Evosite3D

Friday, 16 September 2011

Identifying positive selection in genomic sequences

In this post, I will make a short tutorial on one of my favourite programs, CodeML, which is definitely not the easiest to use.

Theoretical principles:

The selective pressure in protein coding genes can be detected within the framework of comparative genomics. The selective pressure is assumed to be defined by the ratio (ω) dN/dS. dS represents the synonymous rate (keeping the amino acid) and dN the non-synonymous rate (changing the amino acid). In the absence of evolutionary pressure, the synonymous rate and the non-synonymous rate are equal, so the dN/dS ratio is equal to 1. Under purifying selection, natural selection prevents the replacement of amino acids, so the dN will be lower than the dS, and dN/dS < 1. And under positive selection, the replacement rate of amino acid is favoured by selection, and dN/dS > 1.

CodeML and substitutions models:

CodeML is a program from the package PAML, based on Maximum Likelihood, and developed in the lab of Ziheng Yang, University College London.
It estimates various parameters (Ts/Tv, dN/dS, branch length) on the codon (nucleotide) alignment, based on a predefined topology (phylogenetic tree).

Different codon models exist in CodeML. The model 0 estimates a unique dN/dS ratio for the whole alignment. Not really interesting, except to define a null hypothesis to test against. The branch models estimate different dN/dS among lineages (ie ASPM, a gene expressed in the brain of primates). The site models estimate different dN/dS among sites (ie in the antigen-binding groove of the MHC). The branch-site models estimate different dN/dS among sites and among branches. It can detect episodic evolution in protein sequences, as in the interactions between chains in the avian MHC. In my opinion, this is the most powerful application and this is the one used in the Sectome database (to which I contributed during my PhD).

First, we have to define the branch where we think that position could have occurred. We will call this branch the "foreground branch" and all other branches in the tree will be the "background" branches. The background branches share the same distribution of ω = dN/dS value among sites, whereas different values can apply to the foreground branch.

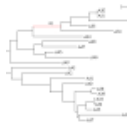
To compute the likelihood value, two models are computed: a null model, in which the foreground branch may have different proportions of sites under neutral selection to the background (i.e. relaxed purifying selection), and an alternative model, in which the foreground branch may have a proportion of sites under positive selection.
As the alternative model is the general case, it is easier to present it first.

Four categories of sites are assumed in the branch-site model:

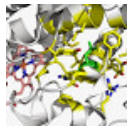
Sites with identical dN/dS in both foreground and background branches:
K0 : Proportion of sites that are under purifying selection (ω0 < 1) on both foreground and background branches.
K1 : Proportion of sites that are under neutral evolution (ω1 = 1) on both foreground and background branches.

Sites with different dN/dS between foreground and background branches:
K2a: Proportion of sites that are under positive selection (ω2 ≥ 1) on the foreground branch and under purifying selection (ω0 < 1) on background branches.
K2b: Proportion of sites that are under positive selection (ω2 ≥ 1) on the foreground branch and under neutral evolution (ω1 = 1) on background branches.

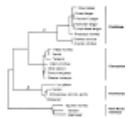
For each category, we get the proportion of sites and the associated dN/dS values.



In this post, I will make a short tutorial on one of my favourite programs, CodeML, which is definitely not the easiest to use. Theoretica...



Tutorial: estimating the stability effect of a mutation with FoldX
Introduction: Here is a brief tutorial on how to use FoldX to estimate the stability effect of a mutation in a 3D structure. The stabilit...



Tutorial on Ancestral Sequence Reconstruction
This tutorial was part of a course on protein evolution done during ECCB 2014 in Strasbourg: http://www.eccb14.org/program/tutorials/pea ...

Follow by Email

Email address... Submit

Subscribe to

Posts

Comments

Search This Blog

Search

Followers

In the null model, the dN/dS (ω2) is fixed to 1:

Sites with identical dN/dS in both foreground and background branches:
K0 : Sites that are under purifying selection (ω0 < 1) on both foreground and background branches.
K1 : Sites that are under neutral evolution (ω1 = 1) on both foreground and background branches.

Sites with different dN/dS between foreground and background branches:
K2a: Sites that are under neutral evolution (ω2 = 1) on the foreground branch and under purifying selection (ω0 < 1) on background branches.
K2b: Sites that are under neutral evolution (ω2 = 1) on the foreground branch and under neutral evolution (ω1 = 1) on background branches.

For each model, we get the log likelihood value (lnL1 for the alternative and lnL0 for the null models), from which we compute the Likelihood Ratio Test (LRT).
The 2×(lnL1-lnL0) follows a χ^2 curve with degree of freedom of 1, so we can get a p-value for this LRT.

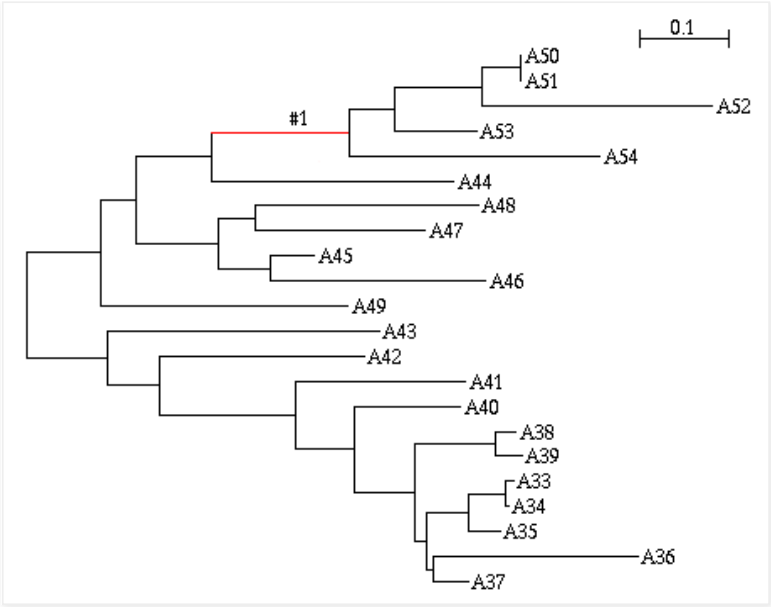
Let's go in details.

File Preparation:

We need four files to run CodeML:

1) The multiple nucleotide (CDS) alignment, in PHYLIP format. CodeML will strictly remove any position that contains at least one gap or an unknown "N" nucleotide:
TF105351.Eut.3.phy

2) The phylogenetic tree in newick format, with the branch of interest specified by "#1"(You can view it with NJplot or FigTree): TF105351.Eut.3.53876.tree



3) A command file where all parameters to run CodeML under the alternative model are specified: TF105351.Eut.3.53876.ctl

4) A command file where all parameters to run CodeML under the null model are specified: TF105351.Eut.3.53876.fixed.ctl

Execute CodeML

Run command file (alternative model):

We estimate the Ts/Tv ratio (fix_kappa = 0) and the dN/dS (fix_omega = 0). The branch-site model is specified by setting the model parameter to 2 (different dN/dS for branches) and the NSsites value to 2 (which allows 3 categories for sites: purifying, neutral and positive selection).

```
      seqfile = TF105351.Eut.3.phy           * sequence data file name
     treefile = TF105351.Eut.3.53876.tree    * tree structure file name
      outfile = TF105351.Eut.3.53876.mlc     * main result file name



        noisy = 9   * 0,1,2,3,9: how much rubbish on the screen
      verbose = 1   * 1: detailed output, 0: concise output
      runmode = 0   * 0: user tree;  1: semi-automatic;  2: automatic
                    * 3: StepwiseAddition; (4,5):PerturbationNNI; -2: pairwise

      seqtype = 1   * 1:codons; 2:AAs; 3:codons-->AAs
    CodonFreq = 2   * 0:1/61 each, 1:F1X4, 2:F3X4, 3:codon table
        clock = 0   * 0: no clock, unrooted tree, 1: clock, rooted tree
       aaDist = 0   * 0:equal, +:geometric; -:linear, {1-5:G1974,Miyata,c,p,v}
        model = 2   * models for codons:
                    * 0:one, 1:b, 2:2 or more dN/dS ratios for branches
      NSsites = 2   * 0:one w; 1:NearlyNeutral; 2:PositiveSelection; 3:discrete;
                    * 4:freqs; 5:gamma;6:2gamma;7:beta;8:beta&w;9:beta&gamma;10:3normal
        icode = 0   * 0:standard genetic code; 1:mammalian mt; 2-10:see below
        Mgene = 0   * 0:rates, 1:separate; 2:pi, 3:kappa, 4:all
    fix_kappa = 0   * 1: kappa fixed, 0: kappa to be estimated
        kappa = 2   * initial or fixed kappa
    fix_omega = 0   * 1: omega or omega_1 fixed, 0: estimate
        omega = 1   * initial or fixed omega, for codons or codon-based AAs
        getSE = 0       * 0: don't want them, 1: want S.E.s of estimates
RateAncestor = 0       * (0,1,2): rates (alpha>0) or ancestral states (1 or 2)
  Small_Diff = .45e-6  * Default value.
   cleandata = 1       * remove sites with ambiguity data (1:yes, 0:no)?
 fix_blength = 0       * 0: ignore, -1: random, 1: initial, 2: fixed
```

Run command file (null model):

