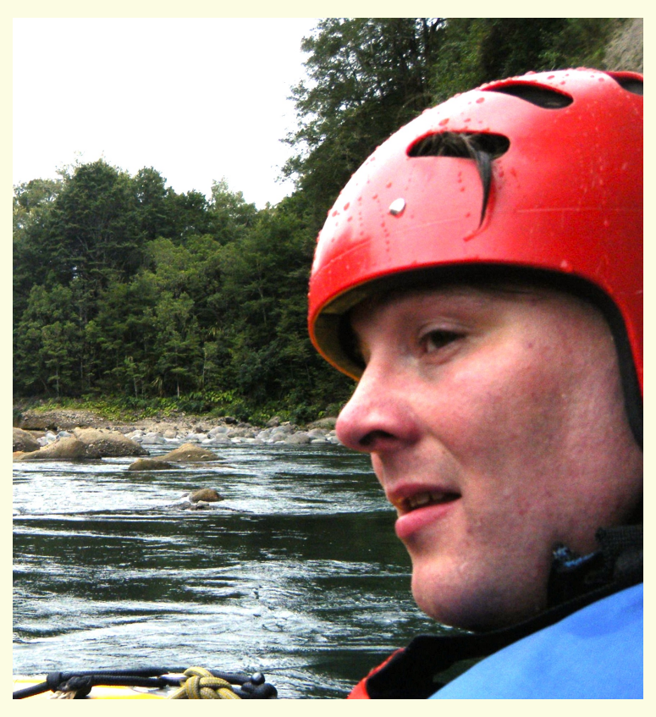
THE UNIVERSITY OF AUCKLAND
FACULTY OF SCIENCE
Department of Mathematics

Phylogenetic inference in the presence of horizontal transfer: Some interesting properties

Steffen Klaere
Computational Evolution Group, University of Auckland

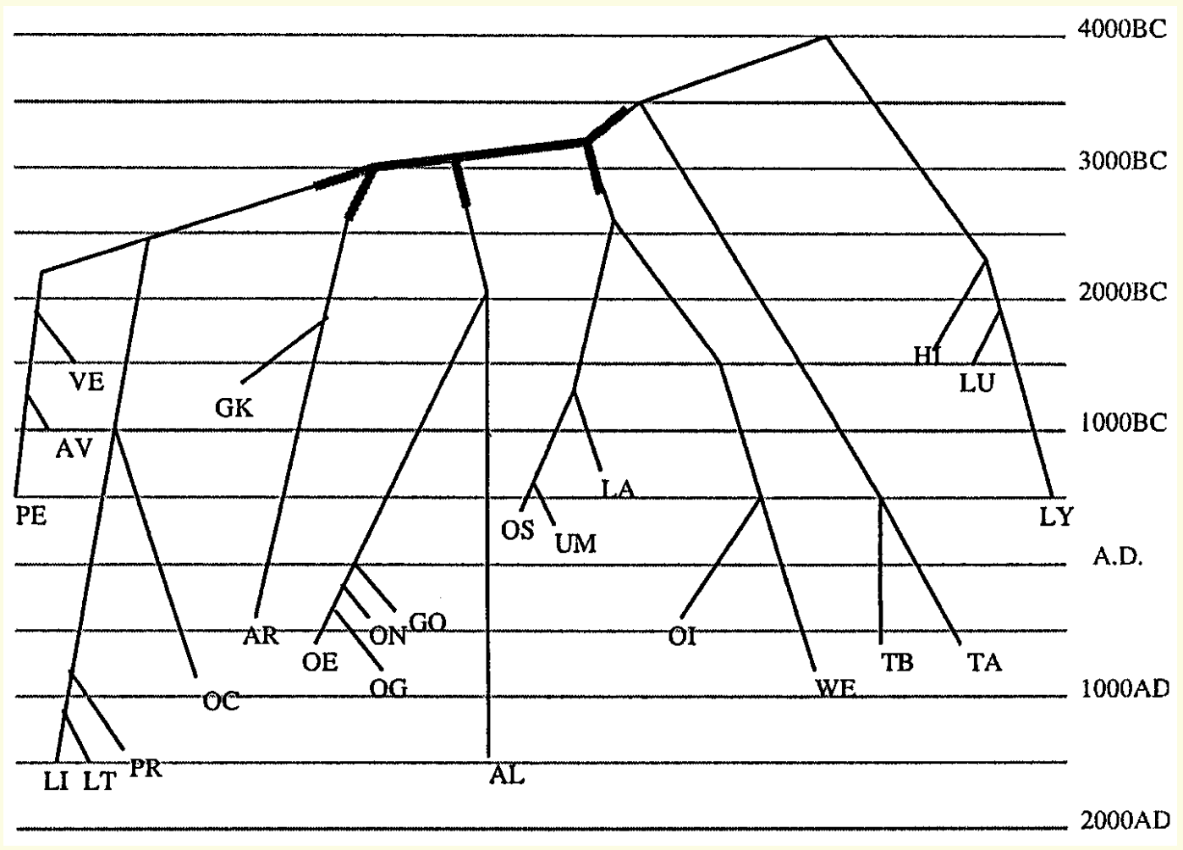


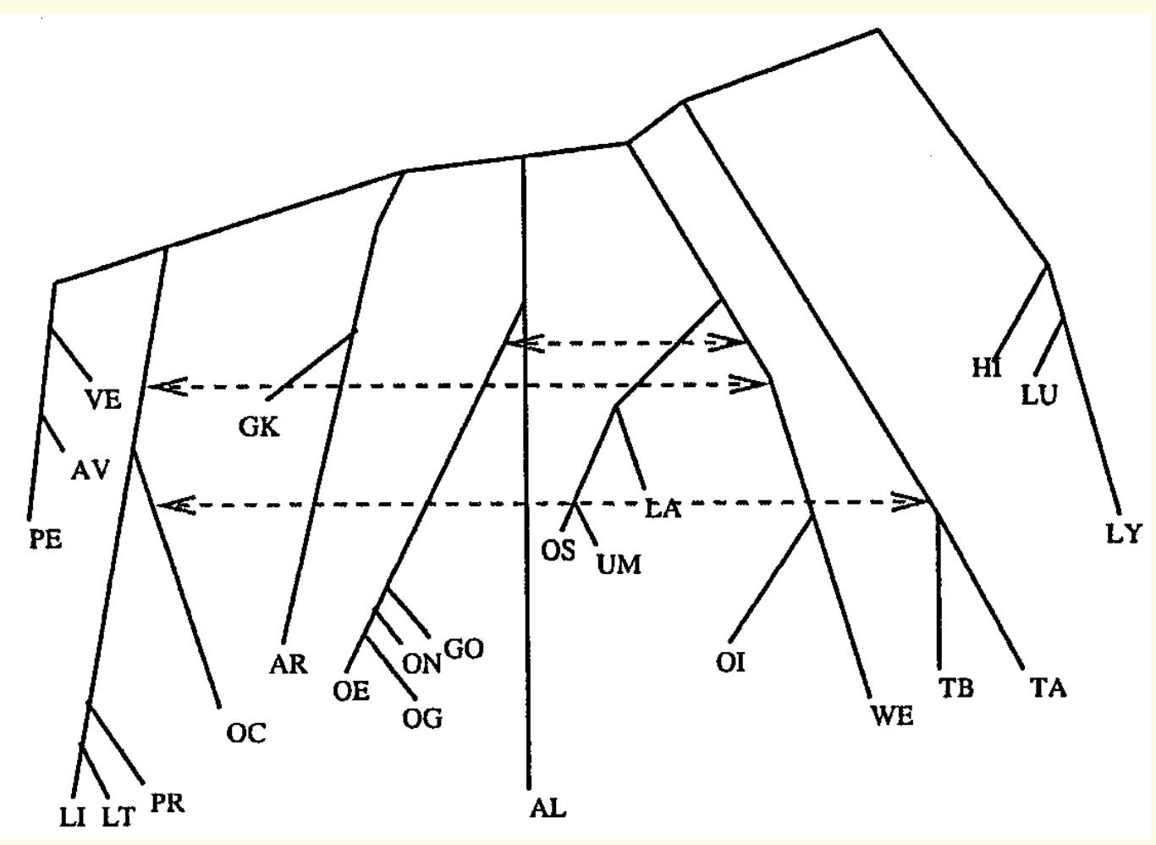
Abstract

The horizontal transfer of genetic material clearly undermines the basic assumptions of phylogenetic inference. Here we explore the extent to which phylogenetic inference might still be carried out even in the presence of 'rampant' lateral gene transfer. Our key question is identifiability: how much horizontal transfer is permissible to still allow the reconstruction of the phylogeny?
We propose a simple model of evolution incorporating both mutation and horizontal transfer. We present various approaches of reconstruction under the model and some identifiability results.
This poster details joint work with David Bryant (University of Otago) and Jamie Kydd (University of Auckland).

Motivation

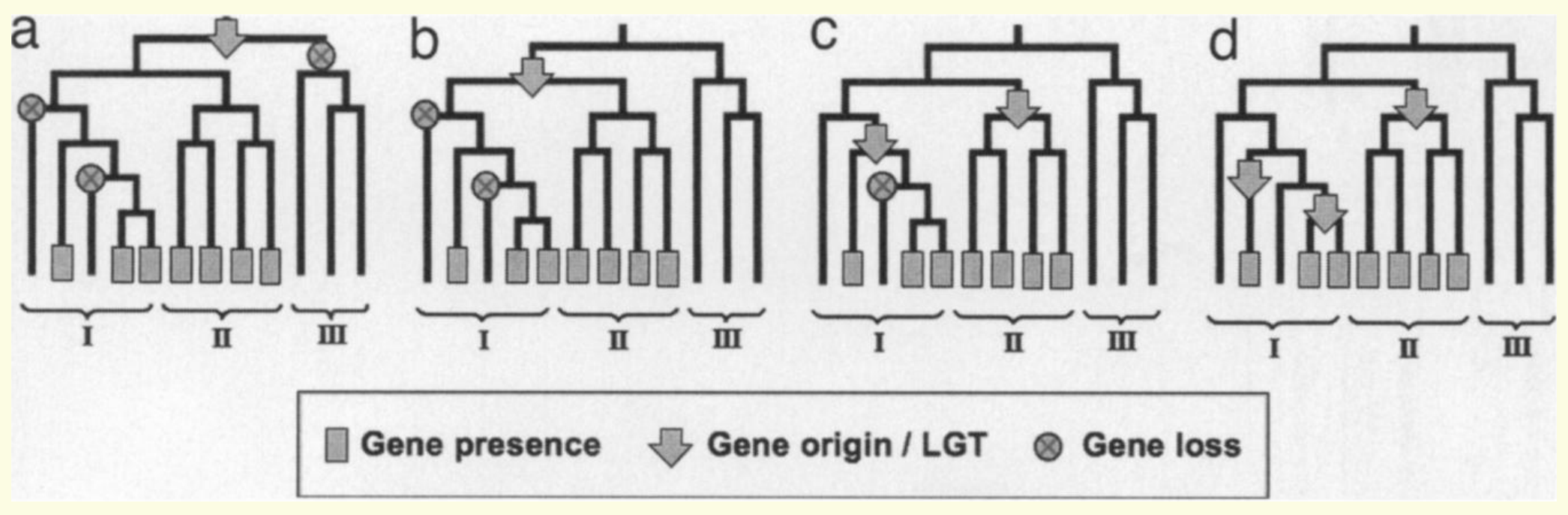
- Usual assumption of stochastic models of phylogeny reconstruction is that changes are solely driven by mutations.
- Under this assumption phylogenies can be reconstructed from sequences (e.g., Steel ,1992).
- Chang (1996) proved that mutation matrices can also be reconstructed.
- Allman et al. (2010) investigated under which conditions phylogeny and model parameters are identifiable from morphological characters.
- Horizontal transfer has been observed in many studies on molecular phylogenies (e.g., Dagan and Martin, 2007) and language evolution (e.g., Nakhleh et al., 2005).
- Horizontal transfer distorts the phylogenetic signal. This ultimately leads to the question of how much transfer is needed such that an underlying phylogeny cannot be reconstructed.
- Tree identifiability under horizontal transfer has not been addressed yet.
- Here we investigate this question for a simple model with transfer between lineages.





Model

- We consider a simple forward model for the evolution of an unlinked bi-allelic marker under horizontal transfer.
- Over time the marker for a lineage can change its state due to mutation or due to a horizontal transfer event in which the lineage copies the state from another, contemporary lineage.
- Mutations occur at a constant rate. The model works with infinite site, infinite allele, and the finite state models.
- During a horizontal transfer event one lineage transfers the state of a single site to another. These transfers occur at a constant rate. However, the rate of transfer can differ between different pairs of lineages.
- At a speciation event one lineage duplicates its state set.
- Rather than work directly with likelihoods we instead examine the probability that a given site is segregating for a set of lineages. This proves to be sufficient for reconstructing full likelihoods.



- We restricted ourselves to this rather simple model. However, an extension to more complicated markers (e.g., whole genome sequences) will be considered in future work.
- Nevertheless, there are situations in which our assumptions might be appropriate, e.g.:
  1. Gene content studies (e.g., Dagan and Martin ,2007; Huson and Steel, 2004a,b). Horizontal transfer in these cases is known as horizontal or lateral gene transfer (LGT).

     How much lateral gene transfer does there need to be before we have no chance of reconstructing the 'original' tree (if there is such a thing)?
  2. Language evolution where the sites correspond to cognates in different languages (e.g., Nakhleh, 2005; Bryant, 2006). Horizontal transfer is here also called borrowing. One possible question one can answer in our framework is:

     Can we infer the extent of borrowing that occurred between languages given only the current distribution of languages?

Results

Identifiability

- Given full knowledge of all parameters the true phylogeny can be reconstructed from the fraction of segregating sites observed.
- If at least the shape of the exchange matrix is known then the true phylogeny and the parameters for mutation and horizontal transfer can be recovered from the fraction of segregating sites observed.

Relation to site pattern probabilities

- Employing the relationship between fraction of segregating sites and site pattern probabilities we are able to infer the probability of a site pattern under a symmetric two-state model.

Reconstruction methods

- The knowledge of site pattern probabilities permits the application of a least squares method.
- The fraction of segregating sites for pairs of lineages can be viewed as a measure of divergence and hence permits the application of least square methods.
- We further derived a formula to compute the covariance between sets of lineages. The covariance matrix can be used in a generalized least square approach.

