Assignment 3 / Biometry and bioinformatics II / 2014

This set of questions is based on data_4.txt

Take a subsample of four animals from the data and by using this data, answer the following questions.

- A) How many conserved sites, how many variable sites, how many parsimony informative sites?

- B) What is relationship between nucleotide differences (among the four animals you are now comparing) at sites which do alter amino acid and those that do not alter amino acid (i.e. is there a difference between observed mutations leading to amino acid differences and those mutations that do not lead to amino acid difference).

- C) Pick up one variable site which is not parsimony informative and explain why it is not parsimony informative.

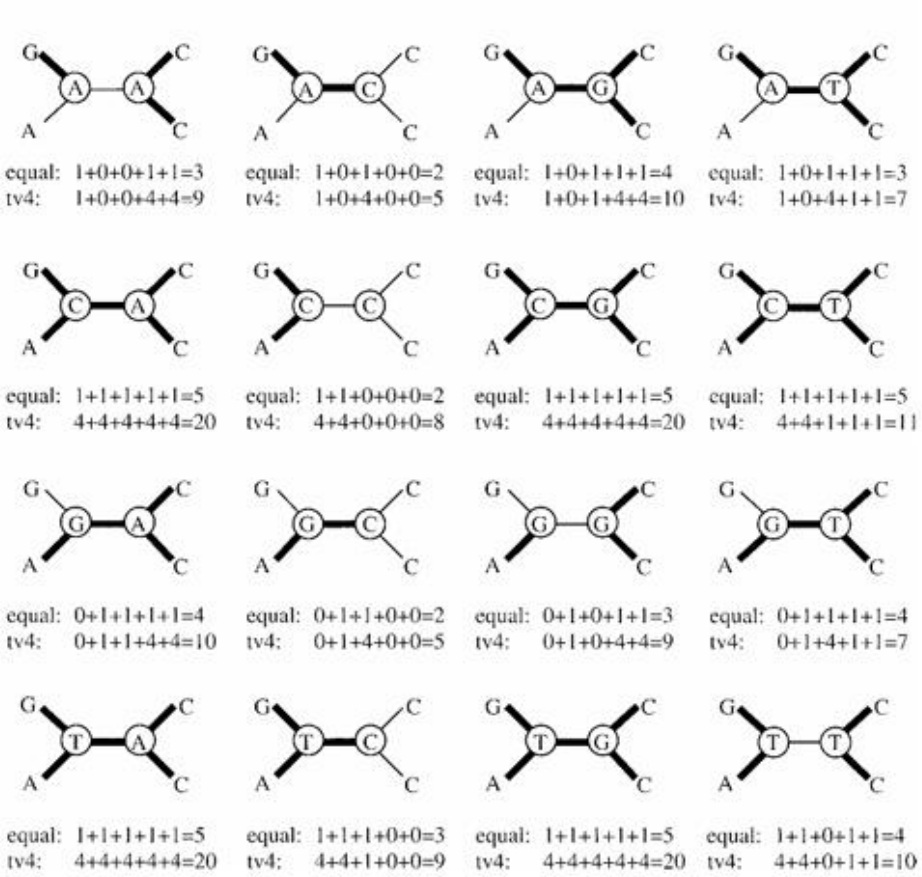
- D) Pick up three parsimony informative sites and construct the unrooted maximum parsimony trees on the basis of these sites. Then combine the information from these separate sites: what is the maximum parsimony tree?

    - N.B. You should use a set of animals which are not *too similar* in order to get a reasonable set for answering to these questions. How to do that? Inspect your clustering! If your first trial set does not include any parsimony informative sites, then your set is too simple and you should take another sample!

- E) Below (next page) is an example using a cost scheme: transversions are weighted 4x. Re-consider your D) (= what you did above). Would you get the same result or a different result by using this kind of a cost scheme?

- F) Calculate the distance matrix by using p-distance and by Jukes-Cantor. You can do this by using MEGA5-facilities. However, in addition to this distance matrix, show at least one calculation by hand (= show, for one species pair, how is their p-distance and Jukes-Cantor –distance calculated).

- G) By using the distance matrix you made in F) (either p-distance or JC, does not matter), construct UPGMA by hand, NOT by MEGA5.



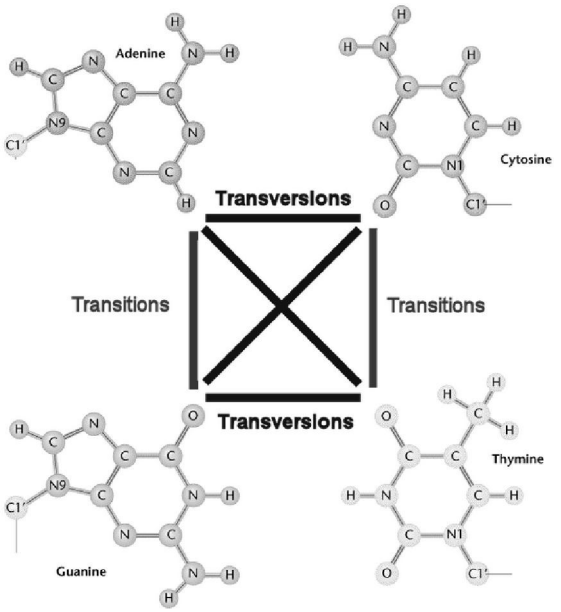
This picture is from *Lemey et al., The phylogenetic handbook, 2009,* www.cambridge.org/9780521877107

Two cost schemes:

- Equal vs. transversions 4x weighted. See next page for clarification on transitions and transversions.

- With equal costs, the minimum length in two steps and this length is achievable in three different ways: internal nodes assignment A-C, C-C and G-C. If a similar analysis for the other two possible trees, ((W,X),(Y,Z)) and ((W,Z),(Y,X)) is conducted, they are also found to have lengths of two steps. *Thus this character (state) does not discriminate among three tree topologies and is parsimony-uninformative under this cost scheme.*

- With 4:1 transversion:transition weighting the minimum length is five steps, achieved by two reconstructions: internal node assignments A-C and G-C. Similar evaluation of the other two trees finds a minimum of eight steps on both trees. This means that two transversions are required rather than one transition plus one transversion. *The character thus becomes informative as some trees have lower lengths than others.*



DNA substitution mutations are of two types:

Transitions are interchanges of two-ring purines (A ↔G) or of one-ring pyrimidines (C ↔T): they therefore involve bases of similar shape.

Transversions are interchanges of purine for pyrimidine bases, which therefore involve exchange of one-ring and two-ring structures.

It is well known that transitions are considerably more common than transversions.