The command file for the null model is the same as for the alternative model, except for two parameters (in red):

1) The name of the output file (outfile) is different.
2) The dN/dS ratio is fixed to 1 (fix_omega = 1).

```
     seqfile = TF105351.Eut.3.phy              * sequence data file name
    treefile = TF105351.Eut.3.53876.tree       * tree structure file name
     outfile = TF105351.Eut.3.53876.fixed.mlc   * main result file name

       noisy = 9   * 0,1,2,3,9: how much rubbish on the screen
     verbose = 1   * 1: detailed output, 0: concise output
     runmode = 0   * 0: user tree;  1: semi-automatic;  2: automatic
                 * 3: StepwiseAddition; (4,5):PerturbationNNI; -2: pairwise

     seqtype = 1   * 1:codons; 2:AAs; 3:codons-->AAs
   CodonFreq = 2   * 0:1/61 each, 1:F1X4, 2:F3X4, 3:codon table
       clock = 0   * 0: no clock, unrooted tree, 1: clock, rooted tree
      aaDist = 0   * 0:equal, +:geometric; -:linear, {1-5:G1974,Miyata,c,p,v}
       model = 2   * models for codons:
                   * 0:one, 1:b, 2:2 or more dN/dS ratios for branches
     NSsites = 2   * 0:one w; 1:NearlyNeutral; 2:PositiveSelection; 3:discrete;
                   * 4:freqs; 5:gamma;6:2gamma;7:beta;8:beta&w;9:beta&gamma;10:3normal
       icode = 0   * 0:standard genetic code; 1:mammalian mt; 2-10:see below
       Mgene = 0   * 0:rates, 1:separate; 2:pi, 3:kappa, 4:all
   fix_kappa = 0   * 1: kappa fixed, 0: kappa to be estimated
       kappa = 2   * initial or fixed kappa
   fix_omega = 1   * 1: omega or omega_1 fixed, 0: estimate
       omega = 1   * initial or fixed omega, for codons or codon-based AAs
       getSE = 0       * 0: don't want them, 1: want S.E.s of estimates
RateAncestor = 0       * (0,1,2): rates (alpha>0) or ancestral states (1 or 2)
  Small_Diff = .45e-6  * Default value.
   cleandata = 1       * remove sites with ambiguity data (1:yes, 0:no)?
 fix_blength = 0       * 0: ignore, -1: random, 1: initial, 2: fixed
```

Launch CodeML:
In Unix (Linux, MacOSX), this will look like:

```
codeml ./TF105351.Eut.3.53876.ctl
codeml ./TF105351.Eut.3.53876.fixed.ctl
```

Analyse results:

1) Assign significance of the detection of positive selection on the selected branch:

Two output files are produced:
TF105351.Eut.3.53876.mlc (alternative model) and TF105351.Eut.3.53876.fixed.mlc (null model).

We retieve the likelihood values lnL1 and lnL0 from TF105351.Eut.3.53876.mlc and TF105351.Eut.3.53876.fixed.mlc files, respectively:

lnL(ntime: 41 np: 46): -4707.210163 +0.000000 (lnL1)
lnL(ntime: 41 np: 45): -4710.222252 +0.000000 (lnL0)

We can construct the LRT:
-->
$\Delta LRT = 2\times(lnL1 - lnL0) = 2\times(-4707.210163 - (-4710.222252)) = 6.024178$

The degree of freedom is 1 (np1 - np0 = 46 - 45).
p-value = 0.014104 (under χ^2) => significant.

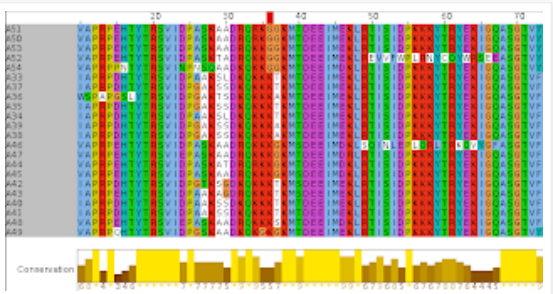
A significant result with the branch-site codon model means that positive selection affected a subset of sites during a specific evolutionary time (also called episodic model of protein evolution).

2) If significant, we can retrieve sites under positive selection:

In the TF105351.Eut.3.53876.mlc, we can retrieve sites under positive selection using the Bayes Empirical Bayes (BEB) method:

```
Positive sites for foreground lineages Prob(w>1):
    36 K 0.971*
   159 C 0.993**
```

Amino acids K and C refer to the first sequence in the alignment.
Position 36 has a high probability (97.1%) of being under positive selection. Position 159 has a very high probability (99.3%) of being under positive selection.



Position 36 shifted from a lysine to a glycine.

In future posts, I will speak about various potential problems (or not) and limits of the inference of positive selection.

RAS

Posted by Romain Studer at 10:05

136 comments:

Anonymous 30 May 2013 at 09:14

Hi Romain, thanks for posting this, it was very helpful. I noticed that you transposed "changing" and "keeping" in the second sentence of theoretical principles. The placement of the '#1' on the diagram is also a little confusing as it applies to the node defining the clade (A44-A50) rather than the branch that it appears to be sitting on. Thanks again.

Reply

Romain Studer 30 May 2013 at 09:24

Thanks for pointing that. I changed the text and the diagram.

Cheers,
Romain

Reply

Sw@ti 26 June 2013 at 04:30

Hi Romain,

Thanks for the information. It was very helpful for me. But now i'm stuck in another problem.
I'm using codeml to estimate omega value for each node of the given tree. I'm doing this for almost 1000 genes. Now I've the result file but I'm a bit confused. In the output file even though the value of dN and dS is 0.000 the omega value is not 0.000? Is it some kind of error or I can use these omega value for my further analysis? The output is like this:

branch t N S dN/dS dN dS N*dN S*dS

26..27 0.000 1308.3 290.7 1.4906 0.0000 0.0000 0.0 0.0
27..10 0.000 1308.3 290.7 1.5623 0.0000 0.0000 0.0 0.0
27..1 0.000 1308.3 290.7 1.4859 0.0000 0.0000 0.0 0.0
31..6 0.006 1308.3 290.7 0.0001 0.0000 0.0119 0.0 3.5
31..32 0.436 1308.3 290.7 0.0425 0.0285 0.6705 37.3 194.9
32..33 0.008 1308.3 290.7 65.0863 0.0034 0.0001 4.4 0.0

Thanks,
Swati

Reply

anna 19 July 2013 at 11:18

Greetings,
I performed branch-site test (model A test2) in codeml. For some genes I have significant LRT but no BEB sites or sometimes only NEB. Could I use these results or should I combine the results of LRT test with BEB?

Thank you in advance.
Ana

Reply

Anonymous 22 August 2013 at 06:59

Hi Romain,

I did some tests with codeml to infer positive selection. I checked convergence by performing the analysis with different starting omega values (i.e. 0.1, 1, 2, 100) and observed that more branches shows positive selection as the initial omega is greater. I didn't find out any clear solution for this. Do you know any paper that discuss this question or have any suggestion?

thanks in advance,
Luana

Reply

Romain Studer 30 August 2013 at 07:10

I would recommend to keep starting with an omega value of 1.
An omega value of 100 is clearly too much to start.
How many branches on how many genes did you check with different omega initial values?

Reply

Anonymous 14 November 2013 at 06:22

Hi Romain,

I am bit confused why do people specify in control file different Omega vales in 0,1,2,10 or so. What does that mean? Please help!

Thanks,
nahil

Reply

Romain Studer 20 November 2013 at 13:39

Hi Nahil,

This is because needs to start from one value, and then refine the parameters to reach the optimum, and so give a more realistic omega. Starting from 1 is a good choice as the model will either go to negative (omega <1) or positive selection (omega>1).

I hope I answered your question,
Romain

Reply

Anonymous 5 December 2013 at 10:18

Hi Romain,

Thanks for this post. It has been very helpful to me. However, I have a slightly different situation and am unsure how to implement the appropriate analyses. I would like to test for convergent selection in several evolutionarily independent lineages. Can I specify multiple lineages as the foreground branches for the branch-site test? Should I just test each lineage separately? Or should I be using a branch test in this type of situation, as positively selected sites may differ between target lineages?

Thanks,
Ben

Reply

Romain Studer 17 December 2013 at 07:14

Hi Ben,

Sorry for the delay.

What I usually do is to test branch by branch. And then use q-value to assess the significance. But I think you can also select all branches at the same time, like in the normal branch models.

There may be the clade-model C and D that were designed for such problem, but I never tried it.
http://www.ncbi.nlm.nih.gov/pubmed/15383915

I hope I answered your question,
Romain

Reply

Natasa 29 April 2014 at 06:35

Hi Romain,
I just started using codeML in PAMLX and am not very experienced with writing codes. How can I test all the branches at same time? How do you specify this in the tree file? I would like to obtain lnL values for all the branches. I used earlier PhyleasProg, but it provides lnL values only for the significant branches.

Thanks!
Natasa

Reply

Romain Studer 22 May 2014 at 12:33

Hi Natasa,

Sorry for the late answer. Actually, you don't specify all branch at the same time, but you recursively analyse branch by branch with CodeML, by moving the "#1" tag everytime. So that requires some code scripting.

In the clade-model, you can the tag "#1" on all branch you would like to test. But I am not familiar with this model.

