Use of phylogenetic invariants to estimate species trees under the coalescent model

Julia Chifman
joint work with Laura Kubatko*

Mathematical Biosciences Institute
*Departments of Statistics and Evolution, Ecology, and Organismal Biology, The Ohio State University

May 25, 2010

Gene Trees vs. Species Trees

- When DNA sequence data from a single gene are used to estimate a phylogeny, a gene tree (a tree representing the evolution of that gene) is estimated.
- Most often, the species tree (a tree representing the actual evolutionary path of the species) is what is desired.
- The gene and species trees often disagree with one another. This is called topological incongruence.

Some Biological Explanations for Incongruence

- Horizontal gene transfer is the transfer of genetic material by processes other than usual reproduction. For example, genetic material can be carried from one cell to another by infectious viruses.
- Hybridization: the genetic process of crossbreeding between genetically dissimilar parents to produce a hybrid.
- Deep coalescence is when two ancestral gene copies fail to coalesce (looking backwards in time) into a common ancestral copy until deeper than previous speciation events.

The Coalescent Process

- The coalescent models the random process by which individual gene histories evolve under the constraints imposed by an overall species tree.

- A consequence of the model is the requirement that gene divergence times pre-date speciation events, resulting in the possibility of variation in the phylogenetic trees for individual genes.

- This possibility exists even when all of the individual genes are compatible with a single, bifurcating species tree.

The Coalescent Process

More specifically...

- The coalescent process specifies that the time to coalescence of j gene copies into $j-1$ is exponentially distributed with mean

$$\frac{j(j-1)}{2}\frac{2}{\theta}$$

- The parameter θ is defined to be $4N\mu$
  - N= the effective population size
  - μ= the mutation rate per site per generation.

- Using this model, Rannala and Yang (2003) derived the gene tree probability density function along a fixed species tree.

The Coalescent Process

- We consider the density of the entire gene tree under the usual assumption that coalescent events are independent across branches
- Then the densities across individual branches can be multiplied to give the density of gene tree
- In this talk we will denote the density of gene tree given species tree by

$$f_{(G,\mathbf{t})|(S,\boldsymbol{\tau})}$$

  - S =species tree
  - G=gene tree
  - $\mathbf{t} = (t_1, \cdots t_{n-1})$ the vector of coalescent times
  - $\boldsymbol{\tau} = (\tau_1, \cdots \tau_{n-1})$ the vector of speciation times

Gene tree - review

- Let T be an n-leaf rooted tree. At the root the distribution of the states is given by $\pi = (\pi_1, \ldots, \pi_k)$.

- Edges e of T are labeled by $k \times k$ transition probability matrices M_e, that reflect probabilities of changes of the states from a node to its child.

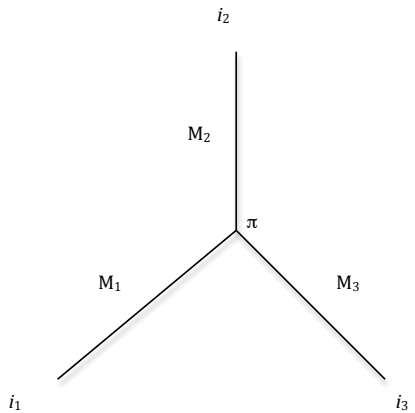
- For example, the Jukes-Cantor (JC69) model specifies a common rate of change among the four nucleotides, leading to the following transition probability matrix

$$M_e = \begin{pmatrix} 1-a_e & \frac{a_e}{3} & \frac{a_e}{3} & \frac{a_e}{3} \\ \frac{a_e}{3} & 1-a_e & \frac{a_e}{3} & \frac{a_e}{3} \\ \frac{a_e}{3} & \frac{a_e}{3} & 1-a_e & \frac{a_e}{3} \\ \frac{a_e}{3} & \frac{a_e}{3} & \frac{a_e}{3} & 1-a_e \end{pmatrix}$$

where $a_e = \frac{3}{4}(1 - e^{-\frac{4}{3}v_e})$ and v_e is the branch length.

