

Prospects and Pitfalls of Whole Genome Phylogeny

Laura A. Salter

Department of Mathematics and Statistics

University of New Mexico

New Mexico Chapter of the American Statistical Association

September, 2004

What is Phylogenetics?

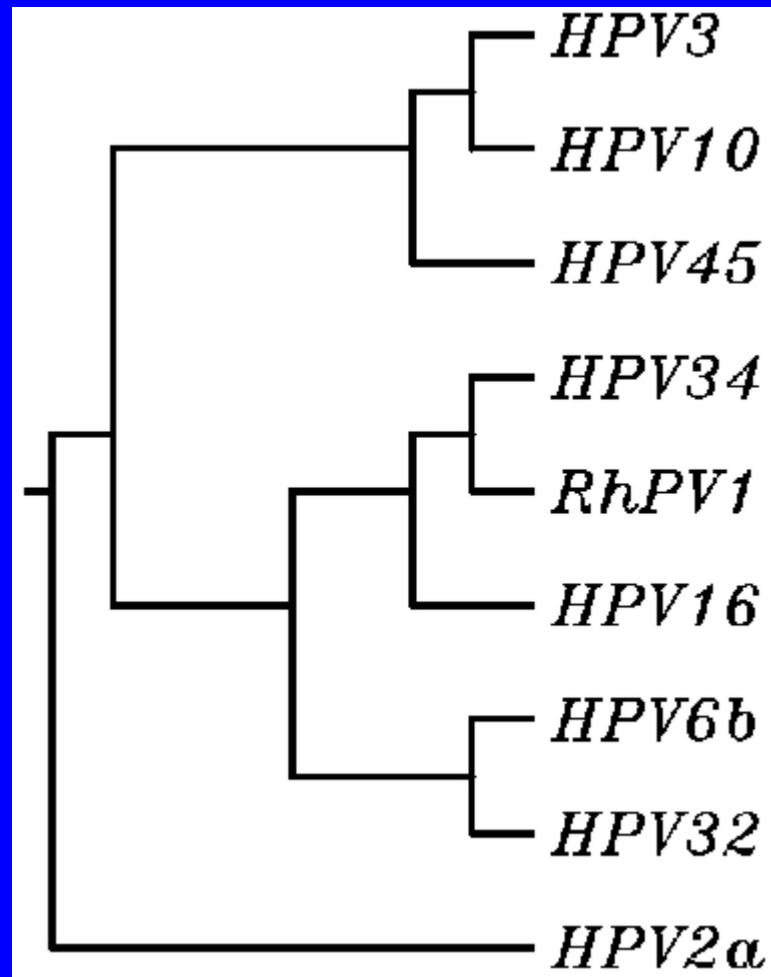
- ▶ Goal: To infer evolutionary relationships among a collection of organisms, based on
 - morphological characteristics
 - genetic information
 - nucleotide sequences
 - amino acid sequences
 - allele frequencies

HPV Example

▶ Example HPV Sequences

Sequence	First 20 Sites
HPV32	ATGTGGCGGCCTAGTGACAA
HPV3	-----CT-----
HPV10	-----CT-----
HPV2a	-----A---A-G
HPV16	-----T-----GGC
HPV34	-----A---C---GGC
HPV45	-----G
HPV6b	-----C---G
RhPV1	-----TC

Example Phylogenetic Tree for the HPV Data



Impact of Genome Sequencing

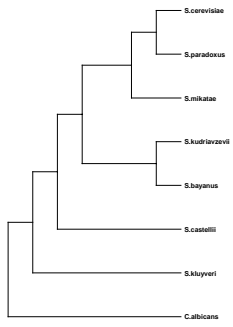
- ▶ Projects like the Human Genome Project are being undertaken for **many** organisms, including
 - mouse
 - *Neurospora crassa* (fungus)
 - *Drosophila* (fruit fly)
 - many others!
- ▶ The result is that significant portions of the genome sequences are now, or will soon be, known for many organisms

