

Testing for Treeness: Lateral Gene Transfer, Phylogenetic Inference, and Model Selection

Joel D. Velasco and Elliott Sober

Introduction – Smoking and Non-smoking Guns: Sometimes it may seem obvious that a lateral gene transfer (LGT) event has occurred. For example, consider the fact that a multi-gene phylogenetic analysis of Rafflesiaceae, a family of flowering plants that includes *Rafflesia arnoldii*, which has the largest flower in the world, places it within the order Malpighiales, a diverse order containing such plants as willow trees and poinsettias. However, analysis of its Mitochondrial gene *nad1B-C* places it within the grape family Vitaceae in the order Vitales, a distantly related taxon (Davis and Wurdack 2004). This discordance is easily explained when we note that Rafflesiaceae is an endophytic parasite and the *nad1B-C* gene analysis places it near its host *Tetrastigma*. The obvious conclusion to draw is that there was an LGT event transferring mtDNA from host to parasite.

Other cases are not so easy. In 1939, Kubo first identified hemoglobin genes in soybean plants (Kubo 1939). At the time, and for the next 40 years, the only other remotely similar genes known were hemoglobin and myoglobin genes found in vertebrates and a few invertebrates. So how did this happen? There are only three possible stories.

In the first, the hemoglobin gene family arose at least twice independently and the leghemoglobins (those in the legumes) and animal hemoglobin genes are not actually homologues. Call this the convergent evolution story. It may seem obvious that this story should not be taken seriously – after all, the probability against convergent evolution on this

scale is staggering. It may seem that hemoglobin genes are so complex that the similarities between the genes in the two groups couldn't be a coincidence. However, the right reply is that however improbable, the coincidence is not literally impossible; that is, the characters could be similar – even identical – just because each evolved independently of each other in exactly the same way. To dismiss this possibility we need to consider the probability of the similarity under other hypotheses.

If convergent evolution did not happen in this case, then the hemoglobin family arose only once in evolutionary history. One way this is possible is that it arose in the distant past prior to the common ancestor of plants and animals. Then the vast majority of lineages lost their hemoglobin genes – or perhaps they simply mutated to the point where there no longer are recognizable copies in the other genomes. With two closely related taxa, it might make sense to explain their similarity by postulating losses in adjacent branches. However, in very distantly related taxa, the loss hypothesis becomes far less plausible, because so many independent losses would be required. But again, this scenario is improbable, not impossible.

The third possibility is LGT; the trait originated just once and subsequently passed from ancestors to descendants, but it also jumped across these vertical branches, passing from a donor to a recipient who was not a descendant of the donor. One way this might happen is for a virus to serve as a vector, by taking a bite out of a donor genome and then inserting it into the genome of a recipient.

Given that there are only three possibilities and that in some cases, two of them are extraordinarily improbable, it may seem obvious that LGT is the only plausible alternative left standing. This is why LGT is a live hypothesis and possible transfer scenarios were invented and investigated in the hope of better understanding the history of hemoglobin (Jeffreys 1982). But starting in the 1980s, hemoglobin genes began to be found in non-legume plants (Appleby et al. 1983). Today, we know of hundreds of hemoglobin genes in bacteria, archaea, plants, fungi, and animals (Vinogradov et al. 2006). This makes the single ancestry story with gene loss far more plausible than it was before since there are more positive instances and fewer gene losses to explain.

While the evidence does not seem all that strong in favor of the LGT explanation today, the more important point is that there will be cases that are not obvious. In the hemoglobin example, our evidence concerning the distribution of the gene across various taxa changed. But even fixing that, we are left with some uncertainty. In some of these cases, the true explanation will be far from clear. For example, the trait might not be that complex – perhaps it is a chunk of DNA twenty base pairs long that exists only on two branches that are separated by three branches in between. We could explain this as an instance of parallel evolution that is improbable, but not outrageously so. Similarly, it could be a case of gene loss – improbable, but not extraordinarily so. Surely we should examine this situation by considering all three possible explanations.

The general point is clear – as with any question about evolutionary history, the key is to look at the evidence and see where it leads. Sometimes there will be clear cases, but we want to go

beyond the clear cases and have a general principle that allows us to assess the evidence for (or against) a historical event and to say how strong or weak it is.

Terminology: We use “tree” to refer to a rooted graph in which branches split but never join as one moves from past to present; each node has at most one parent and at most two descendants. By “network” we mean a tree with reticulations added. In a network, branches join as well as split. In a tree, branches always connect a descendant back to one of its ancestors. These are vertical branches. In a network, some branches are like this, but others connect a node to a nonancestor. These are lateral branches. We can formally define a phylogenetic network as a rooted, directed, acyclic graph, leaf-labeled by a set of taxa, coupled with a set of temporal constraints to ensure that the lateral branches connect nodes that could exist at the same time (Moret et al., 2004).

Parsimony and LGT: Maximum parsimony (MP) is one of the most popular methods for phylogenetic tree reconstruction. According to MP, the best phylogenetic tree is the one that minimizes the number of changes required along the branches. Philosophical justifications for MP have varied, but it has often been justified by appeal to some general methodological principle said to be related to parsimony such as explanatory power or falsifiability (Wiley 1981, Farris 1983, Kluge 2005). Here, a parsimonious tree is better because it requires fewer changes and this is said to increase the explanatory power of the tree.

A natural first step to developing methods for inferring phylogenetic networks is to try to generalize the method of maximum parsimony to cover networks. Hein (1990, 1993) extended

parsimony to cover the case of recombination and Nakhleh et al. (2005) further generalized this and developed computer algorithms for implementing the methods. The general idea is still the same – the parsimony score of a network is the number of changes required along the vertical branches. However, the use of parsimony for phylogenetic networks has a serious problem. If phylogenetic hypotheses are evaluated only by counting the number of homoplasies they require, which is ordinarily equivalent to the total number of changes, then the most parsimonious tree, if it requires any homoplasies at all, can be bettered by adding LGT events, thus reducing the number of homoplasies to zero. The same can be done for all the other possible trees. The result isn't a single phylogenetic hypothesis, but a set of these, each requiring zero homoplasies. The most complicated network with lateral branches connecting every vertical branch to every other will always be among the hypotheses tied for first place. But it is natural to think that networks with more lateral branches are more complex and less “parsimonious” than networks with fewer. This is why the idea that we should minimize lateral branches has also been linked to Ockham's Razor (Nakhleh et al. 2003, Than et al. 2008). General parsimony principles seem to justify minimizing evolutionary changes along vertical branches and minimizing lateral transfer events, but these two goals are in direct conflict.

One possible response is to find the best tree for each gene and then construct the smallest network that contains all of these best trees. This assumes that homoplasies and lateral transfers both reduce the plausibility of a genealogy, but that the former does so far more than the latter. This is an assumption that requires justification.

A second solution is to set a bound for how many lateral events can be postulated and then find the most parsimonious network among those with at most that many lateral branches (Nakhleh et al. 2005, Jin et al. 2007). But how should the appropriate bound be chosen? We don't want to set a bound before examining the data. Nakhleh et al. and Jin et al. determine their bounds based on the idea that adding lateral branches to a network is subject to diminishing returns. Suppose we start with the most parsimonious tree and then add lateral branches one by one, in a certain order. First we add the lateral branch that most reduces the number of homoplasies that need to be postulated, then, using our newly formed network as a base, we search again for the branch that will reduce the score the most, and so on. At a certain point, further additions will bring zero improvement, but before then, an addition may be judged to improve parsimony too little to be justified. How might one choose a threshold that determines when an addition is not worth the candle?

