Codeml

--mounted on **PAML**(Phylogenetic Analysis by Maximum Likelihood)

Developer：Ziheng Yang

Reporter：Longlong Sang

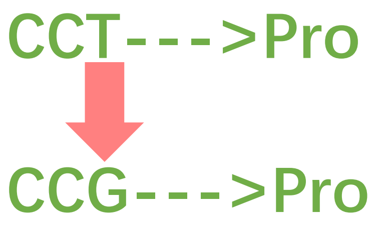
The Lab of Molecular Systematics and Ecology
http://www.lmse.org/

Content

- **Conception**
- **Input files example**
- **Objective**
- **Model interpretation**
- **Summary**

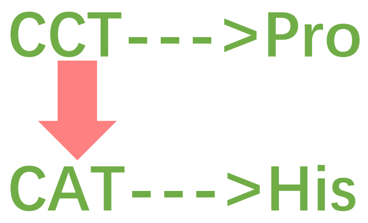
Conception-> Synonymous and Nonsynonymous substitution

- **Synonymous substitution**
  - Nucleotide mutation that not alter amino acids(AA) sequence.
- **Nonsynonymous substitution**
  - Nucleotide mutation that alter amino acids sequence.



CCT--->Pro

CCG--->Pro



CCT--->Pro

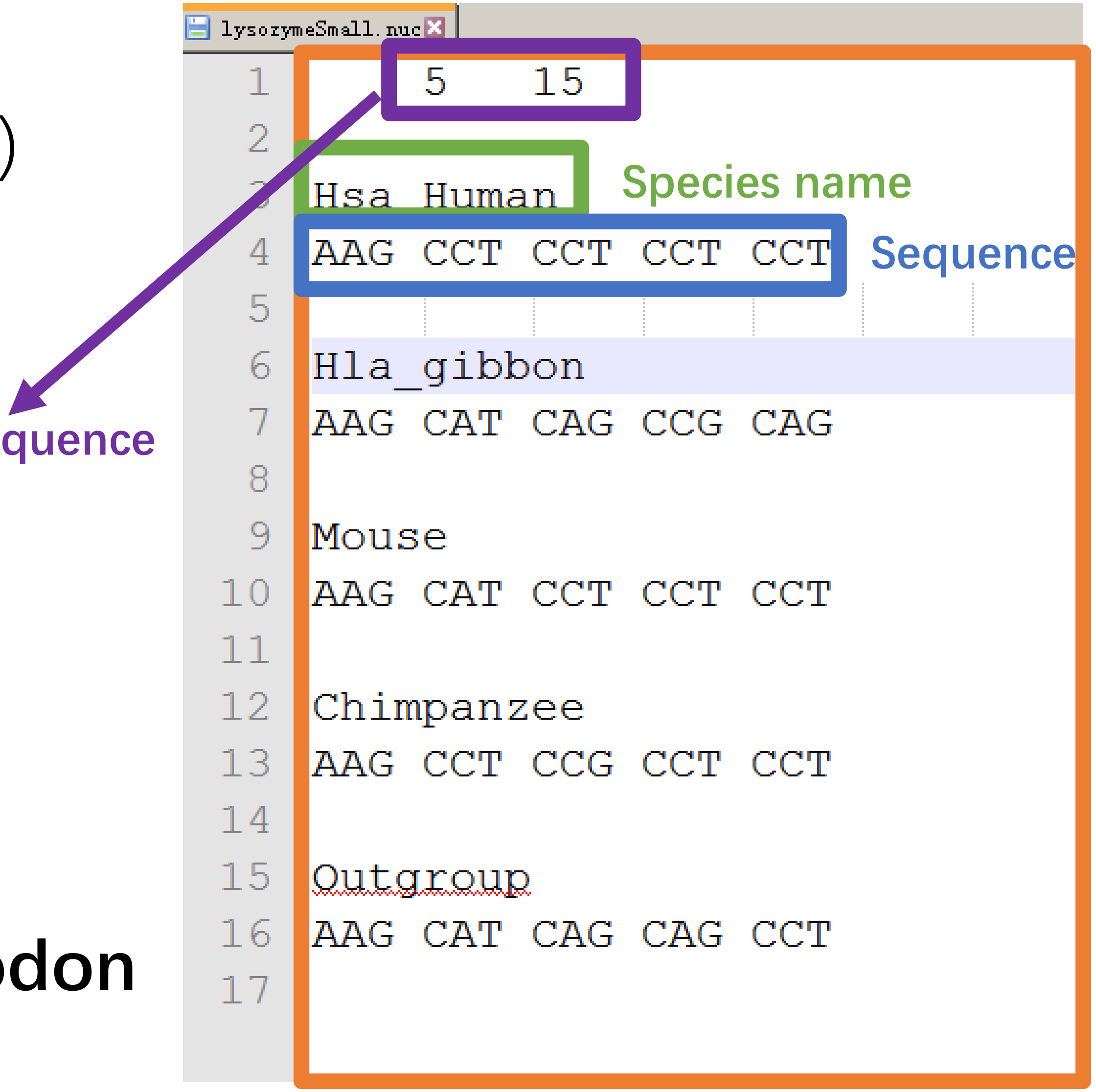
CAT--->His

Omega ω

- **ω = dN/dS**
- **Definition** : The ratio between Nonsynonymous substitution change rate and Synonymous substitution change rate. