SCIE2100/BINF6000 Learning Guide

Week 7: Phylogenetic analysis 1

Outline

- **Scientific motivation**
  - Sequence similarity indicates homology, which in turn implies a historical explanation
  - Evidence of evolution is based on combining existing biology and computational analysis
- **Background**
  - Darwin's On the origin of species (1859) and the concept of Natural selection over time
  - Haeckel's Phylogenetic tree and how it places an evolutionary origin (a common ancestor) around all organisms
  - Phylogenetic analysis requires taxa to be homologous
- **Concepts and terms**
  - Substitution, insertions, deletions and sequence rearrangement
  - Speciation and duplication
  - Homology (revisited), orthology and paralogy
  - Two explanations to why sequences align: evolutionary relationship OR chance
  - A tree is a graphical data structure, representing evolutionary relationships between taxa
  - Terminology: node/branch point, branch, root, leaf, distance, label, extant v ancestor
  - Trees can be unrooted, rooted, with distances undirected or directed in (evolutionary) time
  - Roots can be placed in an unrooted tree by using an outgroup
  - Trees can be ultrametric or additive, which are based on a universal molecular clock (implying a constant rate of evolution) or allow evolutionary change to be modelled additively, respectively
  - Species trees and gene trees
- **Quantifying sequence evolution**
  - Metric of distance v. model of change
  - Distances (p-distance, Poisson corrected distance, Gamma corrected distance)
  - Molecular clock
  - Changes in DNA
  - Evolutionary models of DNA (Jukes-Cantor, Kimura 2/3 parameter models, etc)
  - Rate matrix and probability matrix to model substitution in DNA and amino acids
  - Maximum likelihood for inference (first look)

Reflection

- *What is phylogeny and why is it relevant to science?*
- *How does conservation and by extension phylogenetic analysis provide insights into biomolecular structure and function?*

- *Define homology, orthology and paralogy; explain why paralogous genes are less likely to be similar, than orthologous genes, over the same time frame*
- *Why will phylogenetic analysis fail on non-homologous sequences?*
- *Given a tree, can you say if it is bifurcating? If it is bifurcating, how many internal nodes does it have if there are N leaves?*
- *What is a species tree? What is a gene tree? And what type of information do they represent?*
- *What type of evolutionary event explains why multiple taxa can come from the same genome?*
- *Define the p-distance*
- *What is the Poisson corrected distance correcting for?*
- *What is the Gamma corrected distance correcting for?*
- *Give two reasons why DNA mutations classified as transitions are more accepted by Nature, relative to transversions; also, see example exam question overleaf*
- *For DNA, if substitutions A→C, A→G and A→T happen at rates 0.2, 0.1 and 0.3, respectively, what is the value in the rate matrix for A→A?*
- *For proteins, in very broad terms, how can a rate matrix be created?*
- *What are the two independence assumptions made for Markov chains, for the purpose of modelling evolution? Hint: (aligned) sequences typically consist of multiple positions, and evolution happens over time.*
- *Challenge: Maximum likelihood for phylogenetic tree inference means what? (a) finding the most likely tree, given the sequence content at the leaves, or (b) finding the tree, that assigns the greatest likelihood to the observable sequence content*
- *For more, see example exam questions below*

Resources

- Textbook: Zvelebil M & Baum JO (2008) Understanding Bioinformatics, Garland Science T&F; **Chapter 7 (section 7.1-3), Chapter 8 (section 8.1)**

  The material in Zvelebil and Baum's book is excellent and goes into a fair level of detail. Slides are inspired and to some extent based on this book.

- Online textbook: Kelley & Didulo (2018) Computational Biology: a Hypertextbook; **Chapter 6;** https://search.library.uq.edu.au/primo-explore/fulldisplay?docid=61UQ_ALMA51202119940003131&context=L&vid=61UQ

  Kelley and Didulo's book is basic but serves to illustrate a few practical aspects of phylogenetic analysis.

- Koonin EV (2001) *Genome Biol* 2(4); https://www.ncbi.nlm.nih.gov/pmc/articles/PMC138920/
- Jensen RA (2001) *Genome Biol* 2(8); https://www.ncbi.nlm.nih.gov/pmc/articles/PMC138949/keke

  The two articles above are just for cursory reading and highlight some philosophical standpoints regarding the meaning of orthology and paralogy, clarifying speciation and duplication in trees.