- Computation of site pattern probabilities on gene trees is straightforward once the gene tree and the Markov model have been specified.
- For a particular observation $i_1 \dots i_n$ at the leaves of a gene tree G, where i_m denotes the state at leaf m, let $p_{i_1 i_2 \dots i_n}$ be the probability of site pattern $i_1 i_2 \dots i_n$.
- For example, if G is a tree with three leaves and four states, then probability of a particular observation $i_1 i_2 i_3$ is

$$p_{i_1 i_2 i_3} = \sum_{\ell=1}^{4} \pi_\ell M_1(\ell, i_1) M_2(\ell, i_2) M_3(\ell, i_3)$$



Site pattern probabilities on species trees

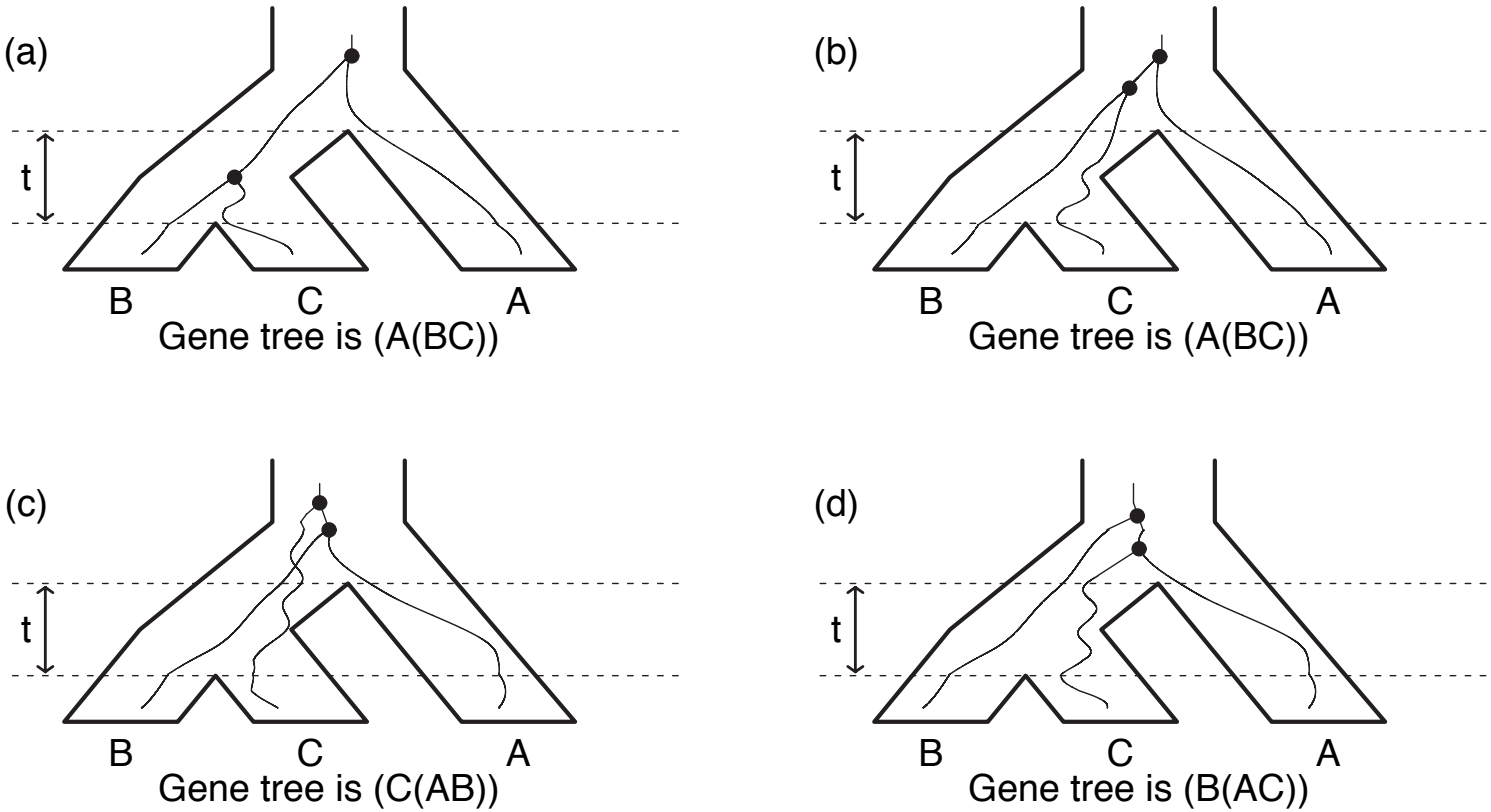
- Gene tree site pattern probabilities can then be used together with the gene tree density $f_{(G,\mathbf{t})|(S,\boldsymbol{\tau})}$, to compute the site pattern probabilities along the species trees.