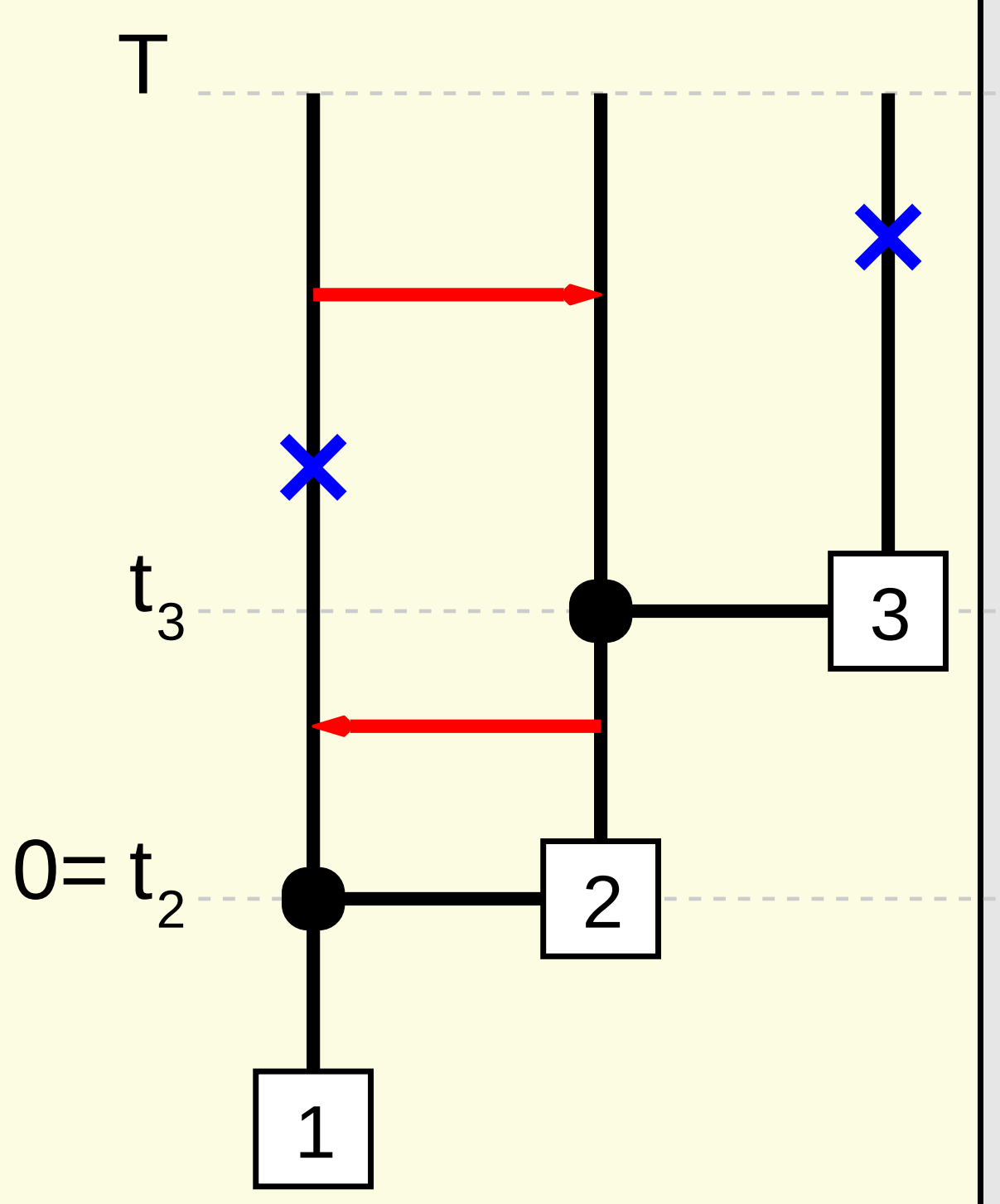
Technical playground

Model setup

- Let $t=0$ denote the time of the first speciation event and T today.
- Let $S_A(t)$ denote the event of observing a segregating site for a subset A of lineages, i.e. $S_A(t)=1$ if the site is segregating and 0 otherwise. Let $s_A(t)$ denote the probability, that $S_A(t)=0$, i.e. that we do not observe a segregating site. Consequently, $1-s_A(t)$ is the probability of observing a segregating site.

There are several events that can occur that will affect the value of $s_A(t)$.

- First regard horizontal transfers:
  1. A lineage a in A copies from a lineage b not in A, an event that occurs at the rate τD_{ab}. The probability that $S_A=0$ is the probability s_{A+b-a} that b has the same state as the other lineages in A, disregarding the lineage a it is copied into.
  2. A lineage a in A copies from a different lineage a' in A, which occurs at a rate $\tau D_{aa'}$. s_A is then given by the probability c_{A-a} that all the lineages in A with the exception of a have the same state. Across all a' in A this becomes $\sum_{a'\in A-a} D_{aa'}s_{A-a}$.
  3. A lineage a in A copies from itself, which occurs at a rate τD_{aa}. s_A does not change.
  4. No lineage in A copies with rate $-\tau|A|$.



- Mutation models can have the following effects
  1. Every time a mutation occurs under the infinite alleles model it establishes an entirely new allele. Thus if a mutation occurs in any lineage in A, which occurs at a rate $\mu|A|$ then $S_A=0$.
  2. In the infinite sites model only two possible states exist, an ancestral state and a derived state. The probability of a subset having the same state around a mutation is equal to that mutation not hitting a lineage in the subset.
  3. In a finite states model a mutation does not affect a subset A with rate $-\mu|A|$. If it affects the subset and $S_A(t)=0$, then $S_A(t+h)=1$. Hence, one has to consider the case that lineage $a\in A$ mutates, $S_{A-a}=0$ and lineage a attains the state of the other lineages through the mutation.

