical species trees in Figure 3.8 share the same topology ((((H,C),G),O),R). Their branch lengths, in coalescent units, differ along the branch ancestral to the HCG clade. Under the coalescent model, both of these species trees give rise to one AGT (fig. 3.3 in Degnan et al. 2009). Indeed, the coalescent process predicts that the proportion of genes with unrooted topology ((H,C),(G,O),R) is higher than the proportion of genes having the species unrooted topology (14.4% vs. 13.2% for tree 1, and 15.0% vs. 11.1% for tree 2, calculated with COAL).

For these two species trees under the coalescent model, the CFs of various clades are shown in Table 3.1. The greedy consensus approach uses these proportions to reconstruct a tree from the clades with the highest CFs until no more clades can be included in the tree, as indicated with bold numbers. If these CFs are inferred perfectly, tree 1 is correctly reconstructed by this method but tree 2 is not: the split GO|HCR is preferred over the split HCG|OR.

Degnan et al. (2009) further showed that the quartet-based consensus method is able to consistently recover the true species tree from CFs of quartets under the coalescent model.

TABLE 3.1. Concordance Factors of Splits on the Full Taxon Set HCGOR, Given the Two Species Trees in Figure 3.8 under the Coalescent Model

Splits, Full Taxon Set	Tree 1 (%)	Tree 2 (%)
HC|GOR	**38.7**	**36.8**
HCG|OR	**25.1**	20.7
GO|HCR	21.2	**23.6**
HG|COR	20.5	18.7
CG|HOR	20.5	18.7

Splits with bold values are included in the species tree reconstructed by the greedy consensus method.

TABLE 3.2. Concordance Factors of Splits on the Four-Taxon Set CGOR

Quartet Resolutions	Tree 1 (%)	Tree 2 (%)
CG|OR	**39.7**	**34.6**
GO|CR	30.2	32.7
CR|OG	30.2	32.7

The split in the true species tree is necessarily supported by the highest proportion of genes under the coalescent model. Bold values indicate which quartets are included in the species tree reconstructed by the quartet-based consensus method.

With the two previous examples, the GO split was reconstructed by the greedy consensus approach on tree 2, but it is not supported from any four-taxon set. For example, the correct four-taxon split CG|OR is preferred over GO|CR based on their concordance (Table 3.2).

The work by Degnan et al. (2009) has important implications for BCA. Even though the estimation of CFs by BUCKy may be accurate, the primary concordance tree built with the currently implemented greedy method may differ from the species tree under ILS. However, this inconsistent behavior can easily be fixed by using a quartet-based method. This method builds four-taxon unrooted trees based on the resolution that is supported by the highest proportion of genes, and then combines the selected quartets into a full tree. If CFs are estimated correctly, then all selected quartets will be consistent with the species tree even in cases when the greedy consensus approach is misleading (Degnan et al. 2009). A future version of BUCKy will implement this quartet-based method for species tree estimation. Note that Ewing et al. (2008) use a similar reconstruction method.

The results by Degnan et al. (2009) prove that concordance trees equal species trees when the coalescent model is correct. Their results also suggest that the best way to summarize gene tree variability into a single bifurcating tree is to use the quartet-based tree reconstruction from CFs. When gene flow, horizontal gene transfers, or hybridization events are prevalent, it is unclear that a single species tree may represent the actual history of the taxa (Galtier and Daubin 2008). A single concordance tree may reflect the vertical phylogenetic signal, but a species *network* might provide a more accurate description of species relationships, with nodes of different types. Speciation nodes may depict the vertical phylogenetic signal, and hybridization nodes may provide the horizontal signal (Baroni et al. 2005; Huson and Bryant 2006; Moret et al. 2004). Hybridization nodes need not be constrained to indicate an exact fifty-fifty chance to choose either of the two parents,

however. More realistic probability models could be developed in the future to represent various kinds of gene flow under a species network (e.g., Jin et al. 2006; Meng and Kubatko 2009). Such network representation is used by Holland et al. (2008), where gene trees embedded in the network are called "principal trees" (see also Than et al. 2008). When a network, rather than a tree, is representative of the species history, it is unclear how the primary concordance tree or primary concordance network should be built from estimated CFs, and there is room for more research in this area.