- Since the true gene tree is unobserved, we must consider all possible gene trees that are consistent with the given species tree, and weight the probability of the site pattern of interest appropriately by the probability of each gene tree under the coalescent model.

- This leads to the following expression for the probability of observation $i_1 \dots i_n$ for species tree (S, τ)

$$P_{i_1 \dots i_n|(S,\boldsymbol{\tau})} = \sum_G \int_{\mathbf{t}} p_{i_{\sigma_G(1)} \dots i_{\sigma_G(n)}|(G,\mathbf{t})} \cdot f_{(G,\mathbf{t})|(S,\boldsymbol{\tau})} d\mathbf{t}.$$

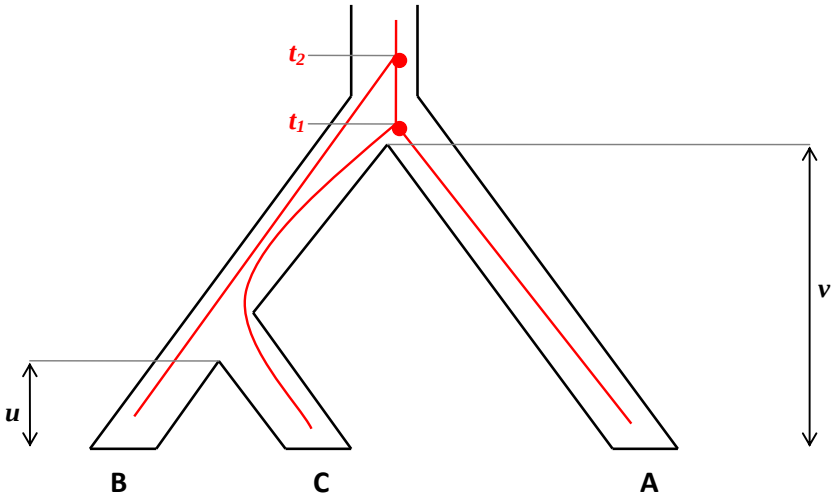
- Where the sum is taken over all gene trees G with corresponding branch lengths $\mathbf{t}$ appropriately integrated out and $i_{\sigma_G(1)} \dots i_{\sigma_G(n)}$ is a permutation of the observation $i_1 \dots i_n$ at the leaves of the species tree S.

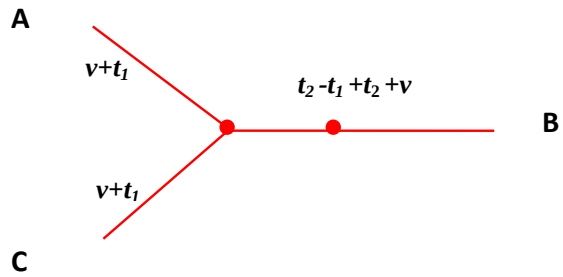

(a)
t
B
C
A
Gene tree is (A(BC))
(b)
t
B
C
A
Gene tree is (A(BC))
(c)
t
B
C
A
Gene tree is (C(AB))
(d)
t
B
C
A
Gene tree is (B(AC))

Example

- Let G_i denote a particular gene tree for history i, then the i^{th} summand of $P_{TGG|(S,(u,v))}$ is:

$$\int_0^{\infty} \int_0^{t_2} p_{GGT|(G_i,(v+t_1,v+t_1,2t_2-t_1+v))} \cdot f_{(G_i,(v+t_1,v+t_1,2t_2-t_1+v))|(S,(u,v))} dt_1 dt_2$$





Example (cont.)