Nakhleh, Jin, and colleagues attempt to determine the appropriate threshold empirically by continually adding lateral branches and tracking how much the addition of each branch reduces the parsimony score. Then, by examining the shape of the curve created by plotting the maximum parsimony score achieved under each addition of a lateral branch, they attempt to estimate which additions capture real transfer events and which merely capture noise. The idea is that we know *a priori* that the graph we obtain when we chart the number of branches added versus the reduction in parsimony score will be a diminishing returns graph. This is a trivial consequence of the order in which we introduce additional lateral branches. But what shape will that diminishing returns graph take? Will there be a sudden and large dropoff between two consecutive branches? Or will the graph exhibit a steady reduction in the number of

homoplasies? Their idea is that adding branches which represent real events will lead to a significant reduction, but adding branches that represent noise will not. Therefore, the first time there fails to be a significant change between two consecutive branch additions is where we infer the true cutoff to be.

While this is an attempt to look at the data to infer the number of transfer events rather than assuming a priori what it should be, this method has its drawbacks. Regardless of how many transfer events really occurred, some branches that model transfer events that didn't really happen will be able to reduce the number of required homoplasies significantly while others won't. So there is the danger of postulating fictitious transfer events with this method. In addition, some real transfer events may not substantially reduce the number of homoplasies required. A transfer event between two closely related branches may not reduce the number of required homoplasies very much – in fact, it may not even reduce the number at all. It is unclear how this procedure can control for the number of type-1 and type-2 errors.

Another way to attempt to solve this threshold problem is to simulate data following a network pattern. Here, for some fictitious data set, we know the right answer concerning how many transfer events there were and where and when they occurred (because the example under analysis is something we invented). Then we simply apply different methods to the data and compare how the methods performed. In this case, we can calculate what the appropriate

threshold should be by noting how many times homoplasies should be avoided by postulating lateral branches so that one gets as close as possible to the target of the true network.¹

This can provide some justification for thinking that certain thresholds are better than others, but it still has a significant problem. Each time we simulate the data, we use a specific model of how traits evolve on a network. For example, we might simply build in the fact that on any particular branch in a set time period, there is a 1% chance of a transfer event occurring that copies the state of one gene on a branch to a site on another branch. On this model, a certain picture emerges of what thresholds will be appropriate. But on another model of transfer that treats transfer events as more or less probable or simply models them in a different way (perhaps one in which there is a correlation between transfer events), it may turn out that a different threshold is better. However, recall that the goal is to use simulations to justify a *single* threshold for a given data set, where the data set and the phylogeny are not examples we invented but are from nature. The problem here is typical of all simulation studies. Simulations simply assume a model for the process generating the data and there is no assurance as to how commonly nature follows the chosen model.

In addition to the problem of finding an appropriate threshold, the particular algorithms developed by Nakhleh, Jin, and colleagues have the problem that they might never even consider the optimal network. This is because they are “greedy” algorithms – that is, they attempt to find the global optimum by repeatedly making locally optimal changes. In general, there is no reason

¹ It may not be possible to recover the true network. The data might be structured so that a false lateral branch will reduce the parsimony score more than one or more of the real lateral branches. In this case, no matter what the threshold, parsimony cannot recover the true network.

to think that the best network with $n+1$ lateral branches can be reached from the best network with n branches. Nakhleh et al. (2005) do use a few empirical checks such as adding branches in a different order and in the cases they looked at, this made no difference. But this will not always work. It is easy to see that this algorithm is not guaranteed to find the best network. The most obvious problem for this approach is that this is a fixed tree method – that is, they start with a single tree and never adjust this tree as they add branches. Contrary to the claim of Jin et al. (2007), in most cases the underlying tree is unknown. This is particularly true in the cases we are considering where there is a significant chance of LGT obscuring the underlying vertical relationships. The best supported tree assuming that there is no transfer at all may not be the vertical component of the best overall network. Rather than using a tree-method to infer the underlying tree and then a network method to infer the lateral events, it would be better to use a single method to infer the total history, including both its vertical and lateral aspects.