3.7 THE CHALLENGE OF DETERMINING LOCI ON WHOLE-GENOME ALIGNMENTS

3.7.1 The Assumption of Homogeneous, Unlinked Loci for GT/ST Reconstruction

To date, almost all methods for GT/ST estimation from multiple loci make the very practical assumption that all sites within a locus share the same tree topology. In other words, each locus is conveniently assumed to have tracked a single topology, or be "topologically homogenous." This is a very restrictive assumption, however, when dealing with very long or whole-genome alignments (Posada and Crandall 2002). Whole genomes can now be sequenced and aligned (Darling et al. 2004; White et al. 2010; Yang et al. 2007). These very long alignments do not come up with predefined loci, unfortunately. If genealogical variation is expected, then it is necessary to locate regions that are topologically homogeneous before using the current methods for estimating species trees and studying gene tree variation. A second problem is that adjacent loci contain spatial dependence, and that current GT/ST methods assume unlinked loci with independent trees given the species tree. This issue of ignoring spatial dependence may not be too severe if this dependence tapers off quickly as we consider more and more distant loci. However, the robustness of current GT/ST methods to violation of the unlinked locus assumption is unknown at this time and is outside the scope of this chapter.

Given a very long alignment, methods like STEM (Kubatko et al. 2009), BEST (Liu et al. 2008), BCA (Ané et al. 2007), or deep coalescence parsimony (Maddison 1997; Maddison and Knowles 2006; Oliver 2008; Page 1998) cannot be applied without first partitioning this long alignment into a number of hypothetically topologically homogeneous loci. Ideally, of course, the inference of loci and the inference of trees should be performed simultaneously, so that the uncertainty of the partition can be considered for inference of locus trees and of the species tree. A number of Bayesian or hidden Markov chain methods have been developed to simultaneously infer recombination breakpoints and the phylogeny at each site of an alignment (e.g., Bloomquist et al. 2009; Husmeier and McGuire 2003; Minin et al. 2005; Suchard et al. 2003; Webb et al. 2009). These methods do not seek to infer the species tree or dominant history, however. We report here on a fast phylogenetic method that has been applied for studying genealogical variability and reconstructing the dominant history of house mice (White et al. 2010).

3.7.2 Detecting Recombination Breakpoints for GT/ST Reconstruction

A fast and easy way to partition a long alignment is to define equal-length loci. Yang et al. (2007) used 100-kb intervals, for instance. Ideally, the choice of the fragment length should strike a compromise between a high probability that each given fragment is topologically

homogeneous (using shorter fragments) and a high phylogenetic informativeness of individual fragments (using longer fragments). In this section, we focus instead on data-driven methods for defining fragments, which rely on the detection of recombination.

There is a confusingly large number of methods for detecting the presence of recombination or the location of recombination points. Posada and Crandall (2001) offer a review and comparison of 14 of these methods, and a more recent nonexhaustive list is maintained by Jun Fan and David Robertson.† Posada et al. (2002) identify five categories of methods: similarity methods, distance methods, phylogenetic methods, compatibility methods, and substitution distribution methods. The number of available methods is partly explained by a number of different goals and different meanings of "recombination breakpoint." Recombination is achieved via meiosis in eukaryotic organisms, and via processes such as homologous recombination, conjugation, transformation, or transduction in prokaryotes. Some detection methods use sequence data from multiple individuals in a single species to estimate the population recombination parameter $\rho = 4N_e r$, where r is the recombination rate per site and per generation and N_e is the effective population size. This parameter ρ considers all recombination events, even though some events do not change the tree topology on either side of the recombination location, and some events do not change the tree branch lengths, as measured in number of generations between coalescent events (Hein et al. 2005). Other methods aim at detecting only those recombination events that actually changed part of the tree: branch lengths and/or topology.