Under the Jukes-Cantor model:

$$\int_0^{\infty}\int_0^{t_2} p_{GGT|(G_i,(v+t_1,v+t_1,2t_2-t_1+v))} \cdot f_{(G_i,(v+t_1,v+t_1,2t_2-t_1+v))|(S,(u,v))}\, dt_1\, dt_2$$

- $p_{GGT|(G_i,\mathbf{t})} = \frac{1}{64} - \frac{1}{32}e^{-\frac{4t_1}{3}-\frac{8t_2}{3}-4v} - \frac{1}{64}e^{-\frac{8t_1}{3}-\frac{8v}{3}} + \frac{1}{32}e^{-\frac{8t_2}{3}-\frac{8v}{3}}$
- $f_{(G_i,\mathbf{t})|(S,\tau)} = (\frac{2}{\theta})^2 e^{-\frac{2}{\theta}(v-u)} e^{-6\frac{t_1}{\theta}} e^{-\frac{2}{\theta}(t_2-t_1)}$
- Integrate and get:

$$\frac{1}{192}e^{\frac{2u}{\theta}}\left(e^{-\frac{2v}{\theta}} - \frac{9(-3+4\theta)e^{\left(-\frac{8}{3}-\frac{2}{\theta}\right)v}}{(3+4\theta)(9+4\theta)} - \frac{18e^{-2\left(2+\frac{1}{\theta}\right)v}}{(3+2\theta)(3+4\theta)}\right)$$

- Apply same process to other histories and add them together to get $P_{TGG|(S,(u,v))}$.

3-leaf species tree for a JC69 model under the coalescent

- When considering a rooted species tree (as required by the coalescent model), the symmetry among nucleotides implied by the JC69 model leads to 4 distinct site patterns out of 64 patterns.

- For example, $P_{ACT} = P_{CGA} = \cdots = P_{GCT}$ under this model, and thus we write this pattern as P_{xyz}, where $x, y, z \in \{A, C, G, T\}$.

- For a simpler output let $\theta = \frac{1}{100}$

- Then, site pattern probabilities along a 3-leaf species tree for a Jukes-Cantor model under the coalescent are as follows.

$$Pxxx = \frac{1}{16} + \frac{225e^{-\frac{8u}{3}}}{1216} + \frac{16875e^{-\frac{4u}{3}-\frac{8v}{3}}}{45904} + \frac{225}{608}e^{-\frac{8v}{3}}$$

$$Pxxy = \frac{3}{16} - \frac{225e^{-\frac{8u}{3}}}{1216} - \frac{16875e^{-\frac{4u}{3}-\frac{8v}{3}}}{45904} + \frac{225}{608}e^{-\frac{8v}{3}}$$

$$Pxyx = \frac{3}{16} - \frac{225e^{-\frac{8u}{3}}}{1216} - \frac{16875e^{-\frac{4u}{3}-\frac{8v}{3}}}{45904} + \frac{225}{608}e^{-\frac{8v}{3}}$$

$$Pyxx = \frac{3}{16} + \frac{675e^{-\frac{8u}{3}}}{1216} - \frac{16875e^{-\frac{4u}{3}-\frac{8v}{3}}}{45904} - \frac{225}{608}e^{-\frac{8v}{3}}$$

$$Pxyz = \frac{3}{8} - \frac{225}{608}e^{-\frac{8u}{3}} + \frac{16875e^{-\frac{4u}{3}-\frac{8v}{3}}}{22952} - \frac{225}{304}e^{-\frac{8v}{3}}$$

