PHYLOGENY ESTIMATION: TRADITIONAL AND BAYESIAN APPROACHES

Mark Holder and Paul O. Lewis

Fenglin Liu 刘凤麟

School of life sciences, Peking University

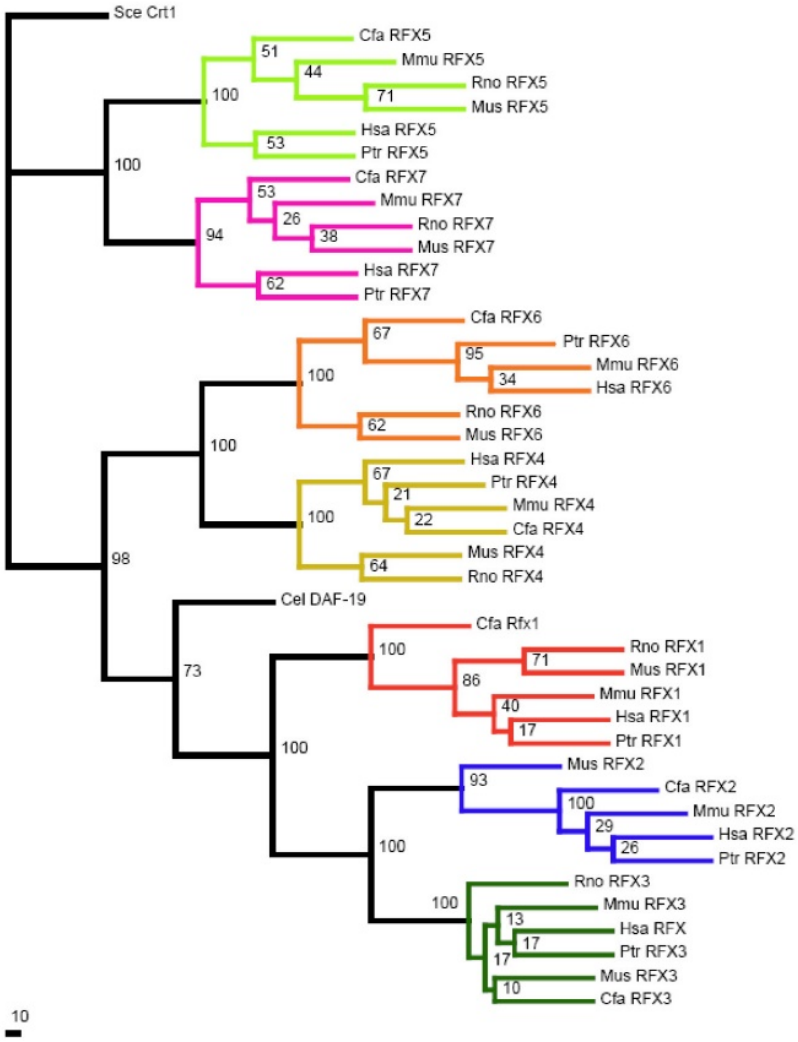
Outline

- What is phylogeny estimation?
- Why estimate phylogeny?
- How to estimate phylogeny?
  - Traditional approaches
  - Bayesian approaches

What is phylogeny?

- Phylogenetics: the study of evolutionary relationships among groups of organisms (e.g. species, populations) or genes, which are discovered through molecular sequencing data and morphological data matrices. (modified from Wikipedia)

- Phylogenetic tree: A graph depicting the ancestor–descendant relationships between organisms or gene sequences. The sequences are the tips of the tree. Branches of the tree connect the tips to their (unobservable) ancestral sequences.



Phylogenetic analysis of mammalian RFX genes.

The species names included in this figure are abbreviated. They are: Mus–mouse (*Mus musculus*); Rno–Rat (*Rattus norvegicus*); Cfa–dog (*Canis familiaris*);

Ptr–chimpanzee (*Pan troglodytes*); Mmu–monkey (*Macaca mulatta*) and Hsa–human (*Homo sapiens*).

