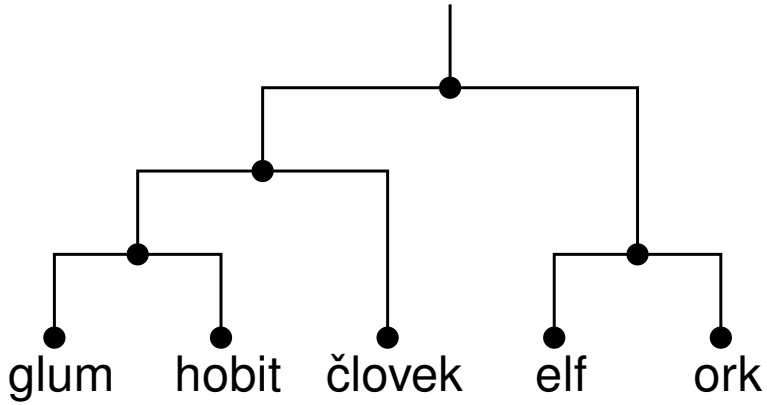
Phylogenetic trees (cvičenie)

Broňa Brejová

29.10.2020

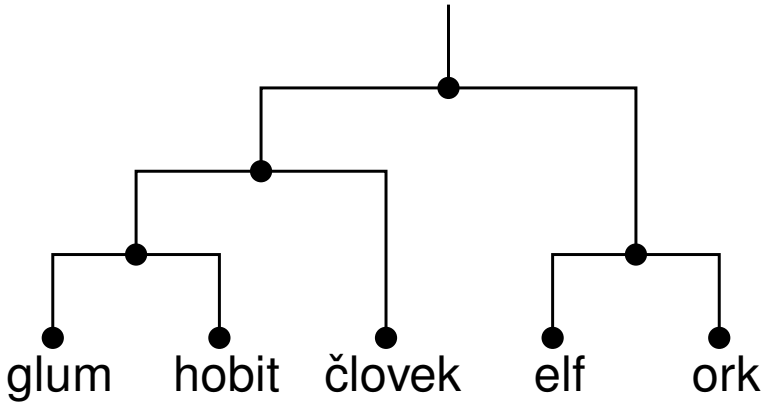
Terminology

- zakorenený strom, rooted tree
- nezakorenený strom, unrooted tree
- hrana, vetva, edge, branch
- vrchol, uzol, vertex, node
- list, leaf, leaf node, tip, terminal node
- vnútorný vrchol, internal node
- koreň, root
- podstrom, subtree, clade



Several facts about trees

- Consider a rooted tree with n leaves, in which each internal node has 2 children. Such a tree always has $n-1$ internal nodes and $2n-2$ branches (why?)
- Consider an unrooted tree with n leaves, in which each internal node has 3 neighbours. Such a tree always has $n-2$ internal nodes and $2n-3$ branches (why?)
- In how many ways can we root an unrooted tree with n leaves?



Unrooted trees

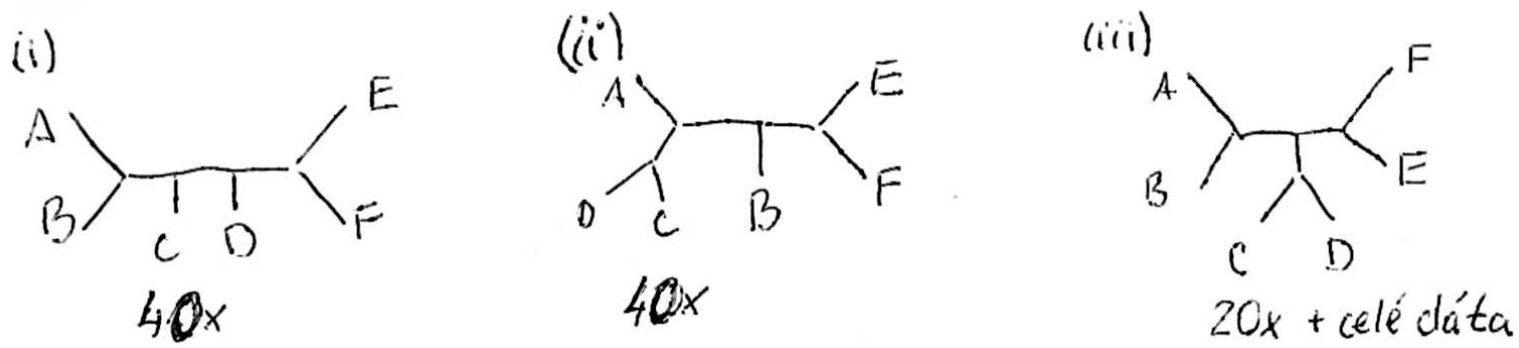
What can we say about relationships from a an unrooted tree of 4 species? Can we say that some two species X and Y are closer to each other than to everybody else?

