IMPRS Workshop: Comparative Genomics

Detecting positive seletion

Excercise

Software (all free)

Bioedit: (http://www.mbio.ncsu.edu/bioedit/bioedit.html) For visualizing and manually editing sequence alignments. The program also provides basic tools for alignment and sequence analyses.

Dnasp: (http://www.ub.edu/dnasp/) for basic analyses of nucleotide alignments.

PAML: (http://abacus.gene.ucl.ac.uk/software/paml.html) The software package Phylogenetic Analysis by Maximum Likelihood contains different programs for analyses of sequence alignments using maximum likelihood

Dataset (from Marshall et al, 2011. Plant Physiol)

Mg1LysM nucleotide alignment (9 sequences)

Mg3LysM nucleotide alignment (9 sequences)

Questions:

1) What is the number of synonymous mutations in the two genes? And the number of non-synonymous mutations?

2) What are the number of non-synonymous and synonymous sites and the number of non-synonymous and synonymous mutations? Use these parameters to calculate Nei and Gojobori's rates of P_N and P_S and the ratio P_N/P_S.

3) Are mutations in the genes randomly distributed?

Methods

1) Begin with the program Bioedit to open the two sequence alignments. Translate the nucleotide sequences into amino acid sequences.

2) Generate an alignment of the transcribed nucleotide sequence and save it in either phylip of fasta format (what is by the way the difference?).

3) Next step is to open the sequences in Dnasp. First define coding region and your population Mg (all nine sequences must be categorized into one group). Now calculate the basic parameters:

Number of haplotypes: The number of different sequence types

Number of segregating sites: The number of polymorphic sites

Total number of mutations: Includes both nucleotide polymorphisms and indel mutations.

Nucleotide diversity Pi: A measure of diversity in the sequence alignment. Takes into account the frequencies of the individual mutations.

Generate a sliding window plot. Is the variable sites randomly distributed?

4) For the coding sequence:

What is the number of non-synonymous sites?

What is the number of synonymous sites?

What is the number of non-synonymous mutations?

What is the number of synonymous mutations?

5) Based on these counts calculate Pn and Ps and the Pn/Ps ratio and discuss if your finding is consistent with positive selection?

6) Visualize the distribution of non-synonymous mutations in sliding window plot

Inference of positive selection with PAML

PAML (Phylogenetic Analysis with Maximum Likelihood) is a software package written by Ziheng Yang in 1993, and further developed in the 90s and years 20**. We will in this course focus on the program called codeml, dedicated to protein sequence analysis.

Codeml is a command line program, meaning that it has to be run from a terminal (under windows : Start => System tools =>DOS shell). In order to work, it requires 3 files :

- A text file containing the sequences, in a dedicated format
- A tree file, describing the relationships between the sequences, as a text file in the so-called parenthetic (aka New Hampshire or Newick) format
- A text file containing several options for tuning the program.

The option file is the most important one. It contains links toward the two others. It contains a series of options in the form 'tag = value'. Everything following a '*' is ignored by the program.

The tree file is either :

- A phylogeny, in case sequences are from different species. The tree therefore depicts the history of species.
- A genealogy, in case sequences are from different individuals of the same species. The tree then depicts the closeness between all individuals.

In many cases the tree is inferred from the sequences themselves, which codeml can do. However, it is rather inefficient for that task, and it is recomended to use external software for this purpose. We here used a software called PhyML, and we provide the corresponding trees for the two data sets.

Estimating omega = dN/dS with maximum likelihood.

The basic function of codeml is its ability to compute the likelihood of a large set of models. We will demonstrate this using the first dataset and a homogeneous model (all positions share the same dN/dS).

1) Go to directory Mg1/M0_profile
2) Edit the codeml.ctl file with a text editor (e.g. Notepad)
3) Find the line setting the initial value of omega, and set it to 1.0, and save the file
4) Using the DOS terminal, 'cd' to the M0_profile directory, and launch codeml by typing 'codeml' in the command line.
5) Codeml outputs several statistics... the interesting one here is the one stating lnL = - ????.

