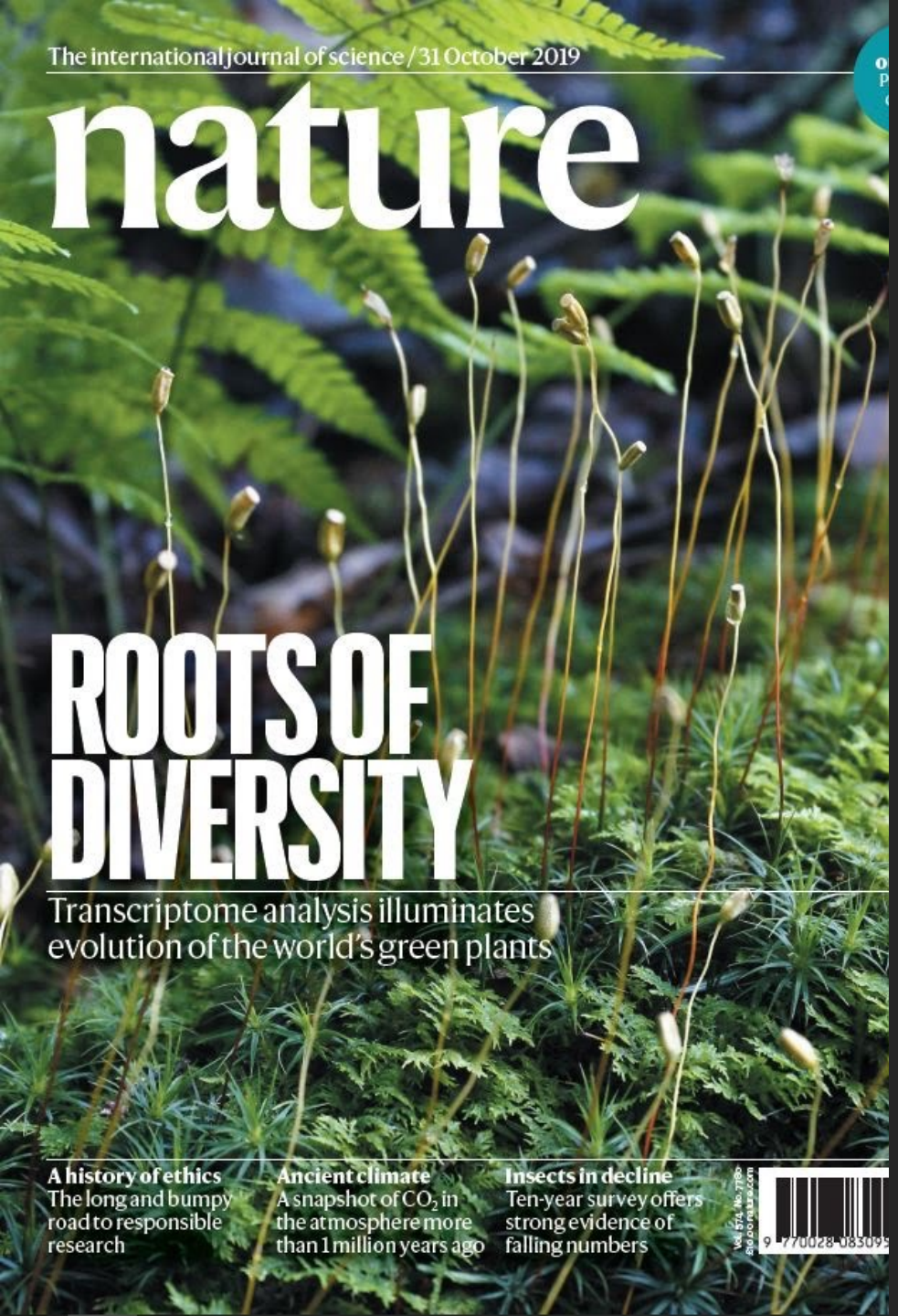