Aftab *et al. BMC Evolutionary Biology* 2008 **8**:226

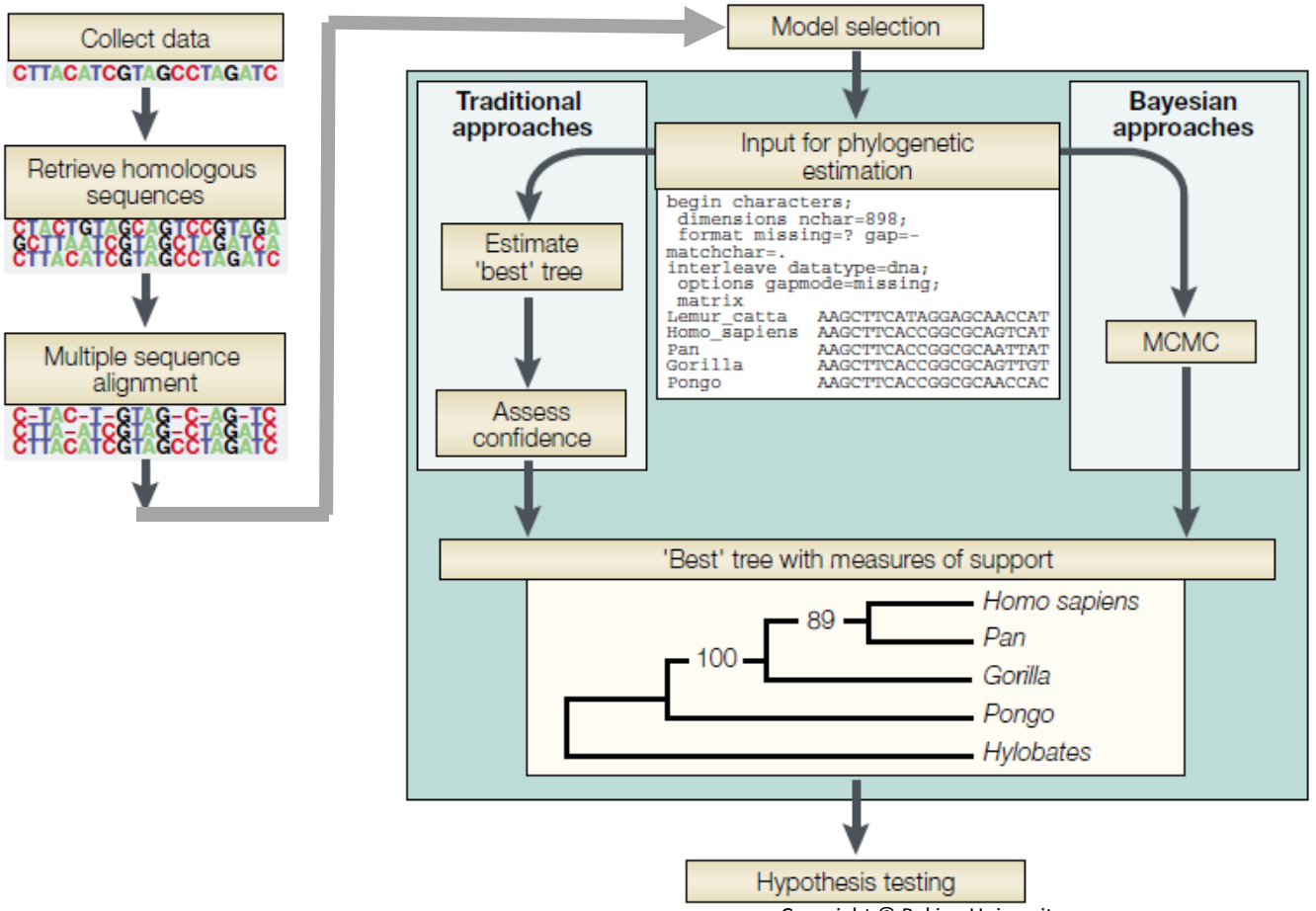
Why do phylogeny estimation?

