PAML ACTIVITIES

2022 Workshop on Molecular Evolution at the MBL

Follow this link to go to back to workshop schedule page.

Follow this link to go to back to the Bielawski faculty page.

Overview

The objective of this activity is to help you understand how to use different codon models, and how to test for selection using PAML (and specifically the CODEML program). The activities are designed to build general analytical skills, and are just as relevant to analyses carried out using other software packages, such as HyPhy.

The tutorial is divided into 4 exercises.

1. Maximum likelihood estimation (by hand)
2. Sensitivity of ω
3. LRTs for alternative hypotheses about temporal changes in selection pressure
4. Test for sites evolving by positive selection in the *nef* gene of HIV-2

The next section (**Accessing the files**) describes how to access and work with the files required for each of the exercises. You will access the files differently depending on whether you are doing the labs at the workshop or at home independently of the workshop.

Note that running the program involves modifying input files and creating output files. *It is best practice to (i) create a separate directory for each exercise or real-data analysis that you want to do and (ii) record and save simple-text notes about the motivation and details of each real-data analysis and store them within the directory that created for each analysis*. The latter is a kind of “read me” file that you will be glad to have after you have done many analyses on your real data and have begun to forget, or mix up, the details.

Here are the slides for the PAML learning activity: 2020_slides(v2)

Here is a copy of the book chapter that accompanies these exercises: Book_Chapter.pdf

This page has links to useful documents and many **additional resources**.

Accessing the files

1. If you are doing the lab AT THE WORKSHOP: On the virtual machines we will be using in the 2022 workshop, there will be a symlink in your home directory named "moledata" that takes you to the course data files. There you will find directories for the various labs (e.g., MSAlab, revbayes, PamlLab, etc.).

To view the list of labs type:

```
ls ~/moledata
```

To view the contents of the Paml Lab type:

```
ls -1 ~/moledata/PamlLab
```

This will reveal the directories for each excercise:

```
ex1
ex2
ex3
ex4
```

The files are already on the virtual machine you are using. *However, you will want to run each exercise in a separate directory that you will create*. So, create a new directory. The name of the new directory is up to you, but pick something informative (e.g., ~/PAML_ex1).

To *copy* the files required for exercise 1 just type:

```
cp ~/moledata/PamlLab/ex1/* ~/PAML_ex1
```

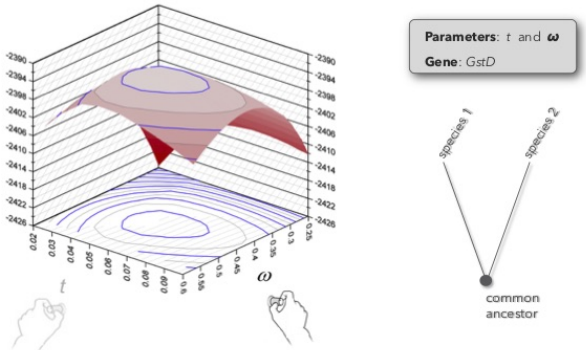
Now you are ready to do exercise 1 within `~/PAML_ex1`

2. If you are doing the lab INDEPNDENTLY of the workshop: You can do this lab by downloading all the necessary files from an archive here, or you can download the files individually for each exercise as you need them here.

Either way, it is still *"best practice" to run each exercise in a separate working directory that you will create* (*e.g.*, `PAML_ex1`), *and work with copies of the required files within that directory*.

Exercise 1

The objective of this activity is to use CODEML to evaluate the likelihood of the *GstD1* sequences for a variety of ω values. Plot log-likelihood scores against the values of ω and determine the maximum likelihood estimate of ω. Check your finding by running CODEML's hill-climbing algorithm.



You will work with THIS "knob"

1. Find the input files for Exercise 1 (**ex1_codeml.ctl, seqfile.txt**) and familiarize yourself with them. Pay close attention to the contents of the modified control file called **ex1_codeml.ctl**.