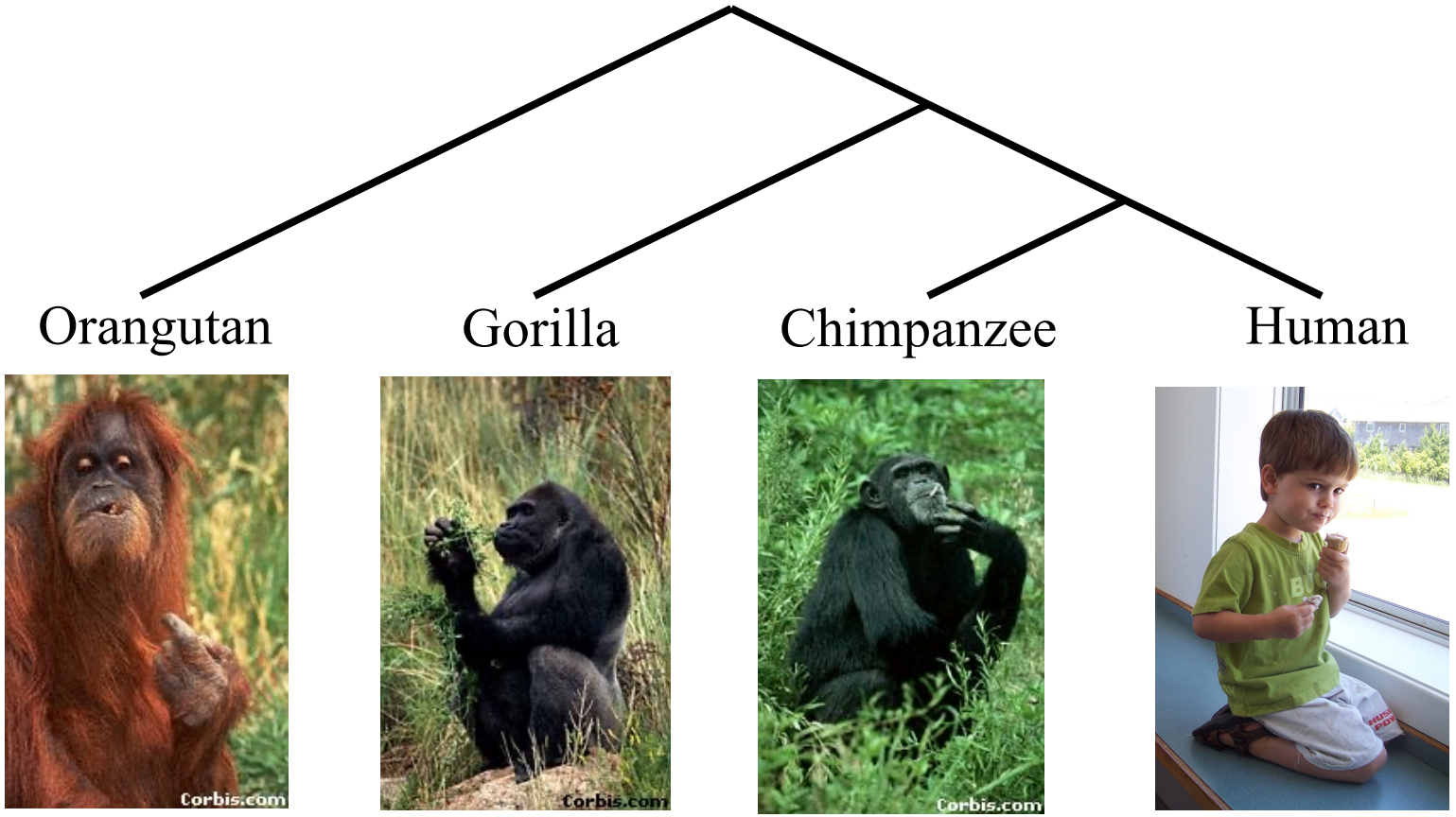