Please don't hesitate to contact me by email (rstuder [at] ebi.ac.uk) if you need more detail.

Reply

Replies

Natasa 26 May 2014 at 01:28

Hi Romain!
Thanks for your response!
I have around 200 genes to analyse, so it would take a long time to do it branch my branch. I will have a look at clade-model.

Natasa

Reply

Romain Studer 26 May 2014 at 04:46

But you will have different results than the branch-site model.

I may have some scripts to automatise the annotation of genes and trees. If you would like to contact me by email (rstuder [at] ebi.ac.uk], I can help you to set up this pipeline.

Reply

Unknown 9 June 2014 at 19:07

This comment has been removed by the author.

Reply

Unknown 9 June 2014 at 19:10

This comment has been removed by the author.

Reply

Romain Studer 10 June 2014 at 02:43

Hi,

Assuming you have all the files in the same folder, the commands you need to run are the following:

codeml ./TF105351.Eut.3.53876.ctl
codeml ./TF105351.Eut.3.53876.fixed.ctl

If you would like to contact me by email (rstuder [at] ebi.ac.uk], I can help you to set up CodeML.

Reply

Unknown 12 June 2014 at 06:29

Oh I just realized you answered my message, I didn't see that and I had just deleted it because I figured out what happened. Seems the code on the paml website for specifying the path wasn't working on my mac. Thanks for the reply though!!! I am now running your tutorial successfully!

Reply

Unknown 15 July 2014 at 12:32

Hi Romain,

I was actually looking for the information on how to set up the tree file in case of multiple hypothesis testing (e.g all branches in a tree), and that's how I came to this site.

So I have to write a script such that each node is set as 'foreground' one at a time. In such case will the log-likelihoods for all rounds be outputted in a single file or multiple result file are produced?

I may need to write an email directly to you for further help, really!

Reply

Romain Studer 15 July 2014 at 13:38

Hi Edson,

You will have multiple result file, two per node (one for null hypothesis, one for alternative hypothesis).

You compute the p-value for each node. Finally, you have to correct for multiple testing (i.e. using qvalue): http://evosite3d.blogspot.co.uk/2013/07/false-discovery-rate-correction-for.html

I hope it helps you.

Romain

Reply

Unknown 16 July 2014 at 01:59

Thanks Romain for your help and for your site.

Reply

Anonymous 2 September 2014 at 05:52

Thanks for the helpful post! I'm trying to run branch-site models now, but I'm a bit confused about one thing.

Do I need significant results for the site-models (M0:M3 and M7:M8) in order to run branch-site models?

Or are these two completely separate tests?

Cheers.

Reply

Replies

Romain Studer 2 September 2014 at 05:55

Hi,

No, there is no need to run site models prior to branch-site model.
Site-models and branch-site models are totally independent.
Site-models intend to find sites that are under recurrent positive selection across the whole tree.
Branch-site models intend to identify sites that have been under positive selection at a particular point of evolution, i.e. on a specific branch of the tree.

Anonymous 2 September 2014 at 06:02

Excellent. Thanks again for the help!
Alice

Reply

Anonymous 10 November 2014 at 11:30

Hi Romain,

Is it sufficient to run this simulation a single time to identify positive selection along a branch? Or, does one need to run the alternative model with different initial values of omega to check that the simulation is converged?

Reply

Replies

Romain Studer 10 November 2014 at 12:39

Hi,

It is better to run multiple times (i.e. 3x or 6x) both null and alternative models. And then take the best likelihood value of each.

Reply

Anonymous 11 November 2014 at 08:31

Ok, that makes sense. Thank you for your help!

Reply

Unknown 19 November 2014 at 06:42

Hello Romain!

Thank you for the excellent article.
I have one question about codeml. If my strains of interest are not in one branch of tree file made by fdnaml program, what should I do? Could I label all my strains with #1, for example it's a part of a tree file:((3#1:0.00339, (26:0.00318,16:0.00428):0.00187):0.00269,(20:0.00687,
((9#1:0.00781,2:0.00719), or should I do something else?

Thank you in advance,
Konstantin

Reply

Replies

Romain Studer 19 November 2014 at 07:17

Hi Konstantin,

You have to make sure that all your nucleotide sequences are in your tree. So you should redo your tree to make sure it is the case.

To be honest, I never used "fdnaml". I generally use PhyML or RAxML when doing phylogenies under Maximum Likelihood.

If you label all strains with "#1", you will evaluate positive selection that continuously occurred in those branches. But the branch-site model wants to find positive selection that occurred at a specific time point (i.e. the branch of interest). So the result will be different.

I am also worried by the fact you are using strains, because they generally don't provide enough divergence. I suspect you will not have enough power to detect a signature of positive selection.

Feel free to contact me by email if you want me to have a look at your data.

Romain

Reply

Anonymous 12 December 2014 at 13:11

Hi Romain,
i have a gene from several species from one plant family and i would like to test whether there was positive selection on a certain branch (containing sequences with a gained additional function) of the alignment (and in addition, which sites are affected in case of significant differences in ω). I did a branch-site model with different ω-values (1,2,5,10), but the tests were not significant. If i, however, run a test of ω variation among lineages (using 'branch models') for the respective branch (setting model = 2 and NSsites = 0) and compare this to the log-likelihood for one ω for all sites and all branches (setting model = 0 and NSsites = 0) which should be the null model of no ω variation among branches i found a highly significant difference between the likelihoods (LRT=28.3, p-value 1.03896e-07). How can i interpret this?

Thank you very much for any help to interpret this result!

Reply

Replies

Romain Studer 12 December 2014 at 16:38

Hi,

For the branch-site model: what do you mean by different values? Are they different starting values?

For the branch model: my interpretation is that by comparing M2 and M0, you found that your branch of interest evolve under different selective pressure. But you cannot conclude it is positive selection or just relaxation of selective pressure. To do so, you need to compare to a null model where your branch of interest is different and its omega fixed to 1.

It is explained in point 1.3:
http://www.ch.embnet.org/CoursEMBnet/PagesPHYL07/Exercises/day2/day2.html

If you want to send me your files (alignment, tree, ctl and mlc files), I would be happy to have a look at them.

Reply

Unknown 2 February 2015 at 18:30

HI Romain,

I'm trying to use this method on a really tiny phylogeny. I only have 4 sequences. Can I even do this? Everywhere I look, people are using this for 1000s of genes, but I mainly just want to look at what changes have happened between my two species of interest. Specifically, I am interested in the branch-site model, because the outgroups are very very long outgroups. I am also wondering whether the branch method would work, but I didn't get any meaningful information when I tried that. Am I asking the wrong questions?

Reply

Replies

Romain Studer 3 February 2015 at 07:31

Hi Rose,

I don't think it will reliable enough. Four sequenced are quite low for these models of codons evolution.

Maria Anisimova did some analysis on the power and accuracy and found that it is not reliable below seven sequences:

http://mbe.oxfordjournals.org/content/19/6/950.full

"Based on our simulations, we make the following generalizations. (1) Prediction of positively selected sites is unreliable when sequences are very similar, and the number of lineages is small (e.g., $S \leq 0.11$ or $T \leq 6$). (2) Increasing the number of lineages is the most effective way to improve accuracy and power. Accurate prediction is possible for data sets comprising very similar sequences if a very large number of lineages have been sequenced. (3) Multiple models should be used in real data analysis to ensure the robustness of the results."

Could you increase your dataset by adding sequences from other close species?

Unknown 14 May 2015 at 20:18

HI Romain,
I never noticed your reply! I have since abandoned that line of thinking, and am looking at different strains of Drosophila where there are plenty of sequences to work with. I'm now using the branch method to test whether different environments have encouraged different selection pressures across the phylogeny. I have been comparing the null of omega being the same over all of the branches with omega being different across all branches. If this is the case I then test my hypothesis of it being different on all of the other branches, determined by what 'group' I have put them into (labelling my groups of interest all with one #, and the other branches as the tree as consecutive # to allow all of them to vary). Should I instead be testing each branch individually by moving the #1 around?

Thanks!

Romain Studer 15 May 2015 at 02:04

Hi Rose,

The model you are trying to use is the free-ratio model, which allow different dN/dS on each branch of the tree. Then you will compare with the one-ratio M0 model. There are two problems in that case:

1) Nearly any models compared to the one-ratio M0 would be significant. But it doesn't mean it will be positive selection, it could be also just relaxation of selective pressure.

2) The free-ratio model will estimate each dN/dS for each branch => this model is too rich in term of parameters and prone to overfitting. It is discouraged.

=> I would recommend your second option, aka testing each branch individually by moving the #1 around. You will use a two-ratio model that allows positive selection, versus a two-ratio model that doesn't allow positive selection. But you could also use the branch-site, which might be more powerful to detected episodic selection.

Romain

Reply

Anonymous 1 March 2015 at 01:46

Hi pal
I got the message: "Hessian matrix may be unreliable for zero branch lengths", and I am not sure what I should do about it. any suggestions?
cheerio!

Reply

Replies

Romain Studer 2 March 2015 at 03:40

Hi,

Do you have identical sequences in your alignment? If yes, I would remove them, from both the alignment and the tree.

Also, what is your value for the "fix_blength" parameter?

Reply

Anonymous 2 March 2015 at 19:26

thanks a lot for the reply.
fix_blength is set on zero, as I am not using any lengths for the branches.

thanks!