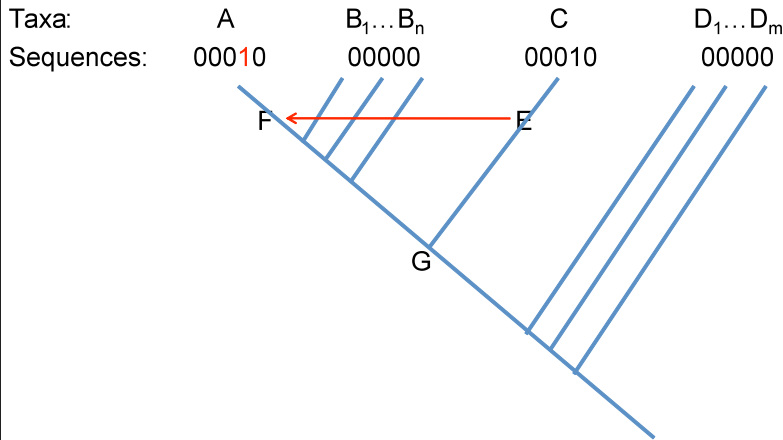
We should not be too critical, though. Nakhleh et al. do define the parsimony score of a network independent of what the particular underlying tree is and it is only for computational convenience that they use a fixed tree method. Switching to a less computationally tractable method such as a simple exhaustive search of all possible networks would solve the problems associated with the greediness of their particular algorithm, but it is important to realize that this would not solve their threshold problems.

The methods developed by Nakhleh, Jin, and colleagues to develop appropriate thresholds are problematic. In fact, the authors themselves later called these methods “*ad hoc*” (Park et al. submitted). Park et al. (submitted) utilizes the same methods and so also faces the problem of

finding appropriate thresholds but in addition to examining the slope of the diminishing returns curve, they also compare individual lateral branches to randomly added branches and determine p-values for each branch and use a bootstrapping procedure to attempt to check the reliability of particular lateral branches. While this is a step forward, given the problems just described for deciding how the penalty for homoplasy compares with the penalty for postulating LGT events, we think another approach merits exploration, one in which model selection ideas are used.

A simple example: We hope that some of the relevant features of conceptualizing how LGT should be brought into the framework of phylogenetic inference are captured by the simple problem depicted in the accompanying figure. The tip taxa and their aligned sequences of binary characters are as shown. The tree given in the figure (which might be very well supported due to characters not shown) treats the 1 at the 4th site in A and C as a homoplasy. If an LGT event is postulated between E and F, two new nodes inserted on the vertical branches leading to A and to C, no homoplasy is needed.

Figure: Should a strict tree phylogeny be supplemented by introducing an LGT event linking E to F to explain the 1 at the 4th site in A and C?



We wish to compare two hypotheses: TREE (which declines to postulate any LGT events) and NET (which says that an event of this sort that links E to F is *possible*).

The first thing to do when thinking within a model selection framework is to conceive of models as propositions that contain adjustable parameters whose values can be estimated by finding their maximum likelihood values. Maximum likelihood methods for phylogenetic networks with lateral transfer in mind were first developed by Jin et al. (2006). Here we describe some of the foundational reasoning involved in developing such methods in a way that makes clear the variety of models possible.

Part of our formula for scoring a network will be its likelihood. That is, we have to know $\Pr_M(\text{Data}|\text{Network})$ – the probability of the tip data given the network on some particular network model. Rather than consider the sequences that attach to all the tip taxa, we will simplify the discussion by considering just the sequences attaching to A and C. These are the data. The same principles can then be extended to calculate the likelihood of the entire network. As is customary, we take TREE to assert that the characteristics of A and C are independent of each other, conditional on the sequence attaching to their most recent common ancestor G. NET denies this. In particular, NET asserts that there are *two* possible influences on the state of F; F's sequence can be affected by the sequence found in its most recent common ancestor G and also by the sequence found in its non-ancestor E.

One option is to assume that an LGT event linking E to F would *guarantee* that F=1 at the 4th site in the sequence. But this is unnecessarily restrictive; the lateral branches that we need to think about in connection with LGT should not be thought of in this way. To see this, consider what it means to draw a vertical branch from an ancestor to a descendant. This does not entail that the character states of the ancestor will, with probability 1, also be found in the descendant. The vertical branch leaves open what probability model we should apply to character evolution in that branch. And there are many choices. We should conceptualize LGT in the same way. When we draw a lateral branch from E to F, this leaves open what the probability is that F=1 at the fourth site in the sequence, given that E=1 at that site. We need to consider probability models of LGT, and there are many.