Impact of Genome Sequencing

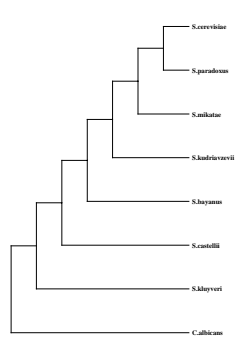
- ▶ It is becoming increasingly common to have a complete data set for a collection of organisms for several genes sampled from different portions of the genome
- ▶ Often, the trees estimated from the individual genes disagree with one another in important aspects of the tree - this is called **topological incongruence**

Example of Topological Incongruence

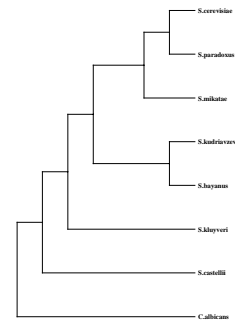
- ▶ Rokas et al. (*Nature*, 2003) examined 106 genes for 8 species of yeast
- ▶ Estimated phylogenies separately for each gene, and examined the level of topological incongruence
- ▶ A total of 12 distinct phylogenies were estimated from the 106 genes
- ▶ **What is the true evolutionary relationship for these 8 taxa?**



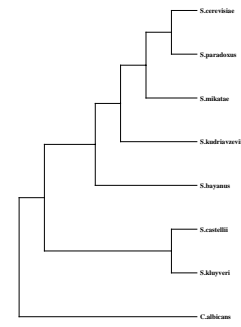
34



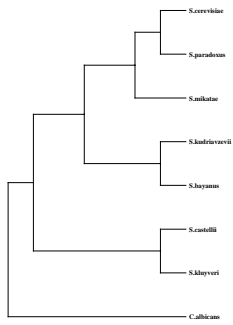
28



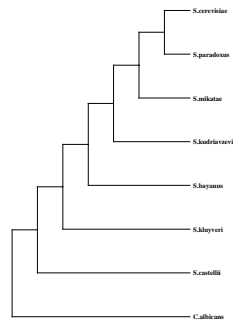