Reply

Jeanne Blanchard 20 March 2015 at 04:27

Hi Romain,
I used the branch-site method. Now I would like to be sure that the selective selection I have detected, is really selective selection and not a result from BGC. For that, I need to know which substitutions occured in my significative sites. As I have the position of the site, I just need to know which sequence is used as the ancestral sequence and from that i will deduce the substitution.

Do you know how I can find which sequence is used as the ancestral sequence?

Thank you a lot,

Reply

Replies

Romain Studer 20 March 2015 at 07:40

Hi Jeanne,

Everything you are looking for are in the rst file. This file contains all the probabilities for a particular amino acid / codon to be present at a specific position at a specific node.

There are also consensus sequences, and all mutations that happened from a node to another one.

Cheers,
Romain

Reply

Unknown 28 May 2015 at 01:19

Hello Romain,

Sorry to bother you. I'm a beginner and I use the branch model on Coding-DNA Sequences (runmode = 0, seqtype = 1, CodonFreq = 2,model = 2,NSsites = 0). I have some questions. My aim is to know the dN/dS on different lineages (species 1 and 2) with the selection directed by an outgroup. I have 1 outgroup, 7 individuals in the species 1 and 4 individuals in the species 2.

(1) Where am I supposed to put the '#1' in the tree file if I want the dN/dS of species 1 and 2? Is my tree correct according that species 2 is more relative to species 1 and that species 1 is more relative to the outgroup (and according that individuals in a species are species really closely relative between them)?

(outgroup#1,
((ind1_species1,ind2_species1,ind3_species1,ind4_species1,ind5_species1,ind6_species1,ind7_species1),
(ind1_species2,ind2_species2,ind3_species2,ind4_species2)));

(2) In more can I really make those analyzes with individual sequences and not species sequences ?

Thank you in advance,
Cheers,

Blaise

Reply

Replies

Romain Studer 28 May 2015 at 02:04

Hello Blaise,

1) To test the branch leading to species 1, you need this tree:
(outgroup,
((ind1_species1,ind2_species1,ind3_species1,ind4_species1,ind5_species1,ind6_species1,ind7_species1)#1,(ind1_species2,ind2_species2,ind3_species2,ind4_species2)));

and to test the branch leading to species 2, you need this tree:
(outgroup,
((ind1_species1,ind2_species1,ind3_species1,ind4_species1,ind5_species1,ind6_species1,ind7_species1),(ind1_species2,ind2_species2,ind3_species2,ind4_species2)#1));

2) The dN/dS is used to compare the rate of substitution. By substitution, it means mutations fixed in the global population (aka the major allele). So you can compare at the species level, but not at the individual level. In your case, I would rather use a McDonald Kreiman test:
http://en.wikipedia.org/wiki/McDonald%E2%80%93Kreitman_test

This test compares the rate of synonymous and non-synonymous substitutions in each populations and also between the two species of interest.

Best regards,
Romain

Reply

Hari 19 June 2015 at 07:12

Hi Romain,

I am trying to run the branch-site model in PAML to test if the dN/dS on a given foreground is significantly higher (or lower) than that of the background. Since the branch-site model output gives me the foreground and background dN/dS (w) values for each of the site classes (i.e. the 4 categories of sites assumed by the model), could I compute the mean foreground w by multiplying the proportion of sites with the w for that category, and summing this product across all categories?

I am also unsure of how to test if the resulting mean foreground w is significantly different from the mean background w. Should I just run the better-fit model, say, a 1000 times to get a distribution of foreground- and background- w values and then do a statistical test? What would you recommend?

Thanks so much! I've been reading all the comments and your responses have been particularly helpful in my understanding of the PAML toolkit!

Best,
Hari

Reply

Replies

Romain Studer 19 June 2015 at 07:49

Hi Hari,

If you want the mean dN/dS, you can try the "branch model" instead the "branch-site model". You only need to replace "NSsites = 2" by "NSsites = 2". And keep ""fix_omega = 0, omega = 1. CodeML will evaluate the dN/dS, starting from 1, and will increase or decrease it. It will give you the average dN/dS on the selected branch.

1) If you want to see if your dN/dS foreground branch is significantly different (lower or higher) of your dN/dS background, you can compare it to the basic one-ratio model M0 (Model = 0, NSsites = 0) by using a likelihood ratio-test (LRT). So no need to run it 1000 times.

2) If you want to see if your foreground branch is under positive selection (dN/dS>1), you can then compute it another null hypothesis using the "branch model" with "fix_omega = 1, omega = 1" (which will be the NULL hypothesis), and compare it to your previous results (the one with "fix_omega = 0, omega = 1).

More details here:
http://www.ch.embnet.org/CoursEMBnet/PagesPHYL07/Exercises/day2/day2.html

Best regards,
Romain

Hari 19 June 2015 at 16:00

This comment has been removed by the author.

Hari 19 June 2015 at 16:02

Thanks Romain, that clears things up a lot.

I ran the branch model (Model = 2, NSSites= 0) on one of my datasets and found that it had significantly higher likelihood ($p<0.05$) than the one ratio model (Model = 0, NSSites = 0), indicating the foreground dN/dS (w) is significantly higher than the background w. Is there a way I can get a confidence interval for w? [I ran the branch model multiple times on the same dataset, and observed that the mean w on the foreground and the background don't change (which is expected I guess, since I do not change the dataset)].

Best,
Hari

Romain Studer 25 June 2015 at 01:31

Hi Hari,

Sorry for the delayed answer.

There are probably a w interval in one of the rst file.

In any case, if your w is between 0 and 1, you can only say it is under purifying selection.

If you found a w (dN/dS) >1, you need to perform another test by comparing your model (Model = 2, NSSites= 0, fix_omega=0) to the model (Model = 2, NSSites= 0, fix_omega=1) .

Best regards,
Romain

Reply

Hari 19 June 2015 at 16:01

This comment has been removed by the author.

Reply

Anonymous 24 June 2015 at 13:19

Hi Romain,

Such a great and helpful post! I was wondering if you could help with a simple question. What is the difference between the Branch-site and Branch model in codeml? I know the branch-site model identifies sites under positive selection on a specified lineage, but what information does the branch model provide that's different from the branch-site model?

Thanks!

Reply

Replies

Romain Studer 25 June 2015 at 01:35

Hi,

The branch model assumes that all sites evolves at the same rate on the specific lineage (it is like an average). Sites on the background have a dN/dS identical (w0), and sites on the foreground have another dN/dS value (w1).

The branch-site model allows that sites to evolve at different rates on the specific lineage:

Sites with identical dN/dS in both foreground and background branches:
K0 : Proportion of sites that are under purifying selection (ω0 < 1) on both foreground and background branches.
K1 : Proportion of sites that are under neutral evolution (ω1 = 1) on both foreground and background branches.

Sites with different dN/dS between foreground and background branches:
K2a: Proportion of sites that are under positive selection (ω2 ≥ 1) on the foreground branch and under purifying selection (ω0 < 1) on background branches.
K2b: Proportion of sites that are under positive selection (ω2 ≥ 1) on the foreground branch and under neutral evolution (ω1 = 1) on background branches.

Im my opinion, the branch-site is more powerful and accurate than the branch model.

Best regards,
Romain

Reply

Unknown 7 September 2015 at 04:35

Hi Romain

Two question on Branch site.
1) Is it possible to use BS by labeling several branches at a time, such as a clade for example as foreground.
2) If BS is significant but the w value of the foreground is = to 1 what does this mean????

Thanks

Reply

Replies

Romain Studer 7 September 2015 at 06:16

Hi,

1) Yes, but your biological questions will be different in that case. Look also at the clade model (model C and D).
A Maximum Likelihood Method for Detecting Functional Divergence at Individual Codon Sites, with Application to Gene Family Evolution
http://link.springer.com/article/10.1007%2Fs00239-004-2597-8

2) This means there is likely an error. Maybe a problem of convergence? Try to run it three or five times. You can also send me you data (alignment + tree + codeml file), I will have a quick look.

Thanks,
Romain

Reply

Anonymous 14 September 2015 at 05:54

Hi Romain

Thanks so much for your answers.

1) could you be more specific in terms of what questions are apply when using one branch-BS or several branches-BS?
I used both the BS and CmC models with similar partitions, got different results and I am trying to Interpret because is some cases both provided significant results, but in other cases BS was significant while CmC was not or vice-versa. In addition the CmC also provided very low values of w even if significant. I guess I am working with a very conserve genome (mitochondrial) not easy.

2) How can I attach the data.

It will be interesting to see your results. I run the BS model using three different starting points of kappas (0.2, 2, 6) then I choose the best Lnl from the Alternative and the Null and calculated the Lrts. Is that the way to do it? For the CmC I used instead different omegas.

can I send you the data by email?

Thanks a lot

Tibisay

Reply

Replies

Romain Studer 15 September 2015 at 01:47

1) When you specify one branch, you ask if something happened at this point (i.e. shift from one function to another). When you specify multiple branches, you ask if a gene is under continuous change (i.e. adaptation to something also changing).

2) Yes, please send them by email. (rstuder [@] ebi.ac.uk).