2. Remember to create a directory where you want your results to go, and place all your files within it. Now open a terminal, move to the directory that contains your files. When you are ready to run CODEML, delete the **ex1_** prefix (the control file must be called **codeml.ctl**). Now you can run CODEML.

3. Familiarize yourself with the results (see annotations in ex1_HelpFile.pdf). If you have not edited the control file the results will be written to a file called **results.txt**. Identify the line within the results file that gives the likelihood score for the example dataset. and record it

4. Now *change and save* the control file and re-run CODEML for a different fixed value of ω. The control file "quick guide" might be helpful here (quick guide). The objective is to compute the likelihood of the example dataset given a fixed value of ω. *Change the control file as follows*:

   - Change the name of your result file (via `outfile=` in the control file) or you will overwrite your previous results!

   - Change the fixed value for omega by changing the value for `omega=` in the control file. The values for this exercise are provided as comments at the bottom of the example control file that has been provided to you.

5. Repeat Step 4 for each value of ω according to the comments of the example control file (*e.g.*, ω = 0.005, 0.01, 0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 2.0).

6. Use your favorite spread sheet or statistical package to plot the likelihood score (y-axis) against the fixed value for omega (x-axis). Use a logarithmic scale for the x-axis (do not

transform *ω*). Your figure should look something like this: ex1 plot template.pdf (note: the data points have been intentionally omitted from this version of the plot; you need to generate the data for yourself).

- For help plotting your results see the additional resources on this page.

7. From your plot, try to answer this question:

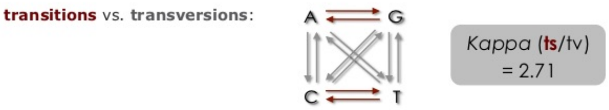
   - *What is the value of ω that will maximize the likelihood score (i.e., the MLE)*?

8. Now change the control file so that CODEML will use its hill-climbing algorithm to find the MLE; set `fix_omega=0` in the control file. Compare the result to your guess from Step 7.

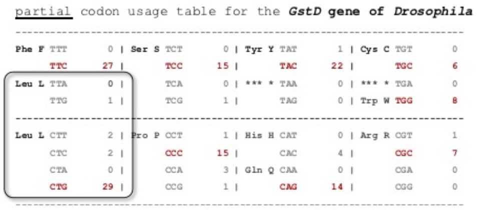
   - *How good was your estimate of the MLE*?

Exercise 2

in this exercise you will investigate the sensitivity of ω to the transition/transversion ratio (*kappa*), and to the assumed codon frequencies (*π*'s). After you collect the required data you will determine which assumptions yield the largest and smallest values of S, and what effect it has on the value of *ω*.



preferred vs. **un-preferred** codons:



```
partial codon usage table for the GstD gene of Drosophila
---------------------------------------------------------------------
Phe F TTT       0 | Ser S TCT       0 | Tyr Y TAT       1 | Cys C TGT       0
      TTC      27 |       TCC      15 |       TAC      22 |       TGC       6
Leu L TTA       0 |       TCA       0 | *** * TAA       0 | *** * TGA       0
      TTG       1 |       TCG       1 |       TAG       0 | Trp W TGG       8
---------------------------------------------------------------------
Leu L CTT       2 | Pro P CCT       1 | His H CAT       0 | Arg R CGT       1
      CTC       2 |       CCC      15 |       CAC       4 |       CGC       7
      CTA       0 |       CCA       3 | Gln Q CAA       0 |       CGA       0
      CTG      29 |       CCG       1 |       CAG      14 |       CGG       0
---------------------------------------------------------------------
```

1. Find the files for Exercise 2 (**ex2_codeml.ctl, seqfile.txt**) and familiarize yourself with them. It would be best to create a new directory for Exercise 2. When you are ready to run CODEML, delete the **ex2_** prefix (the control file must be called **codeml.ctl**).