Phylogenetic Invariants

- Notice that $P_{xxy} - P_{xyx} = 0$.
- Also, since the probability of all possible outcomes must add to 1 we get $\sum_{ijk} P_{ijk} - 1 = 0$
- These polynomial relations are called invariants.
- Phylogenetic invariant is a polynomial in the site pattern probabilities that vanishes when evaluated on any distribution arising from the model i.e., the true phylogenetic tree and associated Markov model

Phylogenetic Invariants

- Phylogenetic invariants and their possible use for inference were first introduced in 1987 by Cavender and Felsenstein, and by Lake
- The general idea is to fix a model and for each tree find and evaluate invariants at observed frequencies of patterns in real data sequenses.
- The best estimate of the phylogenetic tree is the tree for which invariants are "nearly zero."

3-leaf species tree for a JC69 model under the coalescent

- We can make the following change of variables by introducing new parameters

$$x = e^{\frac{-8v}{3}} \text{ and } y = e^{\frac{-4u}{3}}$$

- Then we can express site pattern probabilities as polynomials

$$Pxxx = \frac{1}{16} + \frac{225}{1216}y^2 + \frac{16875}{45904}xy + \frac{225}{608}x$$

$$Pxxy = \frac{3}{16} - \frac{225}{1216}y^2 - \frac{16875}{45904}xy + \frac{225}{608}x$$

$$Pxyx = \frac{3}{16} - \frac{225}{1216}y^2 - \frac{16875}{45904}xy + \frac{225}{608}x$$

$$Pyxx = \frac{3}{16} + \frac{675}{1216}y^2 - \frac{16875}{45904}xy - \frac{225}{608}x$$

$$Pxyz = \frac{3}{8} - \frac{225}{608}y^2 + \frac{16875}{22952}xy - \frac{225}{304}x$$

- Notice that our new parameters x and y, and each P_{ijk} are between 0 and 1.
- For a fixed species tree S we are going to consider a polynomial map from a parameter space to a joint distribution space

$$F_S : [0,1]^{\text{number of parameters}} \to [0,1]^{\text{number of site patterns}}$$

- To keep algebra as simple as possible we are going to extend this map to a complex setting, which is appropriate for studying polynomial maps.

$$\begin{aligned} F_S : \mathbb{C}^2 &\to \mathbb{C}^{64} \\ (x,y) &\mapsto (P_{AAA}, P_{AAC}, P_{AAG}, \ldots, P_{TTT}) \end{aligned}$$

- What we want is the closure of the image $\overline{F_S(\mathbb{C}^2)} =$ an algebraic variety.

- Let $f_1, f_2, \ldots, f_k$ be polynomials in $\mathbb{C}[x_1, x_2, \ldots, x_n]$, then variety is the set of all solutions of the system

$$f_1(x_1, x_2, \ldots, x_n) = \cdots = f_k(x_1, x_2, \ldots, x_n) = 0$$

- With each variety we associate a set of polynomials that vanish on the given variety. This set of polynomials forms an ideal.

- In our setting we will denote this ideal by I_S, where S is the species tree.

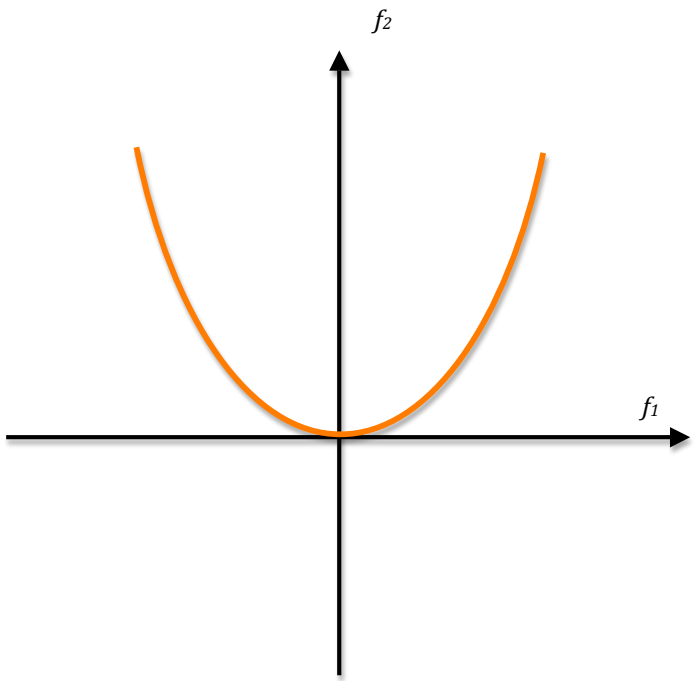
- Thus, we associate to each tree S a parametrized surface V_S = the image of our map.