Romain

Anonymous 1 February 2016 at 16:46

I would like to ask what happens lRT test of null vs. alt model is significant, omega>1 on the forward branch (k2a and/or k2b have >1 values), but there are no significant sites (or that some sites are identified as positively selected, but not with >95% probability).

should this mean that there are positively selected sites, but we cannot detect which, or that we cannot reject h0 ?

assaf

Romain Studer 2 February 2016 at 02:18

Hi Assaf,

Exactly, it means that the LRT test was powerful enough to detect selection on that branch, but the Bayes Empirical Bayes (BEB) was not powerful enough to pinpoint particular site.

It could be due to your dataset. Maybe you don't have enough sequences, maybe the sequences are too divergent, or not enough, or some sequences are particularly different from the others.

I can have a quick look at your data if you want.

For that, I would need:
- Alignment (in nucleotide)

- Tree
- CodeML control file (ctl)
- CodeML result file (mlc)

My email is (rstuder [@] ebi.ac.uk).

Romain

Reply

MEZ 21 September 2015 at 06:41

Hi Romain,

I ran into your blog by searching for help on PAML. Thanks a lot for the helpful info. I have a several questions. I am testing for positive selection on a mammalian receptor. I used a dataset of 35 taxa and an alignment of 3,000 bp long corresponding to the codon sequence. First, I tested for M0 vs M2 constraining a specific order of interest (with a $1 at the beginning of the node to mark all following branches instead of putting a #1 each branch) and the LRT was not significant. Then, I tested M1 vs M2 and in this case the LRT was significant with an estimated omega of 1 for the selected node and of 0.24 for the rest of the tree. Then I tested M7 vs. M8 to know if there were PSS and determine the sites. In this case, LRT was significant and the analysis resulted in a low proportion of sites (approx. 0.1%) evolving with omega>1 with PSS scored both by NEB and BEB. My first question comes in how should I interpret such results? I would assume there is selection acting on the node of interest when compared to the rest of the tree; however I don't understand why M0 vs M2 LRT was not significant like the other tests. Any ideas?

Finally, I tested with the branch-site A model constraining the node of interest again and ran the test with an initial omega value of 0.4 and of 1. I ran each test twice to check for convergence and then I ran a H0 test of the same model with omega fixed to1. In all cases I obtained the same lnL! Even if sites were scored under BEB, I cannot perform the LRT. What are you recommendations? An idea why this happens?

Thanks beforehand!

PhD Student

Reply

Anonymous 9 October 2015 at 08:42

Hi Romain,

I have a doubt related with the tree labeling (for branch-model analysis) in a particular case of my work.

Can I send you the data by email?

Thank you very much,
Daniela

Reply

Replies

Romain Studer 9 October 2015 at 08:44

Hi Daniela,

Yes, of course. I will have a look over the week.

Please send me as much information as possible.

Romain

Reply

Zwätschga 1 March 2016 at 08:06

Hi Romain

Thanks for this post, it was very helpful. I'm a Master student and currently struggling with PAML...
So at first I thought, I'm too stupid to use PAML - but with your dataset I was able to do it.
So I figured, that probably something is wrong with my treefile and I compared yours and mine and there indeed are differences.. I did my tree with MrBayes, but the treefile I got from there looks totally different from yours. I tried different softwares (FigTree export as Newick or Nexus, SeaView export as unrooted tree) and ways to convert it and get the same format as your file - and it looks quite similar but still is not working..
Your treefile looks different in a sense that the grouping/brackets are differently put together. (I don't know how to explain that :-D)

So my question is:
How can I get a treefile (as yours) from my MrBayes files? Or what did you use to create your treefile?

Cheers,
Natacha

Reply

Replies

Romain Studer 1 March 2016 at 08:11

Hi Natacha,

I cannot answer without seeing the tree. Please send me your tree to my email address: rstuder [@] ebi dot ac dot uk.

Even better if you can also send me your alignment and command file.

Thanks,
Romain

Reply

Matt Moore 10 March 2016 at 09:17

This is a great tutorial thanks!

The problem I'm having is interpretation of the results. For example I have bacterial genome sequences from within 3 different hosts (15 genomes each).

So take a gene in one patient for example, I provide a phylip tree and multiple alignment (unrooted) from these 15 genomes and I simply want to know whether that gene is under positive, neutral or purifying selection at all.

Is this possible from the output so that I can write a script that says yes or no, this gene is under positive selection (as I'm automating it over 5,500 multiple alignments

Thanks in advance for any help with this!

Reply

Replies

Romain Studer 21 March 2016 at 06:44

Hi Matt,

Yes, you can do that. You can estimated the global selective pressure (M0) for each gene family (which would contain 15 sequences).

You can also try the site model (M2a versus M1a) or M8vsM8a, which allow different selection at sites.

But you can also take all together and do a branch-model. As you have three hosts, you will have to run automatically your pipeline by selecting each branch at time. Each branch would be the one leading to one particular host. I think it is the most sensitive way to do it.

Romain

Reply

Unknown 23 March 2016 at 07:40

Hi Romain,

Your tutorial is a lifesaver! PAML is such a difficult program to learn and run that I almost gave up until I found your blog. So, thank you so much! I wanted to ask you a quick question. I read the BEB paper you have a link to, but I was wondering how you identified the sites under positive selection. The paper discusses the math behind BEB but I was wondering whether you wrote a custom script to locate site under selection or does PAML do that for the user? Thanks so much for your help!

Taruna

Reply

Replies

Romain Studer 23 March 2016 at 08:30

Hi Taruna,

Quick answer: PAML do that for you.

Once you ran CodeML, it will produce a MLC output file. In this file, at the end, you will have the results for the BEB analysis, as in the example:

2) If significant, we can retrieve sites under positive selection:

In the TF105351.Eut.3.53876.mlc, we can retrieve sites under positive selection using the Bayes Empirical Bayes (BEB) method:

Positive sites for foreground lineages Prob(w>1):
36 K 0.971*
159 C 0.993**

Amino acids K and C refer to the first sequence in the alignment.
Position 36 has a high probability (97.1%) of being under positive selection. Position 159 has a very high probability (99.3%) of being under positive selection.

Best regards,
Romain

Unknown 23 March 2016 at 17:12

Perfect! Thanks, Romain! I'll look at my output files right away!

Reply

Anonymous 8 April 2016 at 09:49

Hi Romain,
Thanks for your example.
I will need to label species from different nodes in the tree, so I was wondering if labeling each branch individually would give the same result. Here are some examples modifying your tree file:

(A)
((((((((((((A37,A36),(A35,(A34,A33))),(A39,A38)),A40),A41),A42),A43),A49),((A46,A45),
(A47,A48))),A44)#1,A54),A53),A52),A51,A50);

lnL(ntime: 41 np: 46): -4707.209700 +0.000000
lnL(ntime: 41 np: 45): -4710.222252 +0.000000

(B)
((((((((((((A37#1,A36#1),(A35#1,(A34#1,A33#1))),(A39#1,A38#1)),A40#1),A41#1),A42#1),A43#1),A49#1),
((A46#1,A45#1),(A47#1,A48#1))),A44#1),A54),A53),A52),A51,A50);

lnL(ntime: 41 np: 46): -4711.333187 +0.000000
lnL(ntime: 41 np: 45): -4711.333187 +0.000000

(C)
((((((((((((A37,A36),(A35,(A34,A33))),(A39,A38)),A40),A41),A42),A43),A49),((A46,A45),
(A47,A48))),A44),A54#1),A53#1),A52#1),A51#1,A50#1);

lnL(ntime: 41 np: 46): -4690.522348 +0.000000
lnL(ntime: 41 np: 45): -4690.533945 +0.000000

(D)
((((((((((((A37#1,A36#1),(A35#1,(A34#1,A33#1))),(A39#1,A38#1)),A40#1),A41#1),A42#1),A43#1),A49#1),
((A46#1,A45#1),(A47#1,A48#1))),A44#1)#1,A54),A53),A52),A51,A50);

lnL(ntime: 41 np: 46): -4711.986242 +0.000000
lnL(ntime: 41 np: 45): -4711.986242 +0.000000

As you can see, they all give different lnLs. Would you know why or another solution to this?

My final goal would be something like this:
((((((((((((A37#1,A36#1),(A35,(A34,A33))),(A39,A38#1)),A40),A41),A42#1),A43),A49#1),((A46,A45),
(A47,A48))),A44),A54),A53#1),A52),A51,A50);

Many thanks
R.G.

Reply

Replies

Romain Studer 14 April 2016 at 03:15

Dear R.G.,

Each labelling are different, and they means different things. Remember that CodeML assign two different classes on background branches and foreground branches. All branches that you tag with "#1" are considered as foreground and then could receive a dN/dS>1.

In your final goal, you are likely test for convergent evolution between species A53, A49, A38, A36 and A37.

I would also tag the branch leading the A36-A37 in that case:
((((((((((((A37#1,A36#1)#1,(A35,(A34,A33))),(A39,A38#1)),A40),A41),A42#1),A43),A49#1),((A46,A45),(A47,A48))),A44),A54),A53#1),A52),A51,A50);

(Sorry, it is quite hard to visualise without knowing your biological hypothesis).

Best regards,
Romain

Reply

Unknown 15 April 2016 at 09:56

Hi Romain. I was wondering if there is a way to specify more than one foreground branch in the tree file when running branch-site model. The PAML manual is bit confusing on this topic. Thanks very much!