For the purpose of reconstructing gene trees and species trees from sequence data, recombination events that did not change the tree topology nor its branch lengths are of *no* interest. This means that the set of important recombination breakpoints is highly taxon-dependent. Consider for instance an original alignment that contains only one sequence from a particular species. Assume that two extra individuals from that species are further sequenced, so that the original alignment is expanded into a larger alignment containing all three sequenced individuals from that species. In this situation, all recent recombination events that affected the relationship between the three individuals need to be detected for the analysis of the larger alignment. On the other hand, these recombination points should be ignored for the analysis of the original alignment since these recombination events did not affect the tree that contains a single individual from the particular species.

Moreover, I argue that it is most important to detect recombination events that affect the tree topology, whereas it is of lesser importance to detect recombination events that only affect branch lengths. Indeed, branch lengths inferred from standard model-based phylogenetic methods represent an average number of substitutions per site, which is the product of substitution rate and of divergence times. These gene tree branch lengths are known to be highly variable: even if divergence times can be assumed to be homogeneous within one locus, substitution rates are known to be highly variable across sites and across lineages (Pagel and Meade 2008; Whelan 2008; Zhou et al. 2007). Therefore, phylogenetic reconstruction methods must account for varying branch lengths within a single homogeneous alignment, whether this is due to substitution rate variability or to divergence time variability from recombination. Not surprisingly, many methods for detecting recombination are sensitive to mutational "hot spots" and other substitution rate heterogeneity if changes in branch lengths are detected as recombination (Grassly and Holmes 1997; Husmeier 2005; McGuire and Wright 2000). For these reasons, methods based on topological changes and insensitive to branch lengths seem most appropriate for the purpose of defining loci for later use by methods for estimating species trees.

† http://www.bioinf.manchester.ac.uk/recombination/programs.shtml

3.7.3 A Minimum Description Length (MDL) Information Criterion

I propose to partition chromosome-wide alignments using a fast MDL approach (Ané and Sanderson 2005). MDL is a widely used tool for model selection. It aims to find an appropriate balance between a good fit to the data and a parsimonious model, just like the commonly used AIC and BIC criteria (Akaike 1974; Schwarz 1978). Here, each partition constitutes a model, including a particular choice for the number of loci and breakpoint locations. Like AIC or BIC, MDL aims to maximize the fit of the partition to the data while penalizing the partition's complexity. The founding principle of MDL is an information theoretic principle, which permits a direct comparison between the complexity of the data and the complexity of the model. The model complexity is measured by the length of the smallest code that can describe the model, and its fit to the data is measured by the length of the smallest code that can describe the data given the model. Using a practical compression algorithm, Ané and Sanderson (2005) showed that the joint complexity of a partition and alignment, or its description length, can be measured as a function of the number of fragments in the partition, and the sum of parsimony scores of each fragment. We consider here a similar criterion:

$$\mathrm{DL} = \underbrace{L_1 + \ldots + L_k}_{\text{fit}} + \underbrace{\lambda k}_{\text{penalty}}$$

where k is the number of fragments in the partition, L_i is the parsimony score of the i^{th} fragment and λ is a penalty parameter that penalizes each fragment. The total parsimony score $L_1 + \ldots + L_k$ of the alignment measures the fit of the model, which here consists of the partition and the k trees. Because the parsimony score is proportional to the negative log-likelihood of the alignment under a no-common mechanism model (Tuffley and Steel 1997), the DL criterion takes the form of a penalized likelihood, just like AIC and BIC. These criteria differ in how they penalize each parameter in the model: AIC uses a constant penalty. BIC uses a penalty that grows with the size n of the data ($\log n$). With MDL, Ané and Sanderson (2005) showed that the appropriate penalty λ for each fragment of the partition grows with the number of taxa ($\lambda \sim$Ntax), as does the complexity of the tree for each fragment. In order to search for the partition with the best, minimum DL, we use dynamic programming (program available upon request). A more detailed exposition of MDL will be published elsewhere.