Bootstrap

- Randomly select several alignment columns, build a tree
- Repeat many times
- Count how many times each branch appears in the trees (branch in an unrooted tree is a split of species into two groups)
- Finally build a tree from the original data and see how often was each branch in the replicates
- We can also build a tree directly from frequent branches
- Bootstrap values estimate confidence, particularly if we have little data (short alignment)
- If the data do not correspond to the assumptions of the used method/model, we can get an incorrect branch with a high bootstrap

Bootstrap

We did 100 bootstrap replicates, obtaining the following results:



Add bootstrap values to the tree (iii)

Which additional branches have support at least 20%?

What would the tree look like if we kept only branches with support at least 80%?

Probabilistic models

Probabilities refer to some thought experiment involving randomness (dice throws, drawing balls from an urn etc.)

We set up these thought experiments in a way that mimics some aspects of reality (properties of DNA sequences, evolution etc.)

The probabilities computed for the though experiment tell us something about the real world.

A famous quotation by statistician George Box "All models are wrong, but some are useful."

Probabilistic models used in the course so far

- Scoring matrices: compare the model of random sequences and related sequences
- E-value in BLAST: random database and query, how many matches with score T do we expect by chance?
- Gene finding: model generating random sequence and annotation. For a given sequence, what is its most probable annotation?
- Evolution, Jukes-Cantor model: model generating one column of an alignment.
  Unknown parameters: tree, branch lengths.
  For a given alignment, which parameters yield highest probability (likelihood) $\max_{param} \Pr(data|param)$

Jukes-Cantor model of substitutions

Probability of observing a change over branch of length t:

$$\Pr(C|A,t) = (1 - e^{-\frac{4}{3}t})/4$$

This applies to every pair of distinct nucleotides.

Probability of not observing a change over branch of length t:

$$\Pr(A|A,t) = (1 + 3e^{-\frac{4}{3}t})/4$$

This applies to every pair of identical nucleotides.

Both cases include also multiple unobserved changes happening at the same nucleotide.

More complex models of substitutions

Not all substitutions are equally frequent:

Transitions (within pyrimidines T<->C, within purines A<->G) are more frequent than transversions (A,G)<->(C,T)

Not all nucleotides are equally frequent in a genome (GC content)

These observations are captured in the HKY model (Hasegawa, Kishino, Yano)

HKY model

Substitution rate matrix (matica rýchlostí zmeny)

$$\begin{pmatrix} -\mu_A & \beta\pi_C & \alpha\pi_G & \beta\pi_T \\ \beta\pi_A & -\mu_C & \beta\pi_G & \alpha\pi_T \\ \alpha\pi_A & \beta\pi_C & -\mu_G & \beta\pi_T \\ \beta\pi_A & \alpha\pi_C & \beta\pi_G & -\mu_T \end{pmatrix}$$

$\kappa = \alpha/\beta$ is the ratio of transition and transversion rates

π_j is the frequency of base j

Rate of substitution from X to Y is the product of π_Y and a factor distinguishing transitions and transversions

The sum of each row is 0 ($\mu_A = \beta\pi_C + \alpha\pi_G + \beta\pi_T$)

The matrix is normalized so that the expected number of substitutions per unit of time is 1

HKY model

Substitution rate matrix

$$\begin{pmatrix} -\mu_A & \beta\pi_C & \alpha\pi_G & \beta\pi_T \\ \beta\pi_A & -\mu_C & \beta\pi_G & \alpha\pi_T \\ \alpha\pi_A & \beta\pi_C & -\mu_G & \beta\pi_T \\ \beta\pi_A & \alpha\pi_C & \beta\pi_G & -\mu_T \end{pmatrix}$$

The matrix has 4 parameters $\kappa = \alpha/\beta$ and three frequencies; we also need time t

More complex model better represents real processes, but we need more data to estimate more parameters

There are many other models with higher or lower number of parameters

Substitution models

Substitution rate matrix (e.g. HKY)

$$\begin{pmatrix} -\mu_A & \beta\pi_C & \alpha\pi_G & \beta\pi_T \\ \beta\pi_A & -\mu_C & \beta\pi_G & \alpha\pi_T \\ \alpha\pi_A & \beta\pi_C & -\mu_G & \beta\pi_T \\ \beta\pi_A & \alpha\pi_C & \beta\pi_G & -\mu_T \end{pmatrix}$$

We have methods for computing $\Pr(Y|X,t)$ for given X, Y, t, and matrix

For example, if ϵ is a very short time, $\Pr(C|A,\epsilon)$ is roughly $\epsilon\beta\pi_C$

This is not true for reasonably long time intervals, therefore we use algebraic methods considering also multiple substitutions at the same nucleotide.