We therefore will distinguish LGT *branches* from LGT *events*. The NET hypothesis postulates an LGT branch linking E to F. Whether all, or any, or none of the character states of F are due to LGT events coming from E is a further question. TREE is a special case of NET, in that TREE says that all such events have a probability of zero. NET has adjustable parameters that describe what might happen on the LGT branch it postulates. Since TREE is a special case of NET, if we evaluate NET and TREE only by their likelihoods, then we are in the same pickle, described above, into which parsimony also lands; TREE can't have a higher likelihood than NET, when each is fitted to the data. Incidentally, distance-based methods such as extensions of weighted-least-squares methods face the same overfitting problem (Markarenkov and Legendre 2004). In the case of parsimony and distance methods, this means using an *ad hoc* stopping rule that dictates when lateral branches should be added. In the case of likelihood methods, as we shall see, there is a well-grounded theory that solves the problem of overfitting.

With respect to the ancestor/descendant branches postulated by TREE, there is a familiar set of options for modeling character evolution. They arise from the different answers one might give to these three questions (Sober 2008):

- Must the different changes that can occur at a given site in a given branch have the same probability?
- Must a given kind of change have the same probability at all sites on a branch?
- Must a given kind of change at a given site have the same probability on different branches?

The Jukes/Cantor model says yes to all three questions (Jukes and Cantor 1969). Other, more complex models say no to some. For example, the Tuffley/Steel no-common-mechanism model says yes to the first, but no to the second and third (Tuffley and Steel 1997).

How should we think about the branch that NET introduces, which links F to its non-ancestor E?

There are various options to consider. Models for lateral branches will have different adjustable parameters from the ones that are possible for vertical branches. In addition to questions about lateral branches that are parallel to those described above concerning vertical branches such as whether to treat different lateral branches identically, different models for LGT are possible, depending in part on how the following questions are answered

- If there is a lateral branch from one vertical branch to another, when does that lateral branch occur?
- If a virus vector draws a sample from the donor's genome on a given branch at some time, some of which then gets transferred to a recipient, how large a sample will the vector deposit?

We think of LGT events as transferring material from donor to recipient virtually instantaneously, so lateral branches, unlike vertical branches, have no temporal duration. And the idea that the material transferred from host to recipient by LGT is a sample from the host's whole genome is very unlike the processes that are usually discussed for vertical branches.

Whatever the details are that go into a given model for LGT, it is important to see that any network model must allow that there are nodes whose states might be influenced by its ancestor and also by a non-ancestor. In our figure, the state of F at a site might be due to the state of E, or

it might be due to the state of G. How might these two possible sources be brought together in a single model? Just as an example, let's consider a model that lumps all the sites on a branch together. We'll use a Jukes/Cantor model for the vertical branch from G to F and a model for the lateral branch E to F that says that all sites in E have the same probability of being transferred to F. We'll call this probability θ . The model of these two possible influences on the state of F is additive, as shown in the following table.

An additive model for NET. Cell entries represent probabilities of the form $\Pr(F=1 \mid M=i \ \& \ E=j)$.		
	E=1	E=0
M=1	$\theta + (1-\theta)P_{11}$	$(1-\theta) P_{11}$
M=0	$\theta + (1-\theta)P_{01}$	$(1-\theta) P_{01}$

The values in the table follow directly from the idea that there are two possible sources for the character state in F and they are mutually exclusive. To calculate the probability that F will be in state 1, first imagine that at both M and E, the character is in state 1. Then, with probability θ , the state at F is a result of a transfer event in which case F will definitely be in state 1 since that is the state at E. If there is no transfer event (this failure having a probability of $1-\theta$) the probability that the character will be in state 1 at F is just P_{11} , which is the probability that the character would go from state 1 at G to state 1 at F by ordinary vertical transmission. Since these are mutually exclusive and exhaustive possibilities, we simply sum their individual probabilities to get the total probability that F is in state 1.