The MDL criterion was used to determine the dominant history of the three closely related species of house mice: *Mus musculus*, *Mus castaneus*, and *Mus domesticus*. White et al. (2010) obtained whole-genome alignments based on Perlegen Sciences Single Nucleotide Polymorphism (SNP) data (Frazer et al. 2007) and on the complete genome sequences of the C57BL/6J strain (Mouse Genome Sequencing Consortium 2002) and of *Rattus norvegicus* as an outgroup (Rat Genome Sequencing Project Consortium 2004). The X chromosome and all 19 autosomes were analyzed, representing a 1.8 billion site alignment across four taxa. MDL was first applied to identify putatively homogeneous loci. BCA was then used on these loci to infer the proportion of each chromosome supporting each topology. Using a penalty of $\lambda = 3$ in MDL, the genome was partitioned into a total of 14,081 fragments of variable sizes. With a lower penalty of $\lambda = 0.9$, about twice as many fragments were identified but the resulting estimated CFs were very similar to those obtained with the higher penalty. White et al. (2010) identified a primary history placing *M. musculus* and *M. castaneus* as sister species in 39% of loci. They also uncovered a pattern of discordance that is inconsistent with the coalescent model, as the two minor histories had significantly different CFs, with higher genomic support for the *M. castaneus-M. domesticus* group than for the *M. musculus-M. domesticus* group.

3.7.4 Comparisons with Other Partitioning Criteria

The MDL approach proposed here has limitations: uncertainty in the number of loci and in breakpoint locations is not assessed and thus ignored in the subsequent GT/ST analysis. One advantage to using a parsimony-based measure of fit, rather than a model-based likelihood, is the computational speed and the corresponding ability to handle very long alignments. RecPars (Hein 1993) is a similar parsimony-based approach, which also seeks a balance between low parsimony scores and few recombination breakpoints. However, it is not clear how the cost of recombination should compare with the cost of substitutions in RecPars, and the algorithm does not scale well with long alignments or large numbers of taxa. Similarly, the program Recco (Maydt and Lengauer 2006) needs a user-defined ratio to weigh the costs of recombination and mutation, and the parsimony-based method RECOMP (Ruths and Nakhleh 2006) needs a user-defined threshold to define breakpoints. MDL provides a way to place the cost of recombination and of homoplasy on an equal footing, that of information complexity. Munshaw and Kepler (2008) also use the MDL principle for detecting recombination breakpoints. Their measure of fit counts different types of substitutions and is related to the parsimony score when the number of steps is small compared with the number of sites. Their method constrains the trees on either side of each breakpoint to differ by a single recombinant node, whereas our MDL criterion does not restrict the fragment topologies in any way. For a given number of taxa, the MDL criterion is very similar to AIC, as both penalize each fragment with a fixed penalty.

GARD (Genetic Algorithm for Recombination Detection), developed by Kosakovsky Pond et al. (2006), uses AIC with a likelihood fit based on a simple model of molecular evolution, which may be sensitive to substitution rate variability. The penalty term in GARD penalizes each branch length parameter of the tree at each fragment, but it does not penalize the complexity of each tree topology as MDL does.

More recently, a number of methods have used probabilistic models for the number and location of recombination breakpoints to account for their uncertainty, using Bayesian inference or hidden Markov models (HMM) (Bloomquist et al. 2009; Husmeier and McGuire 2003; Minin et al. 2005; Suchard et al. 2003). Due to their computational complexity, these methods are either limited to four or five taxa, or they need to be guided by a known phylogenetic tree on parental, nonrecombining sequences. Webb et al. (2009) increase the number of taxa that can be handled by combining an HMM with a Bayesian framework for the state space of this HMM. While these approaches do not seek species tree reconstruction, future developments seem particularly promising for the integrated inference of recombination breakpoints with species tree reconstruction.

