Selecting models of nucleotide sequence evolution

Comparative Genomic Analyses

Rui Borges
Vetmeduni Vienna

Models of sequence evolution

Exercise
Consider the following DNA sequences:

```
           1          11         21         31         41
Sequence 1 GCTAAGAGAG CATACTCTAC GATCTCTAAC CGGTCAGTAT CCACAGATCT
Sequence 2 TGTCCCTCGG CATAGACTTG CAAGTGTCTC GTGTCAGTCG CCTCCGCTCG

           51         61         71         81         91
           CGGACTATTC TATCTGACTA TAGTGTGGCC ACTATTTAGC CCGCTTTTGC
           CTGAATATAA TAACTAACGA TATTGGGGCC ACTATATTGT GAGGATTGGG
```

Use this data to estimate the transition matrix Q.
Which model of the nucleotide substitution models, JC69 or K81, best explain this data?

Selecting models of nucleotide evolution

An attractive procedure to select a model of evolution would be the use of the most complex parameter-rich one: e.g., GTR or UNREST.

- a large number of parameters need to be estimated
- the analyses become computationally difficult
- more error is introduced in each estimate

Likelihood ratio test

The fit of a model can be measured through the **likelihood function**.

- the likelihood expresses the probability of observing the data D, given a model of evolution M and a phylogeny T:

$$L = p(D|\theta, \tau)$$

Likelihood ratio test

Some models of evolution, or some phylogenies, will be more likely than others in explaining the data: **maximum likelihood estimates** (MLE)

- the values of the model parameters, the topology and branch lengths that make the likelihood functions as large as possible

$$\hat{\theta}, \hat{\tau} = \max p(D|\theta, \tau)$$

- for computational reasons we work with the maximized log likelihood

$$\ell = \log p(D|\hat{\theta}, \hat{\tau})$$

Likelihood ratio test

A standard way of comparing the fit to two models of evolution is to contrast their log likelihoods using the **likelihood ratio test** (LRT) statistic:

$$LRT = 2(\ell_1 - \ell_0)$$

- ℓ_1: the maximum likelihood under the parameter richer complex model
- ℓ_0 the maximum likelihood under the simpler model

Likelihood ratio test

When the models compared are nested (i.e., the simple model is a special case of the complex model):

$$LRT \xrightarrow{D} \chi^2(\nu)$$

- $\chi^2(\nu)$ distribution with ν degrees of freedom
- ν equal to the difference in the number of free parameters between the two models

Likelihood ratio test

The χ^2 distribution can be used to select the model of evolution that best fits the data

- LRT $>$ critical value
  the inclusion of additional parameters increases the likelihood of observing the data significantly: the most complex model is favored
- LRT $<$ critical value
  the simpler model is favored

	probability of the upper tail					
df	0.1	0.05	0.025	0.01	0.005	0.001
1	2.7055	3.8415	5.0239	6.6349	7.8794	10.827
2	4.6052	5.9915	7.3778	9.2103	10.596	13.815
3	6.2514	7.8147	9.3484	11.344	12.838	16.266
4	7.7794	9.4877	11.143	13.276	14.860	18.466
5	9.2364	11.070	12.832	15.086	16.749	20.515

Likelihood ratio test

Exercise

The likelihood of two models of evolution (JC69 and K80) was calculated for a multiple sequence alignment.

Model	lnL	np
JC69	-18 562.85	44
K80	-18 551.66	45

Perform an LRT and determine the model of evolution that best describes these sequences?

Hierarchical LRT

It is typically the case that models of sequence evolution are **nested**:

- one model being equivalent to a restriction of one or more parameter values of a more complex model

K80

$$\begin{bmatrix} . & \alpha & \beta & \alpha \\ \alpha & . & \alpha & \beta \\ \beta & \alpha & . & \alpha \\ \alpha & \beta & \alpha & . \end{bmatrix}$$

HKY85

$$\begin{bmatrix} . & \alpha\pi_C & \beta\pi_G & \alpha\pi_T \\ \alpha\pi_A & . & \alpha\pi_G & \beta\pi_T \\ \beta\pi_A & \alpha\pi_C & . & \alpha\pi_T \\ \alpha\pi_A & \beta\pi_C & \alpha\pi_G & . \end{bmatrix}$$

Hierarchical LRT

Exercise
Are the K80 and the F81 models of evolution nested? Justify

$$K80 \qquad \begin{bmatrix} . & \alpha & \beta & \alpha \\ \alpha & . & \alpha & \beta \\ \beta & \alpha & . & \alpha \\ \alpha & \beta & \alpha & . \end{bmatrix}$$

$$F81 \qquad \begin{bmatrix} . & \pi_C & \pi_G & \pi_T \\ \pi_A & . & \pi_G & \pi_T \\ \pi_A & \pi_C & . & \pi_T \\ \pi_A & \pi_C & \pi_G & . \end{bmatrix}$$

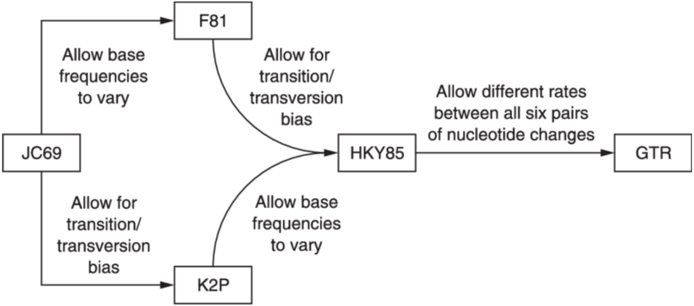
Hierarchical LRT

Several hypotheses can be tested hierarchically to select the best fitting model of evolution:

- are the base frequencies equal?
- is there a transition/transversion bias
- are the transition rates equal?
- are there invariable sites?
- is there rate homogeneity among sites

Hierarchical LRT

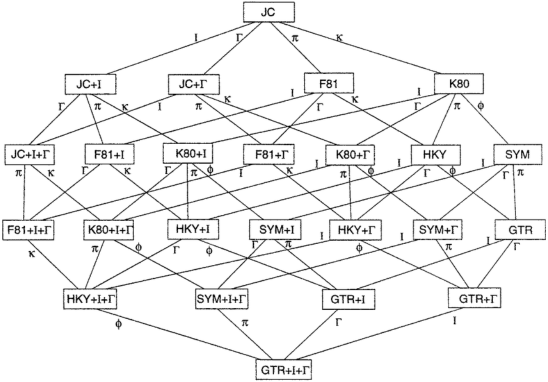
When comparing two nested models through an LRT, we test the different assumptions made by these models about our sequences.


F81
Allow base frequencies to vary
Allow for transition/ transversion bias
Allow different rates between all six pairs of nucleotide changes
JC69
HKY85
GTR
Allow for transition/ transversion bias
Allow base frequencies to vary
K2P

Hierarchical LRT

There are some potential problems with the use of pairwise LRTs in a hierarchical LRT approach.

- many hierarchies of LRTs are possible, and they can result in different models being selected



Literature

The Phylogenetic Handbook by Lemey, Salemi and Vandamme (2009)
Cambridge University Press

- Chapter 10: sections 10.1, 10.2 and 10.3