Using all of these setups one can derive and solve an ODE to compute the probabilities $s_A(t)$. The resulting system is lower block triangular, meaning that in order to compute $s_A(t)$ for some subset A we only need the values for all subsets B with $|B|\leq|A|$.

- At some time t a lineage a will duplicate itself thus creating a new lineage b. Then the probability $s_{ab}(t)=1$. Further for arbitrary subsets A we get:

$$s_A(t)=\begin{cases} s_A(t-\varepsilon), & b\notin A;\\ s_{A-b}(t-\varepsilon), & b\in A \text{ and } a\in A;\\ s_{A+a-b}(t-\varepsilon), & b\in A \text{ and } a\notin A.\end{cases}$$

- In reverse this property for speciation permits us to reconstruct speciation events and finding the smallest time t for which $s_{ab}(t)=1$ and $s_{xy}(t)<1$ for all other lineage pairs.

Relationship between probabilities for segregating sites and site patterns

- In general, the fraction of segregating sites of a subset A can be computed by simply summing up the probabilities for those site patterns which are segregating for A. In consequence, $s_A(T)$ is given as the sum of the probabilities for those site patterns which are non-segregating.
- In a symmetric two-state-model this yields a one-to-one relationship between fraction of segregating sites and site pattern probabilities. For $A\subseteq B$ denote by $F_{A;B}(t)$ the event that the lineages in A are in one state and the lineages in $B-A$ in the other. Let $f_{A;B}(t)$ denote the associated probability. Then, we first observe, that $f_{A;X}(t)$ is the site pattern probability for the site pattern where the lineages in A are in one state and the lineages in $X-A$ in the other. Using Möbius inversion we thus find a way to compute the site pattern probabilities $f_A(t)$ from the segregating sites counts.

Variance and Covariance

- Bryant (2006) employed the probabilities described above to derive the parameters from pairwise comparisons.
- Though the statistical efficiency of such pairwise comparisons is not fully investigated the above calculations permit us to derive for subsets $A,B\subseteq X$ the variance of $S_A(t)$ and covariance between $S_A(t)$ and $S_B(t)$.
- Interestingly, these computations only need the knowledge of $s_U(t)$ for all U with $|U|\subseteq|A\cup B|$.

  Explain that we have not investigated the statistical efficiency of using pairwise comparisons, but that we have derived variance and covariance formulae for these estimators (not necc. shown).

Identifiability

- If we are given the true parameters τ, μ and the true exchange matrix $\mathbb{D}$ then the original phylogeny can be reconstructed.
- The knowledge of the shape of true exchange matrix $\mathbb{D}$ permits the reconstruction of the original phylogeny and the inference of the associated parameters τ and μ.

References

Allman, E. S. and Holder, M. T. and Rhodes, J. A. (2010). *Estimating trees from filtered data: identifiability of models for morphological phylogenetics.* JTB.

Chang, J. T. (1996). *Full reconstruction of Markov models on evolutionary trees: identifiability and consistency.* Math. Bioscie.

Dagan, D. and Martin, W. (2007). *Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution.* PNAS

Huson, D. and Steel, M. (2004a). *Phylogenetic trees based on gene content.* Bioinf.

Huson, D. and Steel, M. (2004b). *Distances that perfectly mislead.* Syst. Biol.

Klaere S. and Kydd, J. and Bryant, D. *Identifiability of phylogenetic trees in the presence of horizontal transfer.* Manuscript in Prep.

Nakhleh, L. and Ringe, D. and Warnow, T. (2005). *Perfect Phylogenetic Networks: A New Methodology for Reconstructing the Evolutionary History of Natural Languages.* Language.

Rivera, M. C. and Lake, J. A. (2004). *The ring of life provides evidence for a genome fusion origin of eukaryotes.* Nature.

Steel, M. A. (1992). *The complexity of reconstructing trees from qualitative characters and subtrees.* J. Class.