The wealth of data provided by chromosome-wide alignments contains ample information regarding the dominant history and the genealogical variability along the genome. However, current GT/ST methods cannot handle this kind of data directly. Coupling computationally efficient methods such as MDL with BCA or STEM provides a first step toward analyzing chromosome-wide alignments for species tree inference.

ACKNOWLEDGMENTS

The author thanks Ingo Ebersberger for kindly sharing the aligned fragments from great apes and rhesus, Bret Larget and David Baum for stimulating discussions, and Laura Kubatko and Lacey Knowles for their very helpful comments on an earlier version of this chapter. This work was partially funded by the United States Department of Agriculture, National Research Initiative Grant 2008-35300-18669.

REFERENCES

Akaike, H. 1974. A new look
tification. *IEEE Transact*
19:716–723.
Ané, C. and M. J. Sanderson
the trees: phylogenetic con
for inferring complex evol
Biology 54:146.
Ané, C., B. Larget, D. A. Bau
2007. Bayesian estimation
trees. *Molecular Biology a*
Baroni, M., C. Semple, and
for representing reticul
Combinatorics 8(4):391–4
Baum, D. A. 2007. Concordar
and the exploration of
56:417–426.
Bloomquist, E. W., K. S. Dorn
StepBrothers: inferring par
recombinant viral sequence
Darling, A. C. E., B. Mau, F.
2004. Mauve: multiple alig
sequence with rearrangem
1394–1403.
Degnan, J. H. and N. A. Ros
species trees with their n
Genetics 2:e68.
Degnan, J. H. and L. A. Salter
under the coalescent proce
Degnan, J. H., M. DeGior
Rosenberg. 2009. Properti
inferring species trees from
58:35–54.
Doolittle, W. F. and E. Baptes
the tree of life hypothesis.
Academy of Sciences of t
104:2043–2049.
Ebersberger, I., P. Galgoczy
Platzer, and A. von Haes
genetic ancestry. *Molecu*
24:2266–2276.
Ewing, G., I. Ebersberger, H.
2008. Rooted triple consen
BMC Evolutionary Biolog
Frazer, K. A., E. Eskin, H. M
Hinds, E. J. Beilharz, R. V.
Morenzoni, G. B. Nilser
Stuve, F. M. Johnson, M
D. R. Cox. 2007. A sequ
8.27 million SNPs in i
448:1050–1053.
Galtier, N. and V. Daubin. 20
in phylogenomic analyses.
the Royal Society of Londo
363:4023–4029.
Grassly, N. and E. Holmes
for the detection of selec

REFERENCES

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19:716–723.

Ané, C. and M. J. Sanderson. 2005. Missing the forest for the trees: phylogenetic compression and its implications for inferring complex evolutionary histories. *Systematic Biology* 54:146.

Ané, C., B. Larget, D. A. Baum, S. D. Smith, and A. Rokas. 2007. Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution* 24:412–426.

Baroni, M., C. Semple, and M. Steel. 2004. A framework for representing reticulate evolution. *Annals of Combinatorics* 8(4):391–408.

Baum, D. A. 2007. Concordance trees, concordance factors, and the exploration of reticulate genealogy. *Taxon* 56:417–426.

Bloomquist, E. W., K. S. Dorman, and M. A. Suchard. 2009. StepBrothers: inferring partially shared ancestries among recombinant viral sequences. *Biostatistics* 10:106–120.

Darling, A. C. E., B. Mau, F. R. Blattner, and N. T. Perna. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* 14: 1394–1403.

Degnan, J. H. and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genetics* 2:e68.

Degnan, J. H. and L. A. Salter. 2005. Gene tree distributions under the coalescent process. *Evolution* 59:24–37.

Degnan, J. H., M. DeGiorgio, D. Bryant, and N. A. Rosenberg. 2009. Properties of consensus methods for inferring species trees from gene trees. *Systematic Biology* 58:35–54.