2. Run CODEML using the settings in the control file for Exercise 2. Familiarize yourself with the results (ex2 HelpFile.pdf). In addition to the likelihood score you must be able to identify the part of the result file that provides estimates of the following:

   - Number of synonymous or nonsynonymous sites (S and N))

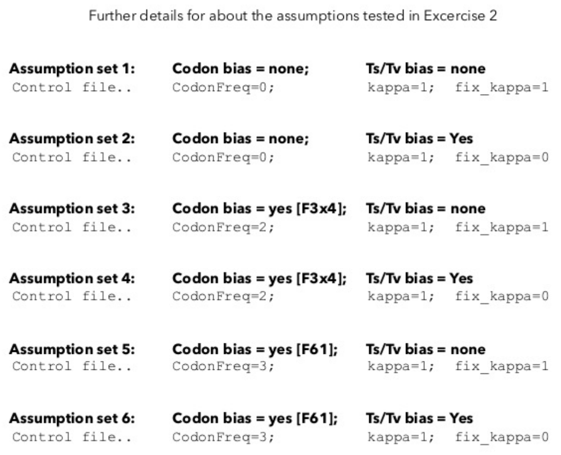
- Synonymous and nonsynonymous rates (dS and dN)

3. As in Exercise 1, you will need to *change and save* the control files and re-run CODEML. The control file "quick guide" might be helpful here (quick guide).The objective is to compute the likelihood of the example dataset under different model assumptions. To do this you must:

   - *Change* the name of the main result file (via `outfile=` in the control file) or you will overwrite your previous results

   - *Change* the model assumptions about codon frequencies (via `CodonFreq=`) and kappa (via `kappa=` and `fix_kappa=`).

4. Repeat step 3 for each set of assumptions about codon frequencies and kappa given as comments at the bottom of the example control file, and shown in the figure below.



Further details for about the assumptions tested in Excercise 2

Assumption set 1:
`Control file..`
	Codon bias = none;
	`CodonFreq=0;`
		Ts/Tv bias = none
		`kappa=1;   fix_kappa=1`

Assumption set 2:
`Control file..`
	Codon bias = none;
	`CodonFreq=0;`
		Ts/Tv bias = Yes
		`kappa=1;   fix_kappa=0`

Assumption set 3:
`Control file..`
	Codon bias = yes [F3x4];
	`CodonFreq=2;`
		Ts/Tv bias = none
		`kappa=1;   fix_kappa=1`

Assumption set 4:
`Control file..`
	Codon bias = yes [F3x4];
	`CodonFreq=2;`
		Ts/Tv bias = Yes
		`kappa=1;   fix_kappa=0`

Assumption set 5:
`Control file..`
	Codon bias = yes [F61];
	`CodonFreq=3;`
		Ts/Tv bias = none
		`kappa=1;   fix_kappa=1`

Assumption set 6:
`Control file..`
	Codon bias = yes [F61];
	`CodonFreq=3;`
		Ts/Tv bias = Yes
		`kappa=1;   fix_kappa=0`

5. In your favorite spreadsheet program create a table like **Table E2** in the slides (see ex2 table template.pdf) and fill it in with your results.

6. Use your table to determine:

   - *Which assumptions yield the largest and smallest values of S*?
   - *What is the effect on ω*?
   - *What model would you choose*?

copy the file /project/sackettl/MolEvol/PAML/exercise3.zip into your work directory, cd to your work directory, and then type