- Detection of orthology and paralogy
- Estimating divergence times
- Reconstructing ancient proteins
- Finding the residues that are important to natural selection
- Detecting recombination points
- Identifying mutations likely to be associated with disease
- Determining the identity of new pathogens

How to estimate phylogeny?

- Assumption
  - As the time increases since two sequences diverged from their last common ancestor, so does the number of differences between them.
- Basic idea
  - Count the number of differences between sequences and group those that are most similar.
- Complexity
  - The rate of sequence evolution is not constant over time.
  - Natural selection or changing mutational biases exist.
  - Many of the sites in a DNA sequence are not helpful.

The phylogenetic inference process

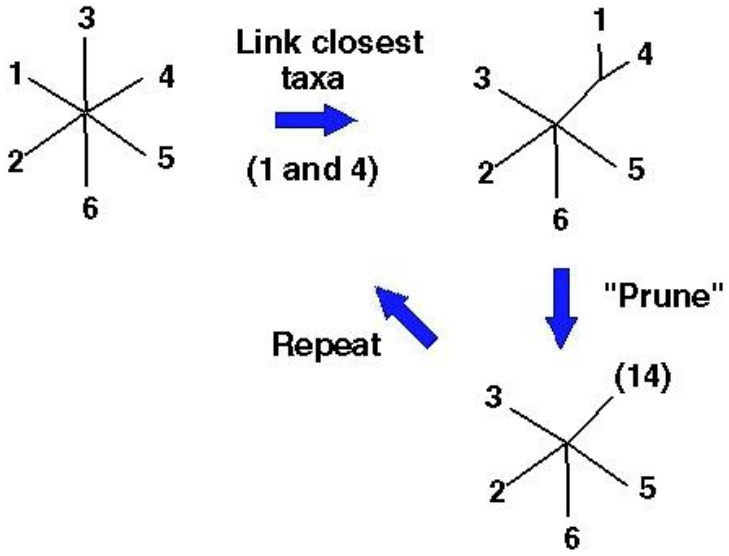


Traditional approaches

- Neighbour-joining (NJ) algorithm

- Tree searches that use an optimality criterion
  - Parsimony
  - maximum likelihood (ML)

Neighbour-joining



- Description
- Advantages
  - Fast
- Disadvantages
  - Information is lost in compressing sequences into distances.
  - Reliable estimates of pairwise distance can be hard to obtain for divergent sequences.
- Software
  - PAUP*
  - MEGA
  - PHYLIP

Parsimony

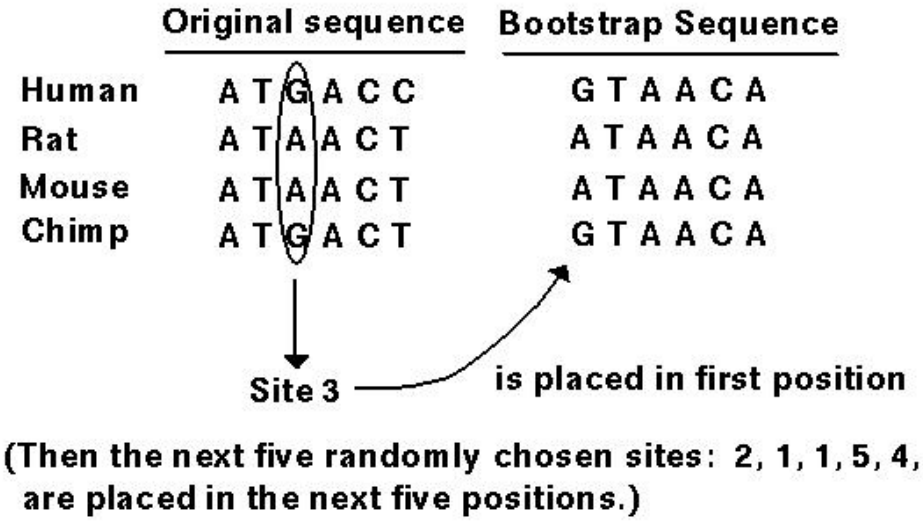
- Description
  - To determine the tree (or trees) that require the fewest number of mutations in order to explain the data that you have.
- Advantages
  - Fast enough for the analysis of hundreds of sequences.
  - Robust if branches are short (closely related sequences or dense sampling)
- Disadvantages
  - Can perform poorly if there is substantial variation in branch lengths.
- Software
  - PAUP*
  - NONA
  - MEGA
  - PHYLIP