12



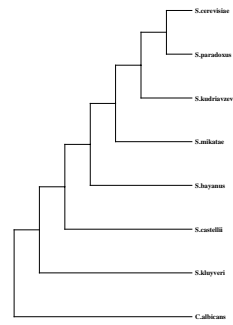
10



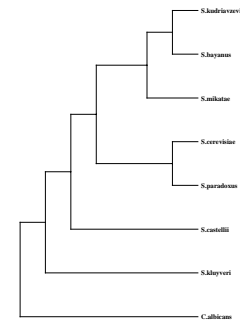
9



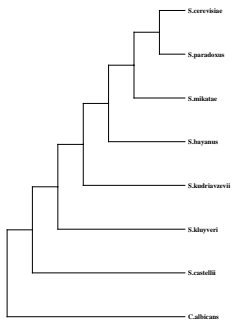
7



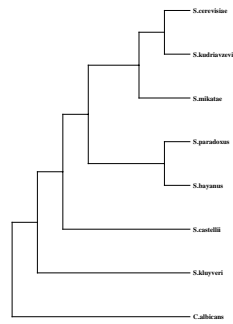
1



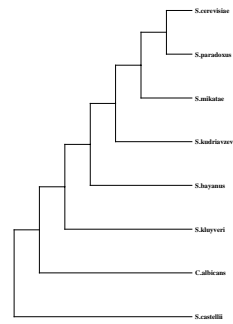
1



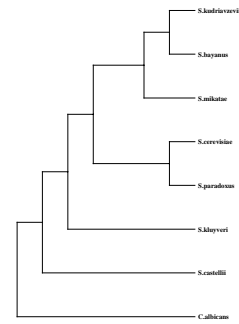
1



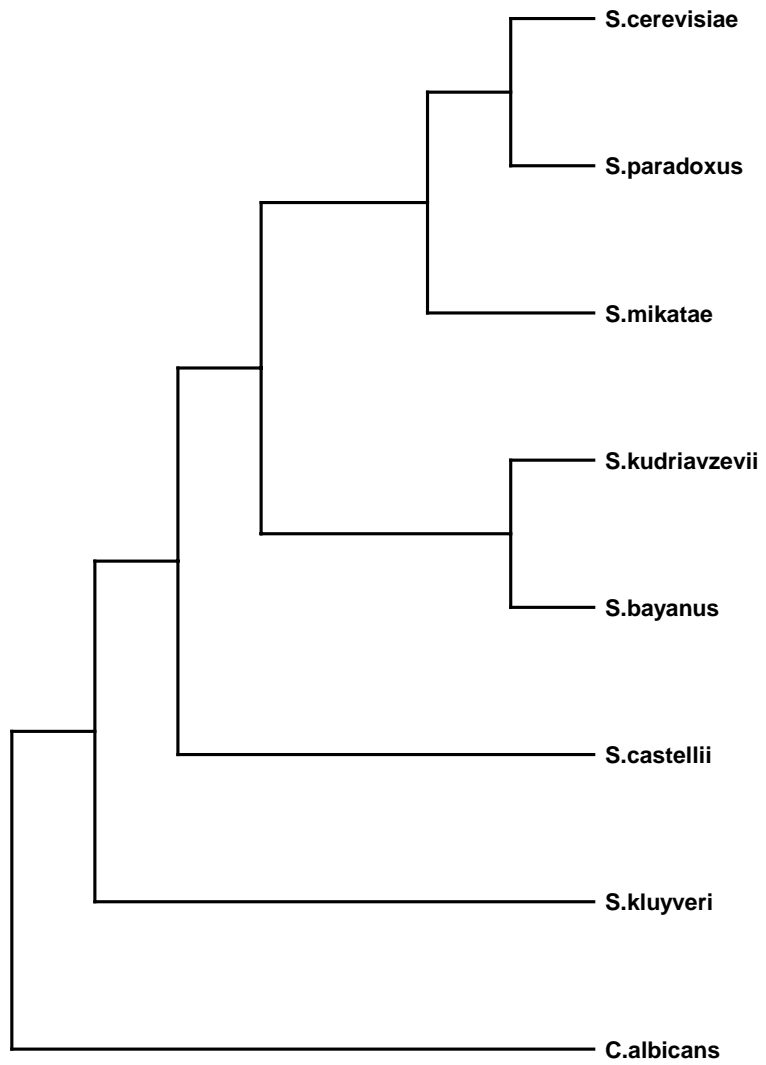
1



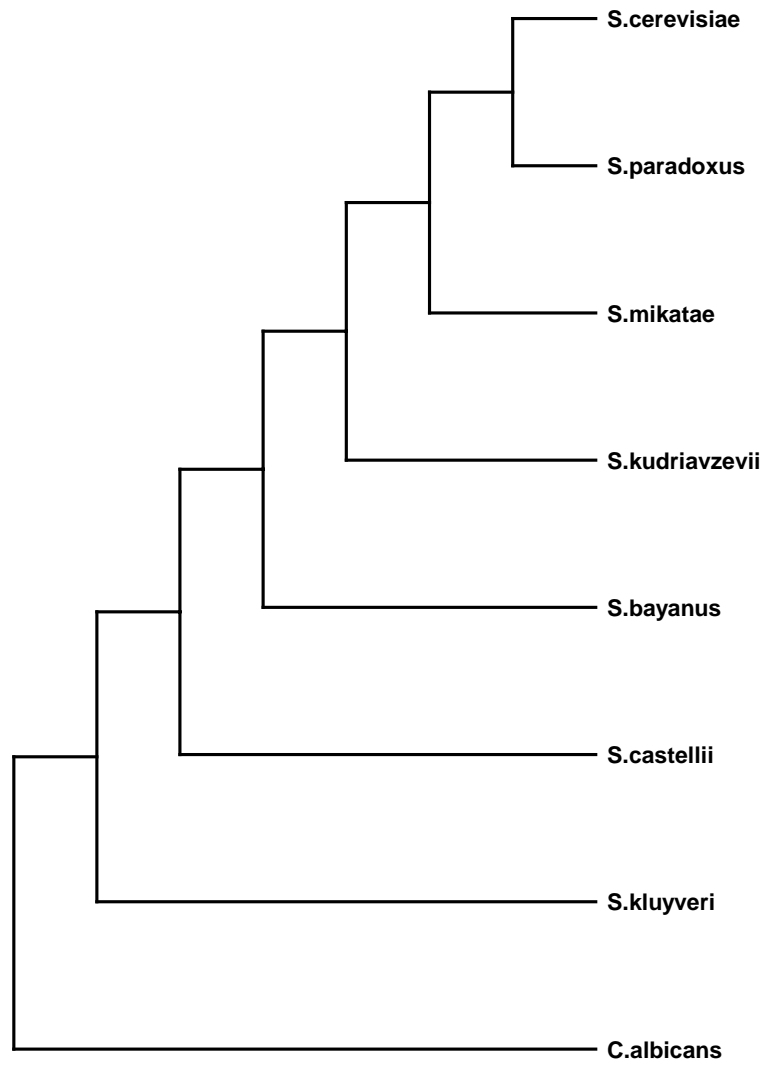
1



1



34



28

Why do we have incongruence?

▶ Is this just stochastic error?

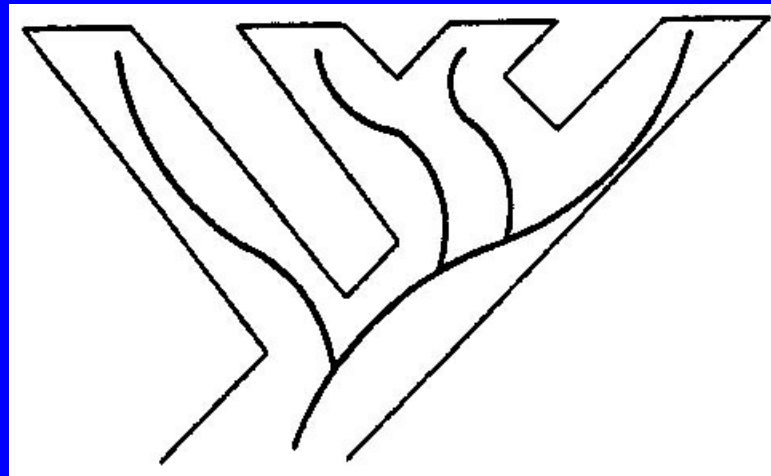
OR

▶ Does this reflect a biological process?

▶ Rokas et al. approach: assume stochastic error is responsible for the differences; in this case, concatenation of the data should result in a single strongly supported tree

Deep Coalescence

Presence of multiple copies of the gene in ancestral population can lead to gene relationships that differ from species relationships



The coalescent process predicts significant variability in gene tree topologies, even when no other processes are at work.

COAL

- ▶ COAL is a program to compute the probability distribution of gene tree topologies for given species trees and branch lengths (written by UNM grad student James Degnan)
- ▶ Use COAL to assess whether coalescence provides a satisfactory explanation of the variation in gene tree topologies observed in the 106 yeast gene data set, and whether it changes our species tree inference

Gene Tree Probabilities

Tree	No. of Genes	Species Tree #1	Species Tree #2
1	34	0.1145	0.0657
2	28	0.0511	0.0845
3	12	0.0195	0.0111
4	10	0.0087	0.0143
5	9	0.0195	0.0111
6	7	0.0087	0.0143
7	1	0.0127	0.0264
8	1	0.0351	0.0094
9	1	0.0087	0.0096
10	1	0.0013	0.0021
11	1	0.0000	0.0000
12	1	0.0060	0.0016

Maximum Likelihood Species Tree

- ▶ Assuming independence of the genes, we can use this information to compute a likelihood for each species tree:

Species Tree #1: $\ln L = -369.845$

Species Tree #2: $\ln L = -379.362$

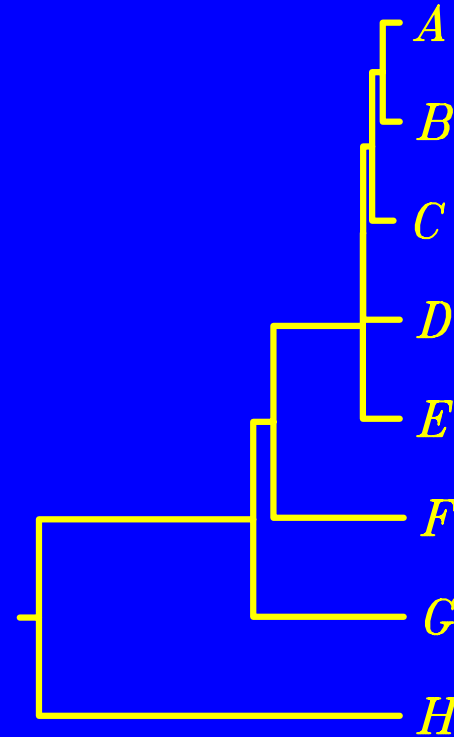
- ▶ Tree #1 is the ML estimate (of the two) for the species tree
- ▶ In this case, the inferred species tree is the same as it is for the concatenation approach

Do the Concatenation and Coalescent Approaches Have to Agree?

NO!

► An example:

Consider this species tree, and let's compute the probability of each of 12 gene trees



Gene Tree Probabilities

Tree	No. of Genes	Example Species Tree
1	34	0.0743
2	28	0.0599
3	12	0.0126
4	10	0.0102
5	9	0.0126
6	7	0.0102
7	1	0.0165
8	1	0.0166
9	1	0.0097
10	1	0.0020
11	1	0.0000
12	1	0.0028

How would this affect concatenation?

- ▶ View DNA sequence data as the result of a two-stage process:
 - Coalescence process generates a gene tree topology
 - Given this gene tree topology, DNA sequences evolve along the tree
- ▶ ML estimation of the **gene tree** is known to be consistent – as more data is added, we become more likely to estimate the tree **that generated the data**

How would this affect concatenation?

- ▶ Thus, as more data are added, we become more likely to estimate the **wrong** tree
- ▶ How likely is this for real data?
 - Depends on tree topology, branch lengths, and population sizes
 - Other evolutionary processes can lead to topological incongruence, such as hybridization, horizontal transfer, and gene duplication or extinction

Open Problems

- ▶ Implement true ML estimation of the species tree - here, several things were assumed:
 - Gene tree topologies known without error
 - Species tree branch lengths known
 - Only considered two most frequently observed gene trees as candidates for the species tree
- ▶ Derivation of conditions under which the concatenation approach will fail
- ▶ Rokas et al. provide a numerical examination of the number of genes required for concatenation to be successful – examine this bound under their assumptions and re-evaluate this for our coalescent approach

Acknowledgements

- ▶ James Degnan – UNM grad student who developed the algorithm and program to compute gene tree probabilities
- ▶ Amy Powell – former UNM Biology grad student who sent me the Rokas et al. paper
- ▶ Students in Phylogenetics Seminars for the past 5 years, who explained the gene tree/species tree problem to me
- ▶ National Science Foundation

References

- ▶ Degnan, J. and L. Salter. 2004. Gene tree distributions under the coalescent process (in revision for *Evolution*).
- ▶ Rokas, A., B. Williams, N. King, and S. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425: 798-804.
- ▶ Jensen, M., H. True, Y. Chernoff, and S. Lundquist. 2001. Molecular population genetics and evolution of a prion-like protein in *Sacchromyces cerevisiae*. *Genetics* 159: 27-535.
- ▶ Maddison, W. xxxx Gene trees in species tree. *Systematic Biology* xxxxxx

The End!