Reply

Replies

Romain Studer 17 April 2016 at 13:50

Hi Taruna,

Yes, you can specify more than one branch as foreground. Just add a "#1" to any of those you would like to test. Remember they will have the same selective pressure, so adding more than one branch in different parts of the tree means your looking after convergent evolution. Depending of what you are trying to do, it might be better to test one branch after the other.

You can find more details on pages 14-16:
http://abacus.gene.ucl.ac.uk/software/pamlDOC.pdf

Best regards,
Romain

Unknown 18 April 2016 at 08:14

Ah I see! I guess I'm trying to test for convergent evolution as well. Thanks for your help!

Reply

Unknown 19 April 2016 at 05:25

Hi Romain. Sorry I ran into another problem. So I tried to run PAML with the same parameters as the small lysosome example that comes with PAML4.8a (except my NSsites = 2), but I kept getting an error that only two branch types are allowed for branch models. The example control file specifies NSsites = 0 and my control file has NSsites = 2. I need to have three branch types (0, #1 and #2). I'm just confused as to why I can't specify NSsites = 2 when testing 3 branch types. Would it be alright if I send you my control file for advice? Thanks!

Reply

Replies

Romain Studer 19 April 2016 at 05:35

This is the problem. You can have only two type of branches for the branch-site model: background and foreground (#1).

And yes, you can send me your files (alignment+tree+control file).

Romain

Reply

Unknown 22 April 2016 at 00:27

Hi Romain.

I am a graduate student of Tohoku University in Japan.

Now, I'm attempting to detect signatures of convergent adaptive evolution of species that adapted to high temperature environment by using branch-site model of PAML program.
Then, I have a ploblem. I don't know what kind of the tree file is suitable for detect convergent adaptive evolution.
This is the topology of the tree.

((1, 2), ((3, (4, 5)), ((6, 7), (8, 9))));

When I would like to detect signatures of convergent adaptive evolution of No. 3 and 8 species, which method is the best one.

1.

tree file: ((1, 2), ((3#1, (4, 5)), ((6, 7), (8#1, 9))));
One run of PAML (put "#1" on No. 3 and 8 branch.)

2.
tree file 1: ((1, 2), ((3#1, (4, 5)), ((6, 7), (8, 9))));
tree file 2: ((1, 2), ((3, (4, 5)), ((6, 7), (8#1, 9))));
Two separate runs of PAML. When both of two runs demonstrate that the gene is under positive selection, I consider it to be related to convergent adaptive evolution.

3.
tree file: ((1, 2), ((3#1, (4, 5)), ((6, 7), 9)));
tree file: ((1, 2), ((4, 5), ((6, 7), (8, 9#1))));
Two Separate runs of PAML. Each tree file for each run doesn't have the other branch adapting high temperature environment.When both of two runs demonstrate that the gene is under positive selection, I consider it to be related to convergent adaptive evolution.

I read papers which detected signature of convergent evolution by using branch-site model but I don't know which method is the best one.

If possible, I would like to receive advice from you.

Please forgive me my lousy English.
I thank you for reading it through.

Shunsuke Kanamori

Reply

Replies

Romain Studer 22 April 2016 at 02:18

Hi Shunsuke,

To test for convergence, I would use the option 1. It means that both branches 3 and 8 have the same selective pressure. And the likelihood test will check if this selective pressure is different from the background.

And it would also be interesting to try option 2. You will check if positive selection is associated with 3 in one step, and with 8 in another step. You can then compare both test, and also compare with option 1.

Best regards,
Romain

Unknown 22 April 2016 at 04:07

Hi Romain,

Thank you for your reply.

I'll start of with the option 1.
After that I will try the option 2.

I'll report the result of the comparison between the option 1 and the option 2 to you after the analysis.
At that time, please discuss it together.

Thanks,

Shunsuke Kanamori

Reply

Unknown 22 April 2016 at 04:06

This comment has been removed by the author.

Reply

Mark 3 June 2016 at 06:42

Hi Romain,

Thanks for this great tutorial.
How do you generally deal with unreliably high dS values? I've been running branch model and branch-site model for an alignment and the outgroup is almost always saturated (dS > 2) but reliably low (dS < 2) for my focal species. Do I need to disregard the full analysis if dS is saturated only for the branch to the outgroup?

Thanks,
Mark

Reply

Replies

Romain Studer 16 June 2016 at 01:32

No, you can keep you analysis, the branch-site is quite robust. But try also by removing the outgroup to see if it makes a different. Unless your outgroup is all genes from one side of the target branch.

Romain

Reply

Unknown 12 June 2016 at 09:50

Hi Roman,

Thanks for the tutorial!!

I am really new in CODEML, so first I will do "biological" questions to you.
1) A33-A54 Could be a family of genes or should be one gene from different species?
2) How close should be the different linages? At Phylum / Class / Order / Family / Genus or Species level?
3)What happens if I have paralogous of the selected gene in the different species?
Thank you very much!!!

Alejandro

Reply

Replies

Romain Studer 16 June 2016 at 01:38

1) They could be paralogous genes or orthologous genes.

2) Hard question: they have to be close enough to avoid problem of saturation, but divergent enough to provide signal. Maria Anisimova published some studies about that:
http://mbe.oxfordjournals.org/content/18/8/1585.short
http://mbe.oxfordjournals.org/content/19/6/950.short
3) It really depends on your biological question. I would do both tests, with paralogs and without paralogs.

Romain

Reply

Anonymous 29 June 2016 at 06:06

Dear Romain Studer

I have a simple question (but serious for me).

When two models(M0=Null, M1=alternative) are compared using LRT, chi square distribution is used. So, p-value is calculated from df and chi-square value. However, I don't understand why some papers used Bonferroni correction for correcting p-value.

Additionally, in below sentence,

"P-values for uncorrected/ Bonferroni corrected χ2 tests, where Pcorr = α/# branch parameters in the model being tested."
If one branch was specified as foreground lineage, what does " # branch parameters in the model being tested" mean? In PAML manual, there is no mention about bonferroni correction. Therefore, I couldn't understand why some papers described the bonferroni correction and how they were calculated.

Thank you.

Kim

Reply

Replies

Romain Studer 29 June 2016 at 08:15

Dear Kim,

I will just summarise the models used here:

All these models are branch models and assume the same ratio amongst sites:

Evolutionary models:

M0: one unique dN/dS value across all branches.
M1: each branch has its own dN/dS value. This model is also called free-ratios.
M2: some branches can have a specific dN/dS value compared to the foreground. This model is also called two-ratios or three-ratios.

We could construct the following Likelihood Ratio Test (LRTs):

M0-M1 = test if any branch are under positive selection => 1 p-value.
M0-M2 = test if foreground branch are under positive selection=> 1 p-value.

The degrees of freedom (df) are the difference between the number of parameters between null and alternative hypotheses. In M1-M0, the df can be as huge as the number of branches, while in M2-M0, the df is around the number of foreground branches.

The test M0-M1 is discouraged because model M1 is very parameter-rich (source: PAML manual).

The test M2 gives a dN/dS per branch of interest. Let'say a tree like: ((A,B),C). You have no prior hypothesis and want to test some branches of interest: You can perform one test with branch A as foreground: ((A #1,B),C). Then you can test for branch B as foreground: ((A,B #1),C). Finally you can test for branch C as foreground: ((A,B),C #1) => You will get three p-values based on the same dataset.

As you don't have prior hypothesis, you need to correct for false-discovery rate. In case of Bonferroni correction, you will have p-value threshold = α/3.

So you need to use multiple test correction when testing multiple branches, and also if you test across

multiple gene families. In that case, you could use q-values.

This problem was very well explained by Maria Anisimova and Ziheng Yang: http://mbe.oxfordjournals.org/content/24/5/1219.short

I have also written a tutorial on the use of q-value package here:
http://evosite3d.blogspot.co.uk/2013/07/false-discovery-rate-correction-for.html

Best regards,
Romain

Anonymous 29 June 2016 at 10:21

HI Romain
Thank you for your clarification.

But I am still confused.

Let'say a tree like: ((A,B),C). You have no prior hypothesis and want to test some branches of interest: You can perform one test with branch A as foreground: ((A #1,B),C). Then you can test for branch B as foreground: ((A,B #1),C). Finally you can test for branch C as foreground: ((A,B),C #1) => You will get three p-values based on the same dataset.

If I just want to test whether the branch B has same omega to branch A and C, do I need to get three p-value based on the same data?
for indtance,
Null model : Bw=Aw=Cw
Alternative model : Bw, Aw=Cw
-2 delta ln = 10
df = 1
I can calculate P-value using chi square(10) and df(1).
in this case, what is the number of branches? three or two?
And if there was 10 species and I wanted to assess three models
null model : all ws are same
alternative model 1 : one branch has different w against other branches
alternative model 2 : two different branches have different ws against other branches

in this case, is p-corrected value α/10 or α/2(α/3)?

Here is similar paper what I want to test.
"The Highly Reduced Plastome of Mycoheterotrophic Sciaphila
(Triuridaceae) Is Colinear with ItsGreen Relatives and Is under
Strong Purifying Selection"

They assessed changes in selective regime in 18 genes in Sciaphila densiflora.