Example

- Let $f_1 = t$ and $f_2 = t^2$

$$F : \mathbb{R} \to \mathbb{R}^2$$
$$t \mapsto (f_1, f_2)$$

- The image of the map is a parabola and associated ideal is $I = (f_2 - f_1^2)$.



- To compute this ideal I_S is very hard.
- For small trees and certain models we can use "Implicitization algorithm for polynomial parametrization" (next slide)
- However, under realistic assumptions for θ, our model will always result in large degrees for trees with leaves greater than 4 and in rational functions.
- Hence, we have little hope of using software packages to compute generating set. Rather, we have to search for natural constructions and possible change of coordinates that will simplify calculations.

Implicitization algorithm for polynomial parametrization

- In principle, given the polynomial parameterization

$$P_1 = f_1(x_1, \ldots, x_n), \ldots, P_m = f_m(x_1, \ldots, x_n),$$

- where $f_1, \ldots, f_m$ are polynomial functions in $\mathbb{C}[x_1, \ldots, x_n]$

- Let $I = \langle P_1 - f_1, \ldots, P_m - f_m \rangle$, and compute the *Gröbner* basis with respect to a *lexicographic* order.

- The elements of the basis that do not involve $x_1, \ldots, x_n$ define the smallest variety containing the parameterization.

3-leaf species tree for a JC69 model under the coalescent

$$Pxxx = \frac{1}{16} + \frac{225}{1216}y^2 + \frac{16875}{45904}xy + \frac{225}{608}x$$

$$Pxxy = \frac{3}{16} - \frac{225}{1216}y^2 - \frac{16875}{45904}xy + \frac{225}{608}x$$

$$Pxyx = \frac{3}{16} - \frac{225}{1216}y^2 - \frac{16875}{45904}xy + \frac{225}{608}x$$

$$Pyxx = \frac{3}{16} + \frac{675}{1216}y^2 - \frac{16875}{45904}xy - \frac{225}{608}x$$

$$Pxyz = \frac{3}{8} - \frac{225}{608}y^2 + \frac{16875}{22952}xy - \frac{225}{304}x$$

3-leaf species tree for a JC69 model under the coalescent

$$G1 = Pxxx - (\frac{1}{16} + \frac{225}{1216}y^2 + \frac{16875}{45904}xy + \frac{225}{608}x)$$

$$G2 = Pxxy - (\frac{3}{16} - \frac{225}{1216}y^2 - \frac{16875}{45904}xy + \frac{225}{608}x)$$

$$G3 = Pxyx - (\frac{3}{16} - \frac{225}{1216}y^2 - \frac{16875}{45904}xy + \frac{225}{608}x)$$

$$G4 = Pyxx - (\frac{3}{16} + \frac{675}{1216}y^2 - \frac{16875}{45904}xy - \frac{225}{608}x)$$

$$G5 = Pxyz - (\frac{3}{8} - \frac{225}{608}y^2 + \frac{16875}{22952}xy - \frac{225}{304}x)$$