Why I love phylogenetics!

Tandy Warnow

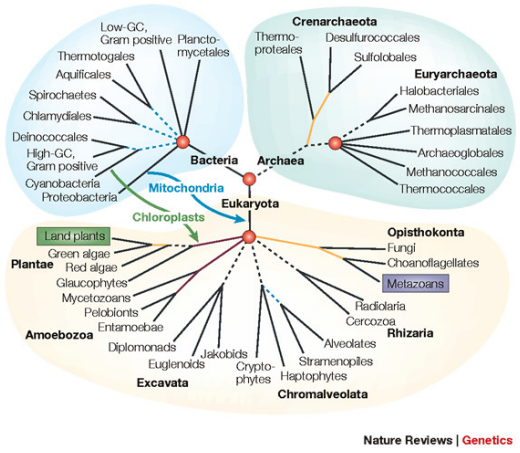
The University of Illinois

Phylogeny (evolutionary tree)



From the Tree of the Life Website,
University of Arizona

Phylogenetic Inference



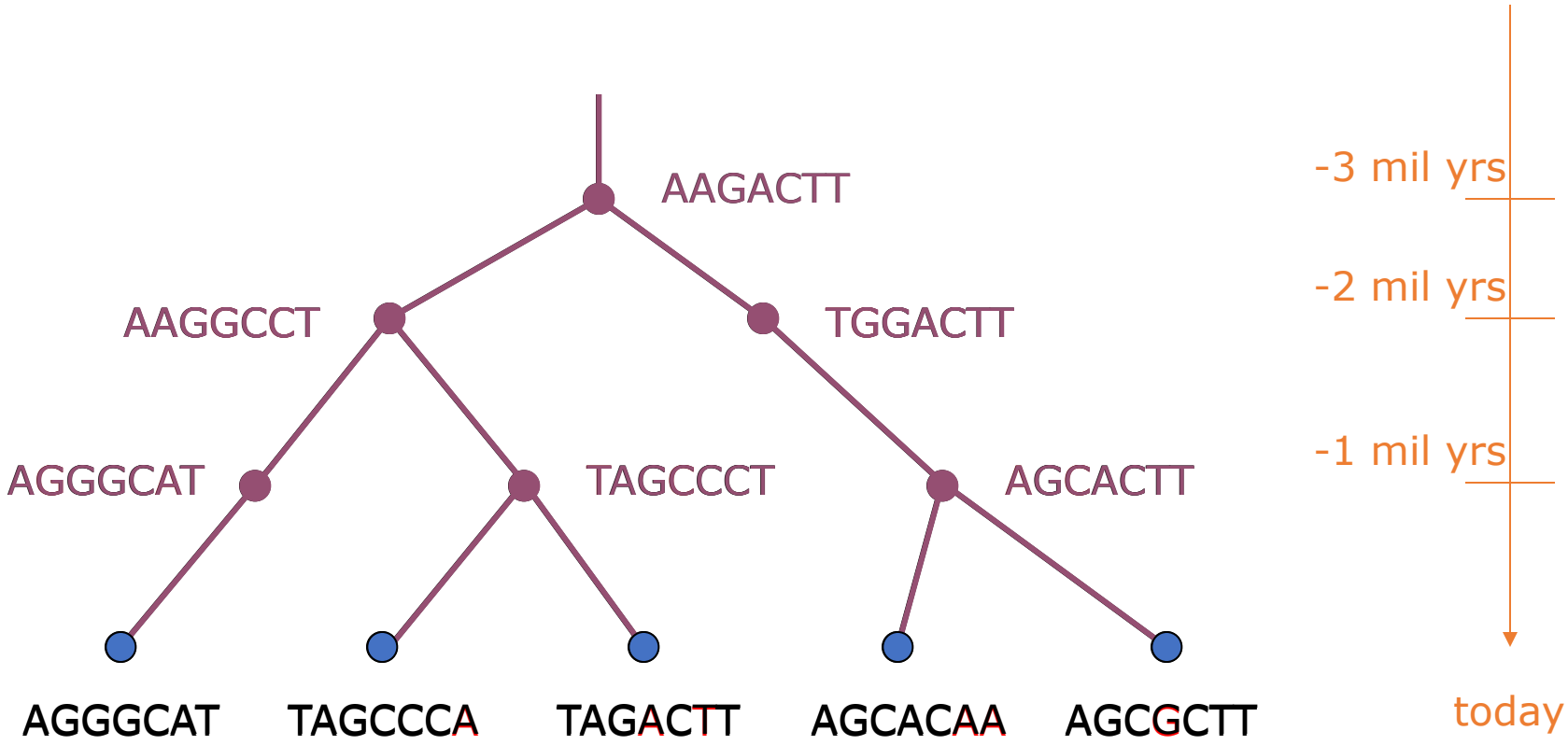
“Big Data”:

- Heterogeneous
- Large
- Noisy
- Error-ridden
- Streaming
- Model-misspecification