In this case, only Sciaphila densiflora was interesting and each gene was separately assessed. So, they compared "one-omega ratio model" vs "two omega ratio model" to detect the change of selective regime in Sciaphila using branch test in codeml.

if all branches have to be tested in this case, p-value threshold may become very very low because there are a number of branches in tree.

Thank you.

Sincerely,

Kim

Reply

Anonymous 30 June 2016 at 00:42

Thank you for your clarification

But I am still confused.

For instance, if I have 10 taxa and 5 genes.

((((a1,(a2,a3)),a4),(a5,(a6,a7)),a9),a10)

a7 is my interesting taxa.

In this case, null hypothesis is same w across 10 taxa (H0).

And alternative hypothesis is a6 has different w against other 9 taxa (H1) or (a6,a7) has different w against other 8 taxa (H2).

I can get p-value from (H0-H1) and (H0-H2). In this case, I will have p-value threshold = α/2 ? or =α/10? I think previous one is right. isn't it?

Sincerely,

Kim

Reply

Replies

Romain Studer 30 June 2016 at 15:24

I would rather use p-value threshold = α/5, because you have 5 gene families. You will perform 10 tests (5* H0-H1 and 5*H0-H2), but H0-H1 and H0-H2 are not the same models.

Romain

Anonymous 1 July 2016 at 05:14

Thank you.

But I just want to know the change of selective regime of each gene. Sorry, I didn't say this.
So numbers of genes seem not to affect to p-value threshold.
In this case, the p-value of H0-H1 and the p-value of H0-H2 is just α?

In the paper,"The Highly Reduced Plastome of Mycoheterotrophic Sciaphila (Triuridaceae) Is Colinear with Its Green Relatives and Is under Strong Purifying Selection",
They compared two models (one ratio model vs two ratio model) for 18 genes like as above example (H0 vs H1). According to your explanation, they should use p-value thresold = α/18 but they didn't. (Supplementary table3)

H0,H1, and H2 in above example are come from same data.

p-values from LRT of H0-H1 and H0-H2 show which model (H0 vs H1 , H0 vs H2) is fit to data.

So I think that p-value threshold of each comparison(H0 vs H1, H0 vs H2) should be divided 2.

Is it wrong?

I am sorry for taking your time.

Kim

Romain Studer 1 July 2016 at 14:42

Dear Kim,

In table S3, this is single test applied on 19 genes families (there are 19 tests, not 18), so it would make sense for me to use p-value/19. Because they tested 19 different families and wanted to find those which are significant.

Another possibility I would have done is to use QVALUE, and set up either a conservative threshold (q<0.05 => in that case, none of the 19 tests are significant) or a less conservative threshold (q<0.10, in that case, 5 tests are now significant).

I agree it is definitively not an easy question to choose which is the best way to correct for false-discovery rate.

Best regards,
Romain

Anonymous 4 July 2016 at 00:55

Dear Romain

How can I thank you your help!

I can make sense.

But I think they might use p-value/2 because they announced model M1 of clpP was significant after Bonferroni correction in table s3.

So it made me confused.

I really thank you for your clarification.

Sincerely,

Kim

Reply

Unknown 20 July 2016 at 09:05

Hello,
I am new with codeml, and when I try to perform an analysis, the program returns me this error:

1153 nucleotides, not a multiple of 3!

C:\pamlX\paml4.9a\bin\codeml.exe -- killed

What does it mean? As I see it, the alignment file must be edited so the length will be a multiple of 3. Am I right?

Sincerely,

Nicolas

Reply

Replies

Romain Studer 20 July 2016 at 12:27

Hi Nicolas,

The models to estimate selective pressure in CodeML works with codons, so nucleotide triplets. This is why it first perform a check if the alignment you provide really code for protein sequence, or if it is just a set of nucleotide sequences. So you need to be sure that the sequences you provide are coding DNA sequences. Be careful of any frameshifts in case you remove one or two columns.

Best regards,
Romain

Reply

Unknown 26 July 2016 at 08:58

Hi Romain,
Thanks for your answer to my previous question.
I have some more questions for you.
I have a sequence file with 4 sequences,which are orthologous for the same gene on different species at the same genus.
If I´m not wrong, I can´t perform an analysis with so few amount of sequences.
Even so,is there a way to analyze them being so few sequences?

Best regards,
Nicolas

Reply

Replies

Romain Studer 27 July 2016 at 01:14

Hi Nicolas,

You can perform CodeML analysis, but you might have problem with power and accuracy.

Accuracy and Power of Bayes Prediction of Amino Acid Sites Under Positive Selection
http://mbe.oxfordjournals.org/content/19/6/950.full

Accuracy and Power of the Likelihood Ratio Test in Detecting Adaptive Molecular Evolution
http://mbe.oxfordjournals.org/content/18/8/1585.full

Could you increase the number of orthologs? Even some close outgroup could help. It will also on your specific biological question(s).

Romain

Unknown 27 July 2016 at 11:26

Hi Romain,
Thanks for the answer. I have already performed an analysis with the four sequences of interest and

their orthologs (16 seqs).
But now I have another problem. When I run codeml under the null hypothesis(fix_omega=1, omega=1), the program return me this error:

Error: p0 + p1 too small for branch&site model?

I don´t know what it means.

Best regards,

Nicolas.

Romain Studer 27 July 2016 at 12:10

Hi Nicolas,

I don't know what is the problem. I can have a look at your data if you want. Please send me the alignment, tree and codeml command file. (email: rstuder [at] ebi.ac.uk)

Best regards,
Romain

Unknown 27 July 2016 at 14:16

Hi Romain,
I run again codeml under the null hypothesis, but changing CodonFreq settings. I was using CodonFreq =3 (codon table), but then I change it to CodonFreq = 2 (F3X4) and the program ends without any error.

Nicolas.

Romain Studer 1 August 2016 at 02:15

Hi Nicolas,

Good to hear it worked.

In general, try to use the parameters I gave in the control files in this tutorial. There are usually working well.

Best regards,
Romain

Reply

Jingwen 6 August 2016 at 11:39

This comment has been removed by the author.

Reply

Jingwen 6 August 2016 at 11:41

Hi Romain,
I am wondering how to retrive the sites under positive selction from BEB and then get the figure you shown in the last by using Jalview. Is there any way to load these sites into Jalview?
Actually, I am quite confused with CodeML, so I used Nielsen Yang.bf in Hyphy instead. It only shows one likelihood.Do you know how to carry out the same test as you described above using Hyphy instead? Thank you.

Best Regards,
Jingwen

Reply

Replies

Romain Studer 7 August 2016 at 11:25

Hi Jingwen,

For Jalview, I identified the sites with BEB>95%, and highlighted them in Jalview, using standard commands from Jalview.

I don't have experience with Hyphy, sorry. You would better contact the authors of Hyphy.

Best regards,
Romain

Jingwen 8 August 2016 at 07:06

Hi Romain,

Thank you so much for your reply. I have found the way to retrive such sites.
Further, I am wondering is it possible to find the sites under positive selection of a specific clade(subtree of a phylogeny)? I tried to use the subtree cut from the whole tree with sequences of this clade to build a site model, but the result seems to be strange since there are so many such sites.

Thank you!
Best Regards,
Jingwen

Romain Studer 16 August 2016 at 02:12

Hi Jingwen,

Sorry for the late answer, I was travelling.

Normally, the branch-site model should help you to find the sites under positive selection. If you are still strugling to understand the results, you can send me the data (files) to me (rstuder [at] ebi.ac.uk) and I will try to have a look at it.

Best regards,
Romain

Jingwen 17 August 2016 at 06:52

Dear Romain,

Thank you for your reply. I have sent an email including the data and contronl files.
I am very appreciated for your help.

Best Regards,
Jingwen

Reply

joanna 13 September 2016 at 07:56

Hi Romain, thanks so much for this website - it's extremely useful! I was wondering if you could comment on the following situation: the LRT test for M7 vs M8 is not significant but there are few BEB sites with probability >0.99 reported. Do BEB sites mean anything at all when LRT test excluded the positive selection? Thanks again for all your time.

Reply

Replies

Romain Studer 13 September 2016 at 08:00

Hi Joanna,

Normally, when the LRT excluded positive selection, we cannot conclude anything. The M7vsM8 is prone to some error. I would recommend to run model M8a to compare to your M8 result. How many sequences do you have?

Otherwise you can send me your files (aln, tree, ctl, mlc) and I can have a look at it if you want.

Best regards,
Romain

joanna 13 September 2016 at 08:07

Thanks for such a quick reply - I think the problem might be that I have only 8 sequences (and gene itself is only about 400 nucleotides long). I will try the M8a vs M8 comparison, thank you very much for the suggestion.

Reply

Unknown 29 September 2016 at 23:38

Hi Romain, thanks for very useful post. For some cases of my analysis, I obtained zero degree of freedom. What does it mean?
lnL(ntime: 14 np: 19): -4167.463182 (lnL1)
lnL(ntime: 15 np: 19): -4173.059304 (lnL0)

Thanks,
Denisa

Reply

Replies

Romain Studer 3 October 2016 at 09:46

Hi Denise,

Sorry for the late answer.

Which model did you compare? What are lnL1 and lnL0 there? To perform a Likelihood-Ratio Test (LRT), you need to have nested model, so the alternative model will have more parameters than the null model.