This corresponds to the value of the log likelihood. Note it with its corresponding omega value (for example in Excel).

6) Reproduce steps 3 to 5 with different values for omega. What is the value of omega which maximizes lnL?

Fitting a model with codeml

Finding the value which maximize the likelihood function is in most cases tedious, as several parameters have to be jointly estimated. Luckily codeml can do that for us...

1) Go to directory Mg1/M0. The codeml.ctl file here has been modified to let codeml find the best parameter values. The modifications are :

fix_kappa = 0

fix_omega = 0

and the line fix_blength = -1 is commented (a '*' was added at the beginning of the line).

2) Run codeml. In the output stands the value for omega (noted w), togetherwith the estimated value sof dN and dS, among others. Codeml also gather results in a file names mlc.txt, which you can open in a text editor.
3) Discuss the value of omega obtained
4) Repeat the procedude for the second data set (Mg3). Compare the value of omega obtained.

More complex models and hypothesis testing

Codeml can fit several types of models. Here we will consider to model which allow each position to chose between distinct selection regime. In the first model, called M1a, each position in the alignment can chose between a value of omega = 1 (neutral evolution) and a value of omega < 1 (purifying selection), which is estimated by the program. This model therefore has 1 additional parameter than the homogeneous one (called M0), namely the

proportion of positions with omega < 1. The second model is called M2a, and allows each position to chose between three possible values for omega, namely omega = 1 (neutral evolution), omega < 1 (purifying selection) and omega > 1 (positive selection), the two latter ones being estimated. This model therefore contains two more parameters than M1a, the proportion of positively selected sites and their corresponding omega value. We will fit these two models to our data :

1) cd to the M1a-M2a directory. The codeml.ctl file has been modified to run the M1a and M2a models instead of M0 (line Nssites = 1 2).
2) run codeml
3) edit the mlc.txt file. Note the lnL values for both models, and the corresponding parameter estimates.
4) Repeat steps 1 to 3 for the second data set. In both cases, which model has the better likelihood ?

To assess whether the likelihood improvement is significant, we use a likelihood ratio test (LRT). The statistics of the test is two times the logarithm of th ratio of the maximum likelihoods of the two models :

$S = 2*\ln(L1/L0) = 2*(\ln L1 - \ln L0)$

L0 being the maximum likelihood of the null model (in our case M1a), and L1 the maximum likelihood of the alternative model (in our case M2a). The resulting S values has to be compared with a chi sqare distribution with n1 - n0 degrees of freedom, n0 and n1 being the number of parameters of the null and alternative models, respectively. The chi2 helper program of the PAML package provides significance thresholds for the chisquare distribution.

5) Perform the likelihood ratio test for both data sets. Conclusions ?

In case positive selection is significant, one can look for candidate positions in the alignment. This is done by computing the probability for each position that its omega is greater than one (Bayes Empirical Bayes approach, BEB). This probability is not a p-value ! Position with a probability greater than 0.9 are generally considered significant and are good candidates for being under positive selection. BEB results are output in the mlc.txt files.

Inferring selection at the divergence level

TODO, need to gather the data set. Only for those who have time.

Additional information

Mg1LysM:	**Mg3LysM**
Exon 1-82	Exon 1-106
Intron 83-134	Intron 107-165
Exon 135-237	Exon 166-297
Intron 238-308	
Exon 309-374	
Intron 375-435	
Exon 436-478	

References:

Marshal et al. 2011. Analysis of Two in Planta Expressed LysM Effector Homologs from the Fungus *Mycosphaerella graminicola* Reveals Novel Functional Properties and Varying Contributions to Virulence on Wheat. Plant Physiol. *156: 756-769*

Nei and Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. *3 (5): 418-426.*

Yang and Nielsen. 2000. Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. *Mol Biol Evol 17 (1): 32-43*