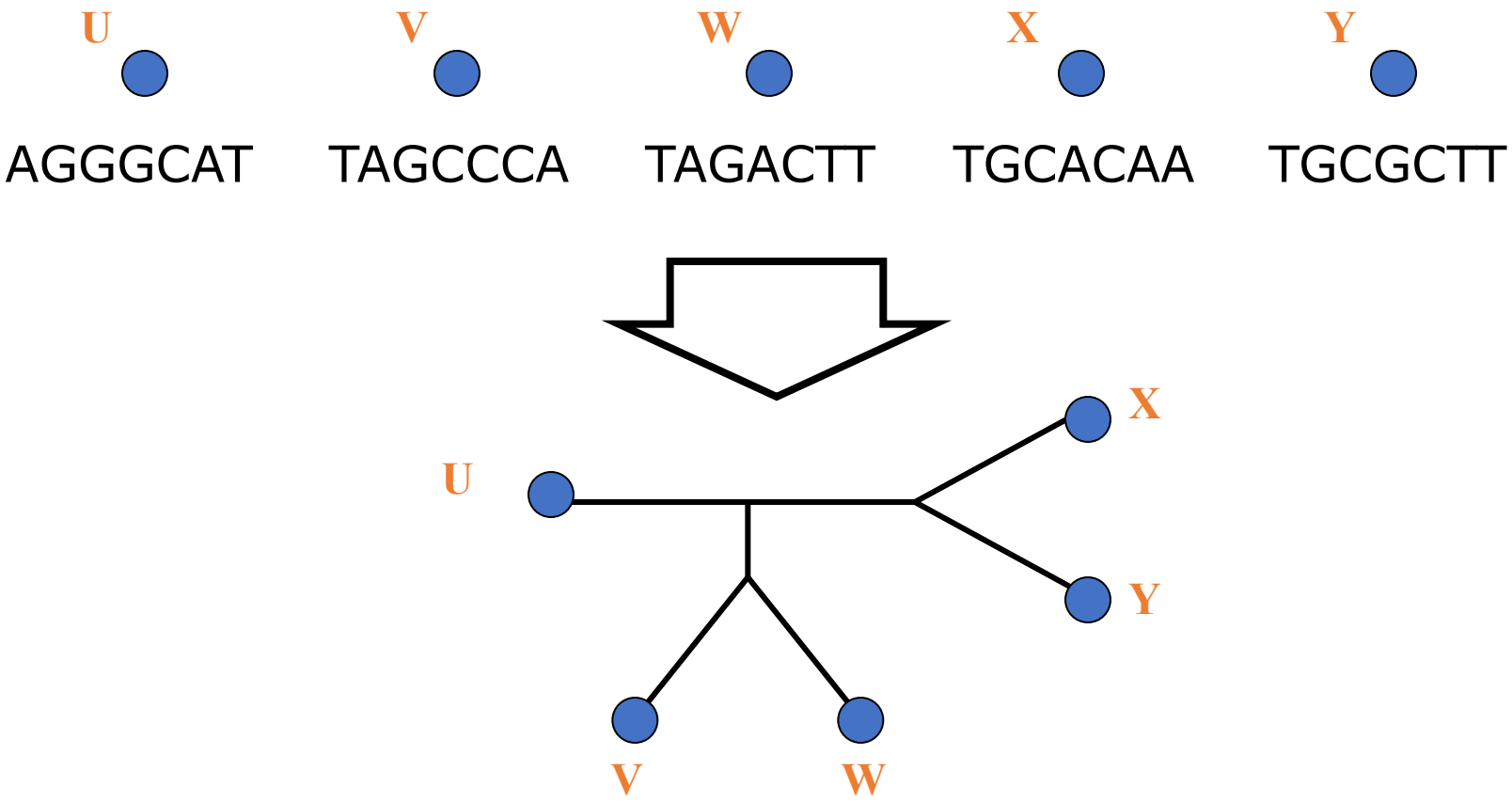
Approaches:

- NP-hard optimization problems and large datasets
- Statistical estimation under stochastic models of evolution
- Probabilistic analysis of algorithms
- Graph-theoretic divide-and-conquer
- Chordal graph theory
- Combinatorial optimization

DNA Sequence Evolution (Idealized)



Phylogeny Problem



Markov Models of Sequence Evolution (Gene Tree)