Doolittle, W. F. and E. Bapteste. 2007. Pattern pluralism and the tree of life hypothesis. *Proceedings of the National Academy of Sciences of the United States of America* 104:2043–2049.

Ebersberger, I., P. Galgoczy, S. Taudien, S. Taenzer, M. Platzer, and A. von Haeseler. 2007. Mapping human genetic ancestry. *Molecular Biology and Evolution* 24:2266–2276.

Ewing, G., I. Ebersberger, H. Schmidt, and A. von Haeseler. 2008. Rooted triple consensus and anomalous gene trees. *BMC Evolutionary Biology* 8:118.

Frazer, K. A., E. Eskin, H. M. Kang, M. A. Bogue, D. A. Hinds, E. J. Beilharz, R. V. Gupta, J. Montgomery, M. M. Morenzoni, G. B. Nilsen, C. L. Pethiyagoda, L. L. Stuve, F. M. Johnson, M. J. Daly, C. M. Wade, and D. R. Cox. 2007. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* 448:1050–1053.

Galtier, N. and V. Daubin. 2008. Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 363:4023–4029.

Grassly, N. and E. Holmes. 1997. A likelihood method for the detection of selection and recombination using nucleotide sequences. *Molecular Biology and Evolution* 14:239–247.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22:160–174.

Hein, J. 1993. A heuristic method to reconstruct the history of sequences subject to recombination. *Journal of Molecular Evolution* 36:396–406.

Hein, J., M. H. Schierup, and C. Wiuf. 2005. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. New York: Oxford University Press.

Holland, B. R., S. Benthin, P. Lockhart, V. Moulton, and K. T. Huber. 2008. Using supernetworks to distinguish hybridization from lineage-sorting. *BMC Evolutionary Biology* 8:202.

Husmeier, D. 2005. Discriminating between rate heterogeneity and interspecific recombination in DNA sequence alignments with phylogenetic factorial hidden Markov models. *Bioinformatics* 21:ii166–ii172.

Husmeier, D. and G. McGuire. 2003. Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. *Molecular Biology and Evolution* 20:315–337.

Huson, D. H. and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23:254–267.

Jin, G., L. Nakhleh, S. Snir, and T. Tuller. 2006. Maximum likelihood of phylogenetic networks. *Bioinformatics* 22:2604–2611.

Kosakovsky Pond, S. L., D. Posada, M. B. Gravenor, C. H. Woelk, and S. D. W. Frost. 2006. Automated phylogenetic detection of recombination using a genetic algorithm. *Molecular Biology and Evolution* 23:1891–1901.

Kubatko, L. S., B. C. Carstens, and L. L. Knowles. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* 25:971–973.

Larget, B. 2008. Bayesian Untangling of Concordance Knots (BUCKy). http://www.stat.wisc.edu/~ane/bucky/ (accessed May 28, 2010).

Liu, L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics* 24: 2542–2543.

Liu, L., D. K. Pearl, R. T. Brumfield, and S. V. Edwards. 2008. Estimating species trees using multiple-allele DNA sequence data. *Evolution* 62:2080–2091.

Maddison, W. P. 1997. Gene trees in species trees. *Systematic Biology* 46:523–536.

Maddison, W. P. and L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology* 55:21.

Maydt, J. and T. Lengauer. 2006. Recco: recombination analysis using cost optimization. *Bioinformatics* 22: 1064–1071.

McGuire, G. and F. Wright. 2000. TOPAL 2.0: improved detection of mosaic sequences within multiple alignments. *Bioinformatics* 16:130–134.

Meng, C. and L. S. Kubatko. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theoretical Population Biology* 75:35–45.

Minin, V. N., K. S. Dorman, F. Fang, and M. A. Suchard. 2005. Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics* 21: 3034–3042.

Moret, B. M. E., L. Nakhleh, T. Warnow, C. R. Linder, A. Tholse, A. Padolina, J. Sun, and R. Timme. 2004. Phylogenetic networks: modeling, reconstructibility, and accuracy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 1:13–23.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562.