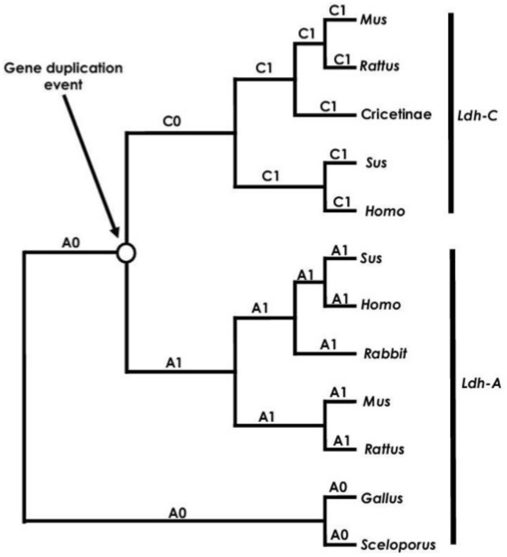
unzip exercise3.zip

Exercise 3

The objective of this exercise is to use three LRTs to evaluate the following possibilities: (1) the mutation rate of Ldh-C has increased relative to Ldh-A, (2) a burst of positive selection for functional divergence occurred following the duplication event that gave rise to Ldh-C, and (3) there was a long term shift in selective constraints following the duplication event that gave rise to Ldh-C.

1. Obtain the files for Exercise 3 from the course web-site (**ex3_codeml.ctl, seqfile.txt, treeH0.txt, treeH1.txt, treeH2.txt, treeH3.txt**). The tree files represent different hypotheses denoted H0, H1, H2 & H3 (see diagram below, or download this file genetree). These evolutionary concepts described above are covered by these four precise hypotheses:

    - H0: Homogeneous selection pressure over the tree (concept = any rate differences are due to changes in mutation rate).
    - H1: Episodic change in selection pressure in Ldh-C (concept = only in the branch that immediately follows the gene duplication event).
    - H2: A long term shift in selection pressure in Ldh-C only (concept = Ldh-C has a novel selection pressure, as compared to its ancestors, whereas Ldh-A remains subject to the ancestral level of selection pressure).
    - H3: A long term shift in selection on both Ldh-C and Ldh-A (concept = both paralogous lineages experience novel selection pressures; i.e., different from each other, and from the ancestor).



H_0: $\omega_{A0} = \omega_{A1} = \omega_{C1} = \omega_{C0}$
H_1: $\omega_{A0} = \omega_{A1} = \omega_{C1} \neq \omega_{C0}$
H_2: $\omega_{A0} = \omega_{A1} \neq \omega_{C1} = \omega_{C0}$
H_3: $\omega_{A0} \neq \omega_{A1} \neq \omega_{C1} = \omega_{C0}$

to execute the program, change to the directory with exercise 3 files and type

/project/sackettl/miniconda3/envs/paml_env/bin/codeml

2. Run CODEML using the settings in the control file for Exercise 3. Familiarize yourself with the results (ex3_HelpFile.pdf). In addition to the likelihood score you must be able to identify the branch-specific estimates of the ω parameter. (In the first run, the branch

specific values for ω will all be the same. In later runs there will be differences among some branches).

3. As in the previous exercises, you will need to *change the control* file and re-run CODEML. The control file "quick guide" might be helpful here (quick guide).The objective is to compute the likelihood, and estimate ω parameters, under different models of how selection pressure changes in different parts of the tree. Because the relevant model information is contained in the tree file, you will need specify one of several tree files in each analysis, and change the control file so that it reads the required tree file.

   - As always, you should change the name of the main result file (via `outfile=` in the control file) or you will overwrite your previous results.

   - Change the model assumptions about branch specific ω values by changing the tree files (via `treefile=` and `model=`) set within the control file.

https://awarnach.mathstat.dal.ca/~joeb/PAML_lab/exercise3/ex3_table_template.pdf

4. Repeat step 3 for each of the four tree files that have been provided to you. Again, keep track of your results by using a table like **Table E3** shown in the slides (see ex3 table template.pdf). In addition, carry out likelihood ratio tests (LRT) of the hypotheses below. See the lecture notes for additional details about LRTs. Use 1 degree of freedom to obtain the *P*-value for each LRT. (Helpful for computing the *P*-value: Chi-Square calculator)

   - H0 vs. H1
   - H0 vs. H2
   - H2 vs. H3

5. Use your table of results to determine:

   - *Which model(s) are supported by the data?*

   TURN IN ANSWERS TO THESE 2 QUESTIONS

   - *What evolutionary scenario is the best explanation of Ldh gene-family evolution?*