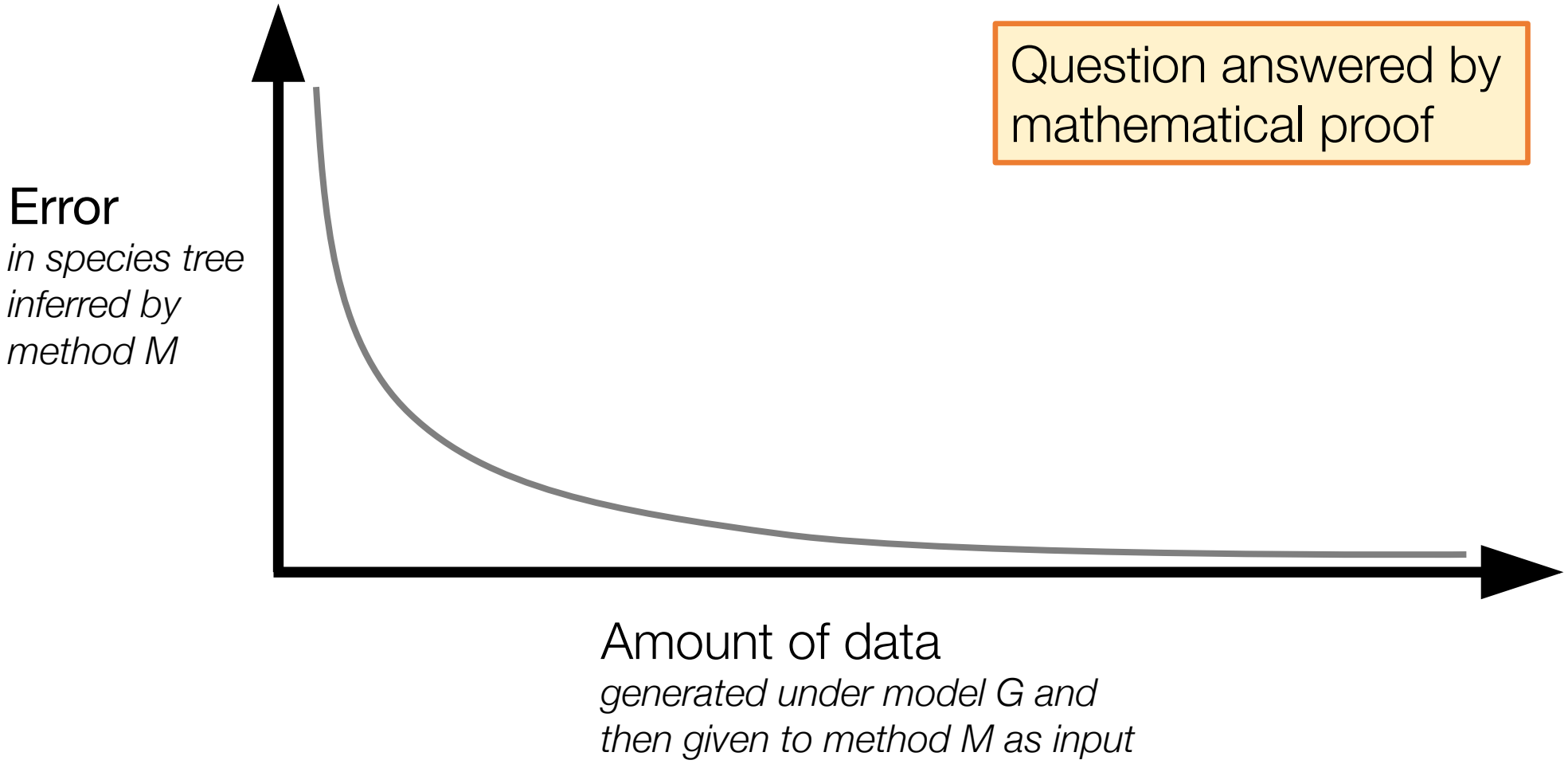
The different sites are assumed to evolve ***i.i.d***. down the model tree, so it suffices to model a single site

Jukes-Cantor, 1969 (simplest DNA site evolution model):

- The state at the root is randomly drawn from {A,C,T,G} (nucleotides)
- The model tree T is binary and has substitution probabilities p(e) on each edge e, with $0<p(e)<3/4$
- If a site (position) changes on an edge, it changes with equal probability to each of the remaining states
- The evolutionary process is Markovian.

More complex models are also considered, often with little change to the theory.