Now suppose that the character is in state 0 at E. In this case, there is only one way to have the character in state 1 at F, since if there were a transfer event, the character at F would be in the


same state it is in at E, namely state 0. So the only way to get $F=1$ when $E=0$ is by way of vertical transmission, and the probability of this is $(1-\theta)P_{11}$. We hope this makes clear why cells in the first column of the Figure have two addends, while those in the second have only one.

The additivity of this model is not an artifact of our using a simple model like Jukes/Cantor. It is simply a consequence of the fact that the state of F has two possible sources; it must come from a lateral transfer event or it must come directly from its ancestor G, and it can't come from both. Notice that what TREE says about the character states of F can be obtained from this model for NET by setting $\theta = 0$.

It is easy to describe a more complex model that is also additive. For example, one might use the Kimura 2-parameter model (Kimura 1980) for vertical branches, in which different kinds of changes (transitions and transversions) are assigned different parameters. And one might contemplate a model for lateral branches in which the relevance of adjacency in a site is recognized (just as they are in models of recombination); sites that are close to each other in a sequence have higher probabilities of being transferred together, whereas more distant sites are more independent. And just as there are models for vertical branches that allow for different branches having different rates of evolution, so there can be models for lateral branches that do the same thing.

If TREE can be equipped with various process models, and the same is true of NET, how shall we proceed? One idea would be to try to isolate the most plausible model for each and then compare these two best cases. We envisage a different procedure. We should consider a variety

of different models for NET. For each of these NET models, we can obtain a model for TREE by setting various parameters equal to zero. Different trees can be obtained from different networks by stripping away all lateral branches. We then should apply a model selection criterion to evaluate these models.

		NET	TREE
More complex  Simpler	Process Models	N3	T3
		N2	T2
		N1	T1

When we look across a row, the TREE model will be simpler than the NET model, in the sense of containing fewer adjustable parameters. On the other hand, the NET model in a row will fit the data better. How do we weight these two considerations? One popular approach to model selection is the Akaike Information Criterion (AIC) (Akaike 1973, 1974). The AIC score of a model is $2k - 2 \ln \{\Pr[\text{data}|L(M)]\}$ where k refers to the number of adjustable parameters in the model and $L(M)$ refers to the likeliest member of the model M . The AIC score of a model is an unbiased estimate of the expected log likelihood of a model with respect to the underlying process that generated the data; this is the content of Akaike's (1973, 1974) theorem. In contrast, the maximum likelihood of a model is a biased estimate of this average – the more parameters in the model, the worse the bias. Akaike showed that we can correct this bias in an exact way,² by imposing a penalty on a model for its complexity. Forster and Sober (1994) describe AIC as

²To derive the AIC, a few background assumptions are needed. For example, there are certain regularity conditions that have to hold for the likelihood function to be asymptotically normal and there has to be enough data to ensure that the likelihood function will approximate its asymptotic properties. See Forster and Sober (1994) and Burnham and Anderson (2002) for more details.

aiming to estimate a model's predictive accuracy; it provides an estimate of how well the model will predict new data when it is fitted to old.

AIC is already widely used in phylogenetic inference partly due to its inclusion in the popular software Modeltest (Posada and Crandall 1998, Posada and Buckley 2004). AIC can tell us which of the models in a given row is better. We also can use AIC to make comparisons between NET and TREE models in different rows. It is true that a sufficiently complex NET model can fit the data perfectly. But that does not mean that it is the best model. AIC permits a comparison among all these models. It is in this respect superior to the likelihood ratio test, which can be used to compare nested models only. A likelihood ratio test can be applied to models in the same row, and to some of the items in the same column, but not to other pairs.