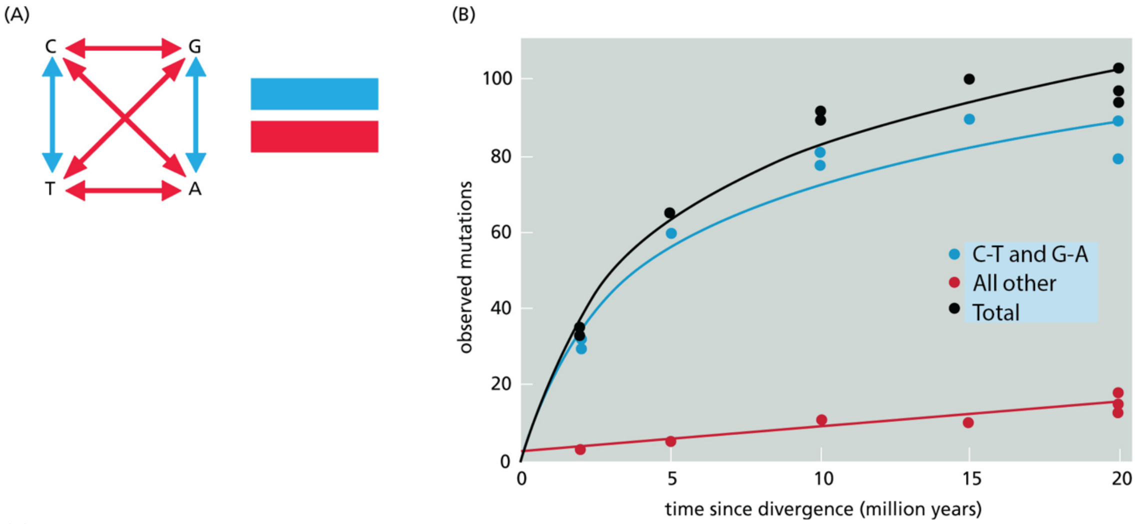
Wikipedia has some very good material on many topics mentioned in this section. If you want to browse Wikipedia pages a good starting point is

https://en.wikipedia.org/wiki/Computational_phylogenetics

Example exam QUESTION 1 PHYLOGENETICS (5 marks)

To model evolutionary changes between the four DNA bases A, G, C and T, it is possible to distinguish between two classes of substitutions as depicted in (A) below: between the two pyrimidines C and T, and between the two purines G and A (in blue), as opposed to all others (in red; changing a pyrimidine to a purine, or vice versa).

This problem is divided in two parts, each of which require you to study the Figure below.



(C)

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix} \quad Q = \begin{pmatrix} * & \alpha & \beta & \gamma \\ \alpha & * & \gamma & \beta \\ \beta & \gamma & * & \alpha \\ \gamma & \beta & \alpha & * \end{pmatrix} \quad Q = \begin{pmatrix} * & \pi_G & \pi_C & \pi_T \\ \pi_A & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \pi_T \\ \pi_A & \pi_G & \pi_C & * \end{pmatrix}$$

JC K81 F81

Part 1 (3 marks): In Figure (B), observed mutations of the blue class surpass those of the red class over evolutionary time. Identify all statements below that correctly explain the trends in (B):

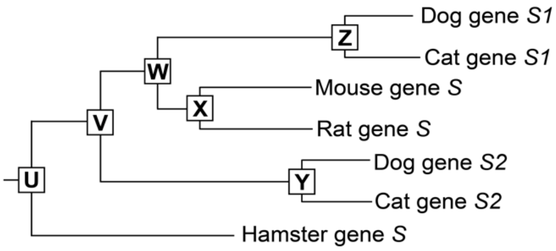
i. For protein-coding regions of the genome, relative to the blue class, substitutions of the red class are more likely to result in non-synonymous amino acid changes and are therefore less tolerated
ii. Relative to the red class, substitutions of the blue class tend to have smaller impact on the fold of the DNA and are therefore less likely to disrupt biological function
iii. The saturation of the observed number of substitutions from the blue class over time is explained by gradual lengthening of genome lengths
iv. The saturation of the observed number of substitutions from the blue class over time is explained by our inability to count actual changes when they occur over and over

Part 2 (2 marks): In Figure (C) three standard DNA rate matrices to model evolutionary change are given (from left to right) JC, K81 and F81; each matrix is specified by parameters as indicated in the elements (asterisks are calculated). (Rows specify source and columns specify target base, ordered as A, G, C and T.) Which of the three matrices has the capacity to distinguish between the red and blue classes of base change?