   - *Is there evidence of positive selection during the history of Ldh evolution?*

   - *Are there any scenarios in which Ldh could have evolved by positive selection that would be undetectable by these LRTs?*

copy the file /project/sackettl/MolEvol/PAML/exercise4.zip into your work directory, cd to your work directory, and then type

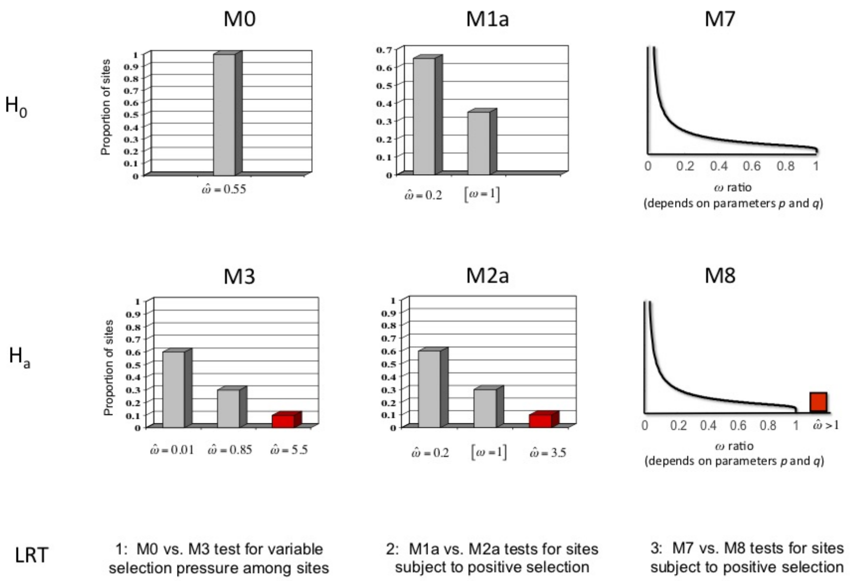
unzip exercise4.zip

Exercise 4

The objective of this exercise is to use a series of LRTs to *test for sites* evolving under positive selection in the nef gene. If you find significant evidence for positive selection, then identify the involved sites by using empirical Bayes methods.

1. Obtain all the files for Exercise 4 (**ex4_codeml.ctl.txt**, **ex4_seqfile.txt**, **treeM0.txt, treeM1.txt, treeM2.txt, treeM3.txt, treeM7.txt, treeM8.txt**). When you are ready to run CODEML, remember to delete the **ex4_** prefix (the control file must be called **codeml.ctl**).

2. If you plan to run two or more models at the same time, then create a separate directory for each run and place a copy of the sequence file, and the required control file and tree file in each directory.



3. As in all the previous exercises, you will need to *change the control file* and re-run CODEML several times. In this case you will be fitting six different codon models (M0, M1a, M2a, M3, M7 & M8) to the example dataset. The control file "quick guide" might be helpful here (quick guide).

    - If you are running your analyses sequentially in the same directory, then you should change the name of the main result file (via outfile= in the control file) or you will overwrite your previous results.

    - Set the tree file with `treefile=`. I have supplied tree files pre-loaded with the ML branch lengths for each model (hence you need to set a different tree for each model). This will greatly speed up your analyses, giving you more “beer time”. See the example control file for more details about treefile names\

    - Set the codon model with `NSsites=`.

    - Fix the value of kappa at the ML estimate with `kappa=`. Again, this will help speed up the analysis. See the control file for the value of kappa for each model.

    - For some models you will also need to set the number of categories (ncatG) in the ω distribution:

        - for M3 set `ncatG=3`
        - for M7 set `ncatG=10`
        - for M8 set `ncatG=10`

after you change your control file, to execute the program, make sure you are in the directory with exercise 4 files and type

/project/sackettl/miniconda3/envs/paml_env/bin/codeml

    - Once the analysis is complete, rename the rst file because subsequent runs will overwrite it!
    - Repeat steps for each of the six codon models listed above.