3-leaf species tree for a JC69 model under the coalescent

Using *Mathematica*:

input: GroebnerBasis[{G1, G2, G3, G4, G5}, {x, y, Pxxx, Pxxy, Pxyx, Pyxx, Pxyz}, {x, y}]

output: $\{9 - 48\text{Pxyx} - 364736\text{Pxyx}^2 + 972800\text{Pxyx}^3 + 547188\text{Pxyz} - 2188768\text{PxyxPxyz}$
$+2432000\text{Pxyx}^2\text{Pxyz} - 1003196\text{Pxyz}^2 + 1945600\text{PxyxPxyz}^2 + 486400\text{Pxyz}^3 - 24\text{Pyxx}$
$-729536\text{PxyxPyxx} + 1945600\text{Pxyx}^2\text{Pyxx} - 1094384\text{PxyzPyxx} + 2918400\text{PxyxPxyzPyxx}$
$+972800\text{Pxyz}^2\text{Pyxx} + 16\text{Pyxx}^2 + 972800\text{PxyxPyxx}^2 + 486400\text{PxyzPyxx}^2,$
$\text{Pxxy} - \text{Pxyx}, -1 + \text{Pxxx} + 2\text{Pxyx} + \text{Pxyz} + \text{Pyxx}\}$

Finch Data

- W097 = Poephila acuticauda
  Q097 = Poephila hecki
  B097 = Poephila cincta

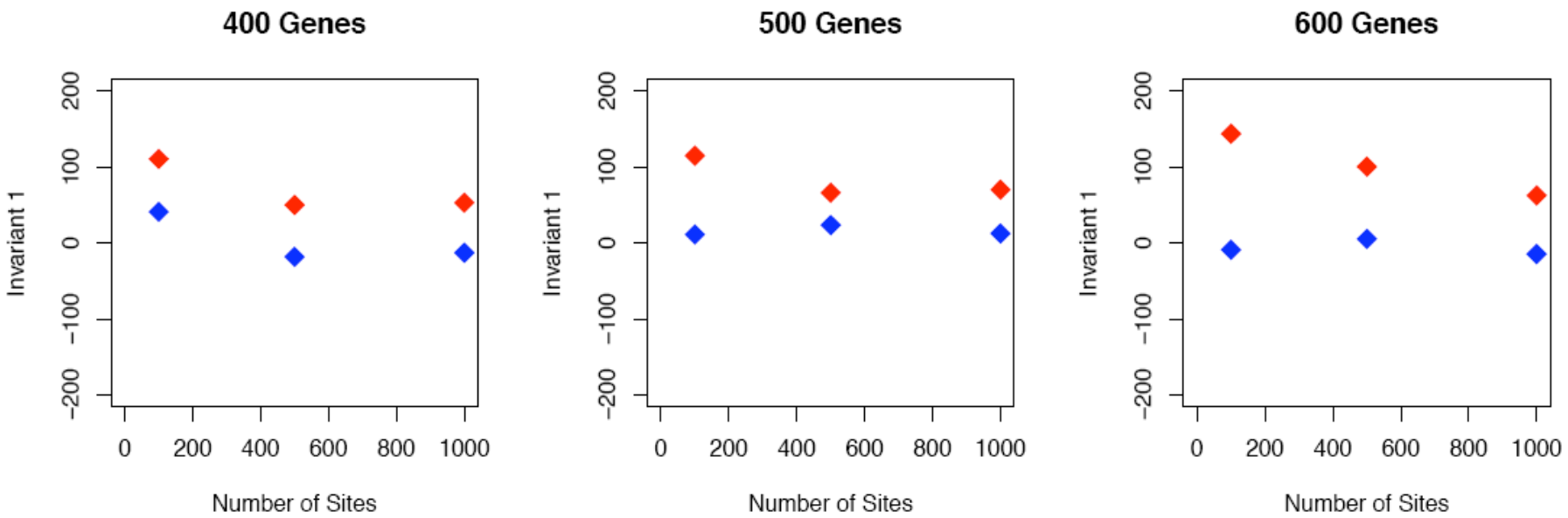
- Site pattern counts under JC69:
  Pxxx 15860
  Pxxy 115
  Pxyx 69
  Pyxx 74
  Pxyz 1

- The tree: ((Q:W):B)

	Cubic invariant	Linear invariant*10000
((Q:W):B)	10.4873	3.10193
((Q:B):W)	20.041	25.4358
(Q:(W:B))	21.0492	28.5378

Results for simulated data: cubic invariant

Blue - true tree
Red - alternate tree



Thank you!