- JC, K81 and F81
- JC and K81
- K81 and F81
- JC
- K81
- F81
- None of them

Example exam QUESTION 2 PHYLOGENETICS (5 marks)

A. The phylogenetic tree below shows the evolutionary relationship of gene *S* in dog, cat, and rodents, rooted with the hamster gene *S* as outgroup. Genes *S1* and *S2* in the same organism are paralogs.



Which of the six nodes (U, V, W, X, Y or Z) describe the following two events? (2 marks)

I. speciation event between Rat and Mouse

II. duplication event between Dog gene *S1* and Dog gene *S2*

B. A *p*-distance is the proportion of sites at which two sequences differ. Based on the four sequences (W, X, Y and Z) below, complete the following *p*-distance matrix of these sequences. (3 marks)

W CAGCATATG

X CATCAACTA

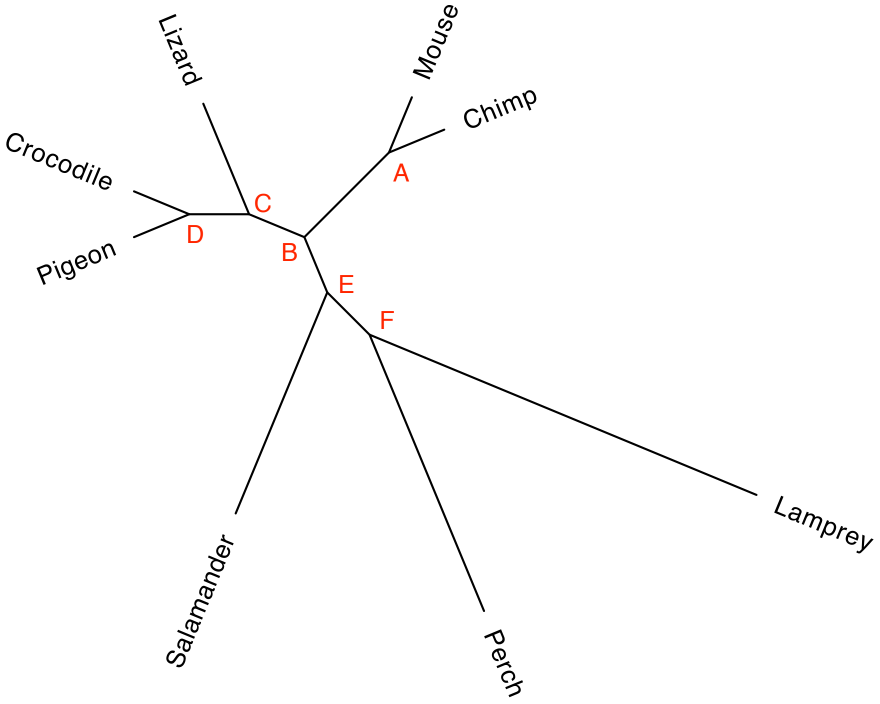
Y CAGCATTTC

Z CTTGTGAAC

	W	**X**	**Y**	**Z**
W	0.00			
X		0.00		
Y				
Z			0.78	0.00

p-distance matrix

Example exam QUESTION 3 PHYLOGENETICS (5 marks)



A. The *unrooted* phylogenetic tree above was inferred from eight orthologous sequences representing different species. Each internal branch point is labelled A-F, and the leaves are labelled with the species names; Lamprey and Perch are fish, the rest are land animals.

Based on this tree, answer the following questions.

I. To root the tree with lamprey as an outgroup, on which branch (A, B, C, D, E or F) should the root be placed?

II. Draw the rooted tree

B. To use molecular data to reconstruct evolutionary history requires making a number of reasonable assumptions. Which of the following statements are INCORRECT? Several statements could apply. (2 marks)

I. The molecular sequences used in phylogenetic construction are homologous.

II. The molecular sequences used in phylogenetic construction share a common origin.

III. Phylogenetic divergence cannot be bifurcating.

IV. Parent branch splits into two or more daughter branches at any given point.

V. The molecular sequences used in phylogenetic construction cannot be paralogous.

C. Which of the following properties of a phylogenetic tree are an indication of insufficient evidence to resolve a tree during phylogenetic inference? (1 mark)

I. Bifurcating branch points

II. Multifurcating branch points