Maximum likelihood

- Description
  - The tree that has the highest probability of producing the observed sequences $P(x_u^{\bullet}|T, t_{\bullet})$ is preferred.
- Advantages
  - The likelihood fully captures what the data tell us about the phylogeny under a given model.
- Disadvantages
  - Can be prohibitively slow (depending on the thoroughness of the search and access to computational resources)
- Software
  - PAUP*
  - PAML
  - PHYLIP

Assessing confidence — the bootstrap

- A high percentage of the bootstrap replicates implies that if another data set were collected, there is a good chance that the group would be recovered.
- Chief drawback: computational burden

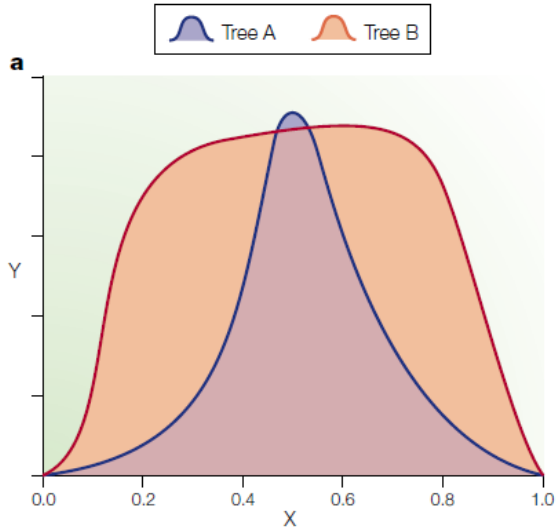


Hypothesis testing

- Use a phylogenetic analysis to determine whether an unknown virus belongs to ‘group A’ or ‘group B’.
- A tree with representatives of both candidate groups and the unknown sample is constructed, and the unknown sequence is intermingled with those from group A.
- The traditional approach involves finding the best tree in which the unknown sample clusters with the group B viruses, and then assessing how much worse this tree is compared to the best tree found in the original search.
- If the placement of the unknown with group B scores much worse than the optimal solution, then the data reject the possibility of the unknown sample actually belonging to group B.