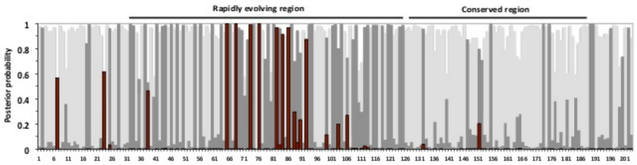
4. Keep track of your results (ex4_HelpFile.pdf) by using a table like **Table E4** shown in the slides (see ex4 table template.pdf).

    Models 2 and 3 table says should have PSS — count # positively selected sites in rst file

5. In addition, carry out the following likelihood ratio tests:
    - M0 vs. M3 (4 degrees of freedom)
    - M1a vs. M2a (2 degrees of freedom) You do not need to do the LRTs - just record the likelihood
    - M7 vs. M8 (2 degrees of freedom)

6. Lastly, open the rst file generated when you ran model M3 (ex4_rst-HelpFile.pdf). Locate the columns of posterior probabilities for each site under the three site-categories of this model. Use these data to produce the plot for the nef gene like the one shown below (*your plot will look different than the one shown below*).

Example of a posterior distribution of selection pressure among sites

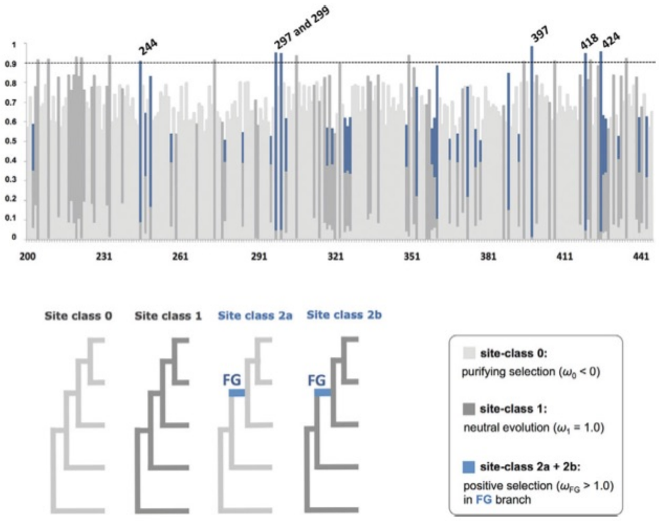


NOTE: This is **NOT** the distribution for the *nef* gene

7. As a final step, try to synthesize all your results and attempt a biological interpretation of the sort that you would want to publish within an actual research paper. The following two general questions should help get you going. I strongly encourage you to do this last step in collaboration with other workshop students; talk it through!
    - *What biological conclusions are well-supported by these data?* TURN IN AN ANSWER TO THIS QUESTION
    - *What aspects of the results can you interpret according your prior biological knowledge of this, or similar, systems?*

NEXT STEPS...

Now that you have some experience with codon models, you might want to try out a tutorial covering more advanced topics. The advanced tutorial focuses on **detecting episodic adaptive evolution** via "Branch-Site Model A".



The tutorial also includes **additional activities** about:

- identifying and labelling phylogenetic tree branches for input to branch-site codon models
- detecting instabilities in your parameter estimates
- carrying out robustness analyses
- use of smoothed bootstrap aggregation (SBA) to correct for parameter estimate uncertainty and instability in codon models

The protocols for each activity are published in *Protocols in Bioinformatics* (UNIT 6.15). This unit also presents **recommendations for “best practices” when carrying out a large-scale evolutionary survey** for episodic adaptive evolution by using PAML.

You can take a look at the PDF file for *Protocols in Bioinformatics* UNIT 6.15 here: UNIT 6.15

The files required for this “advanced lab” are available via this Bitbucket repository: repository-link