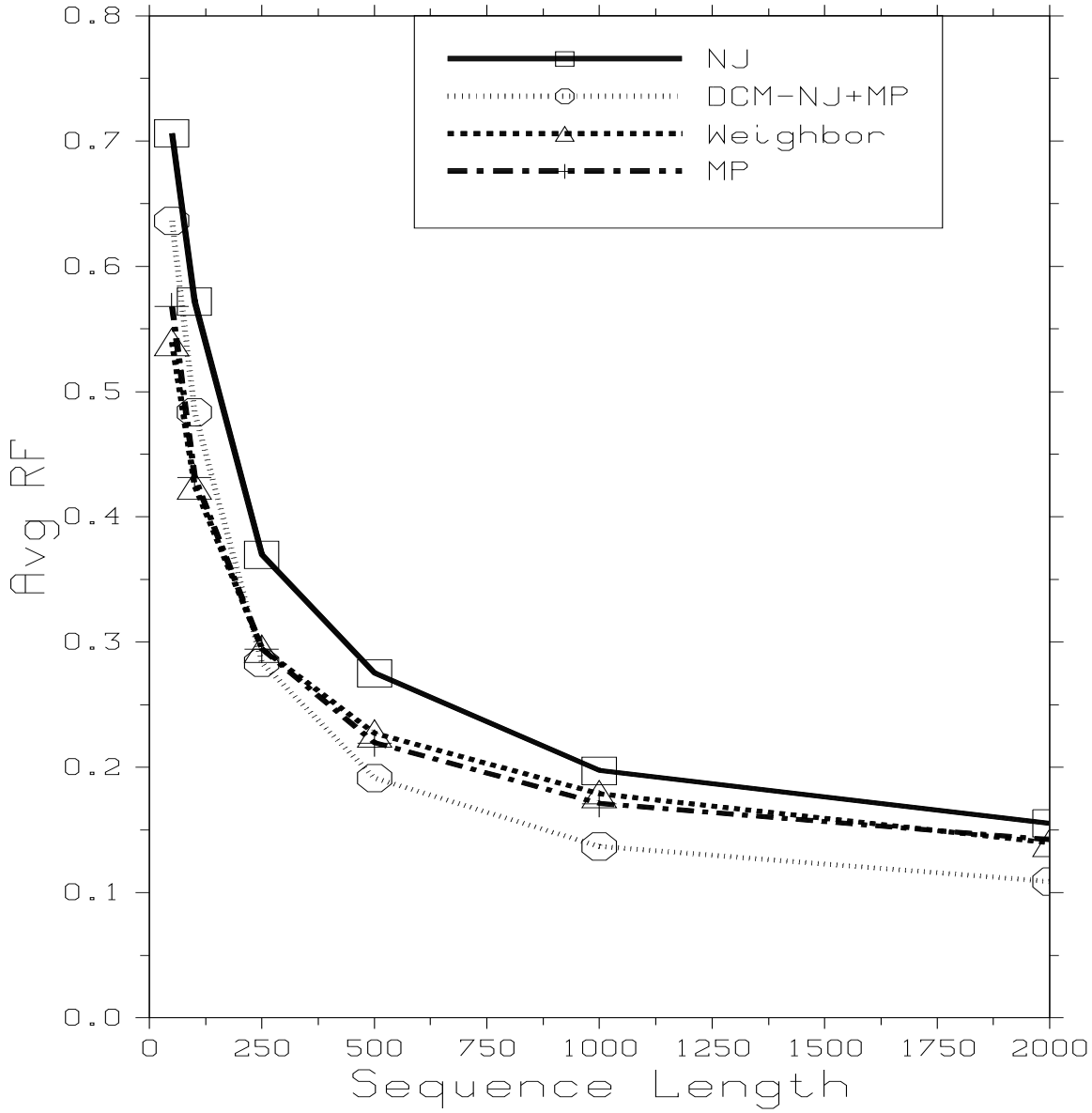
Is method M statistically consistent under model G?



Neighbor Joining vs Maximum Parsimony

- Neighbor joining (distance-based) is polynomial time and is proven statistically consistent
- Maximum parsimony (Hamming Distance Steiner Tree Problem) is NP-hard and statistically inconsistent

Therefore, we should use Neighbor Joining… right?



(a) 400 taxa

Methods:
NJ: Neighbor Joining (polytime and consistent)
DCM-NJ+MP: divide-and-conquer
Weighbour: polytime and consistent
MP: Maximum parsimony (NP-hard, inconsistent)

Y-axis: sequence length
Y-axis: Tree error rate (fraction of edges missing)

Note: NJ has higher error than MP on these data

So, predictions from theory do not seem to work!

Figure from Nakhleh et al., Pacific Symposium Biocomputing, 2002.

Why I love phylogenetics

- Phylogeny estimation is a non-trivial and complex statistical estimation problem
- Theory and empirical evaluation are both needed – and they inform each other.
- These insights lead to advances in methods, which in turn enable biologists to make more accurate scientific discoveries.