Bayesian phylogenetics

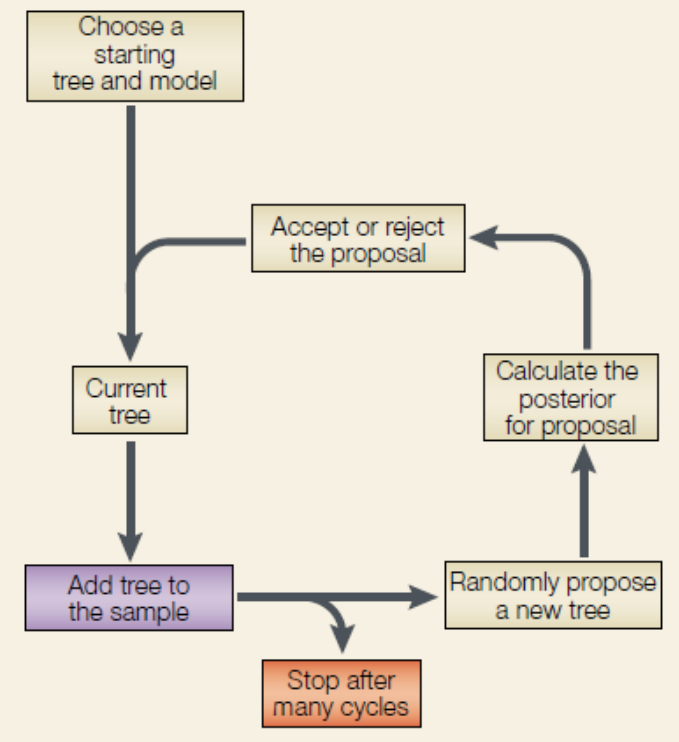
- Description
  - To maximize the posterior probability $P(\text{tree}|\text{data})$
  - $P(T, t_\bullet|x^\bullet) = \dfrac{P(x^\bullet|T, t_\bullet)P(T, t_\bullet)}{P(x^\bullet)}$
- Advantages
  - It has a strong connection to the maximum likelihood method.
  - The primary analysis produces measures of uncertainty.
  - It allows complex models of sequence evolution to be implemented.
  - It doesn't rely on the molecular clock assumption to estimate divergence times.
  - The nuisance parameters are integrated out(marginalized) to obtain the marginal posterior probability of a tree.

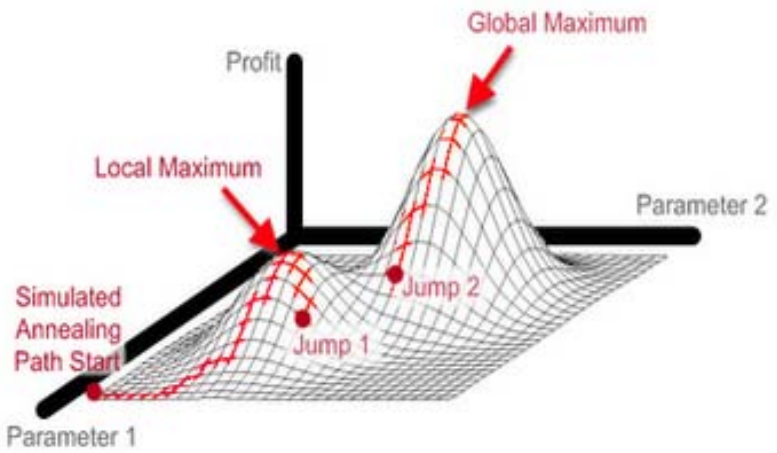


Bayesian phylogenetics

- Disadvantages
  - The prior distributions for parameters must be specified.
  - It can be difficult to determine whether the MCMC approximation has run for long enough.
- Software
  - MrBayes
  - BAMBE

Markov chain Monte Carlo





http://www.stanford.edu/~hwang41/

Conclusion

- The estimation of phylogenies has become a regular step in the analysis of new gene sequences.
- MCMC-based approaches are extending the field by answering previously intractable questions.
- These new techniques seem poised to teach us a great deal about the tree of life and molecular genetics.

References

- Holder M, Lewis P O. Phylogeny estimation: traditional and Bayesian approaches[J]. Nature reviews genetics, 2003, 4(4): 275-284.
- Aftab S, Semenec L, Chu J S C, et al. Identification and characterization of novel human tissue-specific RFX transcription factors[J]. BMC evolutionary biology, 2008, 8(1): 226.
- http://www.zoology.ubc.ca/~bio336/Bio336/Lectures/Lecture14/Overheads.html
- http://www.stanford.edu/~hwang41/
- Richard Durbin et al. 生物序列分析. 2010

Group members

- 熊罗星 (Discussion, revising ppt)
- 胡致远 (Discussion, revising ppt)
- 皮航宇 (Caption)
- 李子逸 (Discussion, revising ppt, caption)
- 张金阳 (Caption)
- 陈思雨 (Caption)
- 刘凤麟 (Discussion, making ppt, presentation)

Thank you!