Best regards
Romain

Unknown 4 October 2016 at 00:58

It was branch-site model NSsites 2, but I've already found what was wrong and everything works perfectly now :)

Best regards,
Denisa

Reply

Rajesh Kumar Gazara 10 October 2016 at 08:27

Hi Romain,

First of all, I would like to say thanks about this blog and the important information. I tried with my sequence files and I am getting this runtime problem:

Counting codons..
NG distances for seqs.:
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112

49728 bytes for distance
200080 bytes for conP
0 bytes for fhK
5000000 bytes for space

Species 0.456?

I don't know where I did something wrong.
Do you know why I am getting this error??

Reply

Replies

Romain Studer 10 October 2016 at 08:32

Hi Rajesh,

I guess the problem comes from your tree format. Do you have any boostrap information from the tree? If yes, this could be the problem. Otherwise, you can send me your files (alignment, tree and control file) and I can have a look at them later today.

Romain

Rajesh Kumar Gazara 10 October 2016 at 08:42

Hello Romain,

Thank you very much. I am checking according to your suggestion. if I will not be able to solve this issue. I will send the files that you have mentioned. Thanks one again.

Rajesh

Reply

Rajesh Kumar Gazara 10 October 2016 at 10:46

Hi Romain,

Thanks a lot. I fixed it, you were right :)

Reply

Replies

Romain Studer 10 October 2016 at 12:32

You are welcome!

Romain

Rajesh Kumar Gazara 18 October 2016 at 04:31

Hi Romain,

If you don't get any significant positive selection site using codeml but you are still getting functional divergence site from different tools?? What does it suggest?? Do you have any idea about it?

Romain Studer 18 October 2016 at 09:49

Hi Rajesh,

It suggests that either CodeML doesn't have enough power on the dataset (i.e. too close or too divergent sequences) or that the other tools are wrong. Remember that all these tools (CodeML, Diverge, etc...) are predictive methods, so they are as wrong as any models. They will give some clues on which residues could be under positive selection. Then ideally you should go back to the lab to make some experiments to test them in vitro / in vivo. Or least develop some hypothesis based on 3D structure or knowledge you have on these sites.

Best regards,
Romain

Reply

Rajesh Kumar Gazara 18 October 2016 at 11:32

Hi Romain,

Thank you for your suggestion!

Reply

Anonymous 2 December 2016 at 05:03

Dear Romain,

I tested a few genes for positive selection, and in one (rbcL) there is positive selection, however in the .mlc file the BEB method (and NEB) are missing. Do you have to specify anything specific in the ctl-file for the BEB-method to be printed? I already tried changing verbose to 2.

I am using Paml4.8 and these were the commands in the ctl-file used for the alternative model:
seqtype = 1
cleandata = 1
verbose = 2
noisy = 9
runmode = 0
method = 0
clock = 0
getSE = 1
RateAncestor = 0
CodonFreq = 2
estFreq = 0
model = 2
aaDist = 0
NSsites = 2
icode = 0
Mgene = 0
fix_kappa = 0
kappa = 2
fix_omega = 0
omega = 1
fix_alpha = 1
alpha = 0
Malpha = 0
ncatG = 4
ndata = 1
Small_Diff = 5e-7
fix_blength = 0

Thanks for any help,
Best regards,
Paul

Reply

Replies

Romain Studer 2 December 2016 at 05:52

Hi Paul,

No, nothing to add in the ctl file to get the BEB values.

You use model=2, NSsites=2, so the branch-site model. It could be that codeML was able to identify positive selection but was unable to identify exactly which sites are under positive selection (no BEB score >50%). So BEB and NEB are missing in the mlc file. In the rst file, you could find all BEB values for any sites for any class.

If you still have trouble, feel free to send me an email (rstuder [@] ebi.ac.uk), with mlc, ctl and rst files.

Best regads,
Romain

Anonymous 6 December 2016 at 03:06

Hi Romain,

Thank you for your reply. I already have found the problem, there were a few species which had identical dna sequences for the gene rbcl, after removing those 'duplicates' codeml did calculate and include the BEB (and NEB)-scores in the output file.

Best regards,
Paul

Reply

Rajesh Kumar Gazara 16 December 2016 at 09:25

Hi Romain,

Do you know about extended clade model? If yes, can you tell me what parameter should I use or any link related to extended clade model's example??

Thanking very much. Looking for positive response.

Regards,
Rajesh

Reply

Replies

Romain Studer 19 December 2016 at 11:23

Hi Rajesh,

No, I don't have any experience using these clade models, which is an extension of the clade models C. Actually, there are very few positive results in the literature using these clade models, to my knowledge.

The reference paper:
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0018630

Best regards,
Romain

Reply

Rajesh Kumar Gazara 22 December 2016 at 07:17

Hi Romain,

Thanks for reply. I will read this reference paper.

Regards,
Rajesh

Reply

Unknown 7 November 2017 at 11:54

This is a super helpful tutorial! Thanks a bunch. I am wondering if the indicated foreground branch includes just the one branch labeled, or also includes the branches nested within?

In my particular analysis, I want to know if a particular family of New World monkeys is under positive selection for my gene locus.

Do I use branch sites for this or the clade models?

Reply

Replies

Romain Studer 9 November 2017 at 11:43

Dear Susanna,

The indicated foreground is only the one branch labelled. All other branches before and after the adaptive event are considered to be under negative selection/neutral evolution.

The test that also used the nested branches is the clade models, but I don't use it, and there are very few examples in the literature.

In your case, I recommend the branch-site test.

Best regards,
Romain

Unknown 12 January 2018 at 06:47

Thanks for all your help Romain. I have a follow up question that is similar to the thread below.

I want to use a species tree for my analysis, but I dont have distance estimates or large enough portions of the genomes to make branch length estimates for a few of my study species. But I have the sequence of my focal gene, which is only 27 bp long. so branch length estimates will not be strong regardless. Plus, a bunch of sequences are identical across species, which is sort of why I am doing this analysis. But Ive seen above that having identical sequences can cause errors?

my question is: is it legitimate to use a species tree and with my only my focal gene sequences for estimating branch lengths?

Reply

Anonymous 29 November 2017 at 13:28

Dear Roman,

this is rally super helpful tutorial. Thanks for providing it. I am confused about the tree which should I use gene tree of species tree. I am working on 12 bacteria from same genus (different species). I generated the species tree by concatenating single copy orthologs. I need to run codeml for all the ortholog clusters. I have the sequence alignments and ML gene tree for each alignments. I want to compute PS using branch-site model. My question is should I use same 'species tree' for all the clusters I have or I should use different corresponding gene tree of each alignment?

With regards
Singh

Reply

Replies

Romain Studer 29 November 2017 at 23:32

Dear Singh,

I would personally use the species tree. If you are sure they are orthologs, it will be more robust.

Best regards,
Romain

Anonymous 30 November 2017 at 04:16

Thank you Romain!

Reply

Anonymous 13 December 2017 at 04:18

Dear Romain,

As suggested by you and several papers I end up with the using species tree for detecting PS. However, in case of genes where I found the evidence of recombination, I have different phylogeny for each nonrecombinant fragment. As such gene trees could significantly deviate from the phylogenetic patterns species tree, in this case, is it OK to use species tree for PS analysis?

With regards
Singh

Reply

Unknown 21 June 2018 at 16:33

Dear Romain,

I would like to have your suggestion about the parameter runmode=2. I run codeml with runmode=2. Is it possible to use best trees from output of codeml to be used in next codeml run in runmode=0 ?

Reply

Unknown 9 October 2018 at 05:35

Dear Romain,

I would like to know the meaning of "check convergence" in the output of branch-site model. It is asking me to check for convergence of the parameter estimates in PAML or to check for convergent selection?? Sorry for the basic question but I am confused.

thanks a lot

Reply

Vignesh 20 February 2019 at 08:27

Hi Romain,

I have a doubt regarding varying initial omega values.
If I run model M0 and use the branch lengths from that model as initial values for other models, do i need to run codeml multiple times because in this case it will use the same initial values for ML estimation of parameters and every run will give the same result more or less.

Reply

Lowzenza 13 August 2019 at 19:03

Hey Romain,

DO you know of any tutorials to test for rate of synonymous substitutions or codon bias for highly conserved genes? I read about fmutsel but there is no explanation on how to set up the control file to run it properly in the PAML documentation.

None of my genes are under episodic positive selection so now I want to test if maybe codon bias is occurring instead.

Reply

Mah 6 June 2020 at 08:16

Hi, Thanks for the helpful information.
I have SNPs alignment files (in Phylip format) for different genes of a gene family. The SNPs have been identified from different accessions (882 accessions) of the same species. My objective is to identify positive selection and see if any alleles or genes have been under positive selection among those accessions. For example for gene #1 I have the alignment file taken from 882 samples and so on.

I want to use CODEML for this but I am not sure how because basically I don't know if I can do it without choosing foreground and backgrounds? cause I don't know which one is more interesting!
would be grateful for your help!
Mah

Reply

Replies

Romain Studer 14 June 2020 at 07:52

Hi,

I don't know if CodeML can handle SNP. Maybe better to look to other tools, like HyPhy, if they can handle it.

Best regards,
Romain

Reply

Enter your comment...

Comment as: Google Accour

Publish Preview

Newer Post Home Older Post

Subscribe to: Post Comments (Atom)

Powered by Blogger.