Cladistic parsimony versus model selection parsimony. Cladistic parsimony scores phylogenies by counting homoplasies. Another sort of minimization criterion would be to score phylogenies by counting homoplasies and lateral branches as well, with some weight assigned to how much the one matters compared with the other. Model selection involves parsimony considerations, but one is not counting homoplasies or lateral branches at all. Rather, what one counts are parameters. And furthermore, parsimony, in this sense, is only one consideration, not the whole show, according to model selection criteria. Fit-to-data (i.e., likelihood) matters as well, with AIC describing the appropriate trade-off between parsimony and fit-to-data.

Cladistic parsimony counts the homoplasies required by a phylogenetic tree whereas model selection parsimony counts the parameters in a model. The first of these has been investigated

from the point of view of likelihood with the goal being to determine under what circumstances the parsimony ordering of a set of tree hypotheses coincides with their ordering in terms of likelihood. Felsenstein (1973) and Tuffley and Steel (1997) have each identified assumptions about the evolutionary process that guarantee that cladistic parsimony and likelihood will be ordinally equivalent, and it is now widely recognized that there also are assumptions in which the two orderings will disagree. See Sober (2004, 2008) for discussion. Given that cladistic parsimony sometimes can be viewed as a reflection of likelihood, it is interesting that Akaike's theorem establishes that both likelihood and number of adjustable parameters are relevant to estimating the predictive accuracy of a model; in this sense AIC can be seen as a procedure that takes account of both kinds of parsimony – that which reflects likelihood and that which reflects the number of adjustable parameters.

Other sources of reticulation: Our approach to LGT also applies to any other cases of reticulation such as recombination and the problem of inferring hybridization events. The genetic material that is transferred in an LGT event would not be passed on without recombination and methods that are useful for detecting one can be used for detecting the other (Chan et al. 2009). There are more significant physical differences between LGT and hybridization. In the former, there is a donor and a recipient; in the latter, there is a melding of two objects to form a third (as in sexual reproduction). But from a mathematical point of view, these differences disappear. In a tree (in the sense we defined), every offspring node has just one parent (its most recent ancestral node). To introduce the possibility of hybridization into a tree model, you allow that some offspring may have two parents. This complicates the model,

but allows it to fit the data better. The point of contact between hybridization and LGT is that TREE is a special case of HYBRID, just as TREE is a special case of NET.

While our approach is designed to accommodate actual reticulate evolution, it will also infer reticulation when faced with a variety of processes such as certain combinations of gene duplication and loss and lineage sorting. Distinguishing among these different underlying processes is a difficult task that we do not attempt to address here. What we hope to have provided in this paper is a useful way of looking at a certain class of problems in phylogenetic inference. By thinking about different possible models of lateral gene transfer and utilizing tools from model selection theory such as AIC, phylogeneticists can recover reticulation in a more theoretically grounded way and further, can recover the underlying vertical history at the same time.

Acknowledgments: We thank David Baum, Rob Beiko, Ehud Lamm, Bret Larget, and Mike Steel for helpful discussion.

References:

- Akaike H (1973) Information theory and an extension of the maximum likelihood principle. Pages 267–281 in Second International Symposium on Information Theory. Akademiai Kiado, Budapest.
- Akaike H (1974) A new look at the statistical model identification. *IEEE Trans. Aut. Control* 19:716–723.
- Appleby CA, Tjepkema, JD, Trinick MJ (1983) Hemoglobin in a nonleguminous plant *Parasponia*: possible genetic origin and function in nitrogen fixation. *Science* 220, 951–953.
- Burnham KP, Anderson DR (2002) Model selection and multimodel inference: A practical and information-theoretic approach (2nd edition). Springer-Verlag, New York
- Chan CX, Darling AE, Beiko RG, Ragan MA (2009) Are protein domains modules of lateral genetic transfer? *PLoS ONE* 4(2):e4524
- Davis CC, Wurdack KJ (2004) Host-to-Parasite Gene Transfer in Flowering Plants:

- Phylogenetic Evidence from Malpighiales. *Science* 305:676
- Farris JS (1983) The logical basis of phylogenetic analysis. In: Platnick, N.I., Funk, V.A. (Eds.), *Advances in Cladistics II*. Columbia University Press, New York, pp. 7–36. Reprinted in E. Sober (ed.), *Conceptual Issues in Evolutionary Biology*, Cambridge: MIT Press, 1994 pp. 333-362.
- Forster M and Sober E (1994) How to tell when simpler, more unified, or less *ad hoc* theories will provide more accurate predictions. *British Journal for the Philosophy of Science* 45: 1-35.
- Hein J (1990) Reconstructing evolution of sequences subject to recombination using parsimony. *Mathematical Biosciences*, 98:185–200
- Hein J (1993) A heuristic method to reconstruct the history of sequences subject to recombination. *J. Mol. Evol.*, 36:396–405
- Jeffreys AJ (1982) Evolution of globin genes. In: Dover, G.A., Flavell, R.B. (Eds.), *Genome Evolution*. Academic Press, New York, pp. 157–176.
- Jin G, Nakhleh L, Snir S, Tuller T (2006) Maximum Likelihood of Phylogenetic Networks. *Bioinformatics* 22(21):2604-2611.
- Jin G, Nakhleh L, Snir S, Tuller, T (2007) Inferring phylogenetic networks by the maximum parsimony criterion: A case study. *Molecular Biology and Evolution*, 24(1), 324-337.
- Jukes TH, Cantor CR (1969) Evolution of protein molecules. *Mammalian-Protein Metabolism*. Academic Press., New York, pp. 21–132.
- Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16, 111–120.
- Kluge AG (2005) What is the rationale for ‘Ockham’s Razor’ (a.k.a. Parsimony) in phylogenetic inference?. In: Albert, V. (Ed.), *Parsimony, Phylogeny, and Genomics*. Oxford University Press, Oxford, pp. 15–42.
- Kubo H (1939) Über hämoprotein aus den wurzelknöllchen von leguminosen. *Acta Phytochim. (Tokyo)* 11, 195–200.
- Markarenkov V and Legendre P (2004) From a phylogenetic tree to a reticulated network. *Journal of Computational Biology* 11(1): 195-212.
- Moret, BME, Nakhleh L, Warnow, T, Linder, CR, Tholse A, Padolina A, Sun J, Timme R (2004) Phylogenetic networks: Modeling, reconstructibility, and accuracy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1), 13-23.
- Nakhleh L, Sun J, Warnow T, Linder CR, Moret BME, Tholse A (2003) Towards the development of computational tools for evaluating phylogenetic network reconstruction methods. In *Proceedings of the PSB03*. Kauai, Hawaii.
- Nakhleh L, Jin G, Zhao F, Mellor-Crummey J (2005) Reconstructing phylogenetic networks using maximum parsimony. In: Markstein V, editor. *Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference (CSB2005)*; August. p. 93–102.
- Park HJ, Jin G, Nakhleh L (2009) On the Significance of Phylogenetic Networks Inferred by Maximum Parsimony (submitted)
- Posada D, Crandall KA (1998) Modeltest: testing the model of DNA substitution. *Bioinformatics* 14 (9): 817-818.
- Posada D, Buckley TR (2004) Model selection and model averaging in phylogenetics: advantages of the AIC and Bayesian approaches over likelihood ratio tests. *Systematic Biology* 53: 793-808.

- Sober E (2004) The Contest between Likelihood and Parsimony. *Systematic Biology* 53: 6-16.
- Sober E (2008) *Evidence and Evolution – the Logic Behind the Science*. Cambridge: Cambridge University Press.
- Than C, Ruths D, Nakhleh L (2008) PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* 9:322
- Tuffley C and Steel M (1997) Links between maximum-likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology* 59:581-607.
- Vinogradov SN, Hoogewijs D, Bailly X, Arredondo-Peter R, Gough J, Dewilde S, Moens L Vanfleteren JR, (2006) A phylogenomic profile of globins. *BMC Evol. Biol.* 6, 31–47.
- Wiley E (1981) *Phylogenetics: The theory and practice of phylogenetic systematics*. New York: Wiley-Interscience.