Munshaw, S. and T. B. Kepler. 2008. An information-theoretic method for the treatment of plural ancestry in phylogenetics. *Molecular Biology and Evolution* 25:1199–1208.

Oliver, J. C. 2008. AUGIST: inferring species trees while accommodating gene tree uncertainty. *Bioinformatics* 24:2932–2933.

Page, R. D. 1998. GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics* 14:819–820.

Pagel, M. and A. Meade. 2008. Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 363:3955–3964.

Posada, D. and K. A. Crandall. 2002. The effect of recombination on the accuracy of phylogeny estimation. *Journal of Molecular Evolution* 54:396–402.

Posada, D. and K. A. Crandall. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proceedings of the National Academy of Sciences of the United States of America* 98: 13757–13762.

Posada, D., K. A. Crandall, and E. C. Holmes. 2002. Recombination in evolutionary genomics. *Annual Review of Genetics* 36:75–97.

Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428:493–521.

Rodriguez, F., F. Wu, C. Ané, S. D. Tanksley, and D. M. Spooner. 2009. Do potatoes and tomatoes have a single evolutionary history, and what proportion of the genome supports this history? *BMC Evolutionary Biology* 9: 191.

Ronquist, F. and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.

Ruths, D. and L. Nakhleh. 2006. RECOMP: a parsimony-based method for detecting recombination. *Proceedings of the 4th Asia Pacific Bioinformatics Conference*. London: Imperial College Press, pp. 59–68.

Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6:461–464.

Suchard, M. A., R. E. Weiss, K. S. Dorman, and J. S. Sinsheimer. 2003. Inferring spatial phylogenetic variation along nucleotide sequences: a multiple changepoint model. *Journal of the American Statistical Association* 98:427–437.

Than, C., D. Ruths, and L. Nakhleh. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* 9:322.

Tuffley, C. and M. Steel. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology* 59:581–607.

Webb, A., J. M. Hancock, and C. C. Holmes. 2009. Phylogenetic inference under recombination using Bayesian stochastic topology selection. *Bioinformatics* 25:197–203.

Wendel, J. F. and J. J. Doyle. 1998. Phylogenetic incongruence: window into genome history and molecular evolution. In P. Soltis, D. Soltis, and J. J. Doyle, eds. *Molecular Systematics of Plants II*. New York: Springer, pp. 265–296.

Whelan, S. 2008. Spatial and temporal heterogeneity in nucleotide sequence evolution. *Molecular Biology and Evolution* 25:1683–1694.

White, M. A., C. Ané, C. N. Dewey, B. Larget, and B. Payseur. 2009. Fine scale phylogenetic discordance across the house mouse genome. *PLoS Genetics* 5(11):e1000729.

Wu, F., L. A. Mueller, D. Crouzillat, V. Petiard, and S. D. Tanksley. 2006. Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the Euasterid plant clade. *Genetics* 174:1407–1420.

Yang, H., T. A. Bell, G. A. Churchill, and F. Pardo-Manuel de Villena. 2007. On the subspecific origin of the laboratory mouse. *Nature Genetics* 39:1100–1107.

Zhou, Y., N. Rodrigue, N. Lartillot, and H. Philippe. 2007. Evaluation of the models handling heterotachy in phylogenetic inference. *BMC Evolutionary Biology* 7:206.

CHAPTER 4

PROBABI
TOPOLO
SAMPLIN

4.1 INTRODUCTI

Phylogenetic stu
whereas populat
tion (intraspecif
promises to hel
1994).

Traditiona
nucleotides, am
tree is estimated
sequence is inte
variation within
typically has a l
exists between s
the genetic vari
variation that e
such as that occ
(Maddison and

Previousl
the multispecie
is one gene sam
Rosenberg 200
(Rosenberg 20
lineage sample
tree inference f
per species (De
al. 2009), parti
different specie
(Kubatko et al.
1989). The tre
across multiple
2010) or Globa
consistent esti

Estimating Species
Laura S. Kubatko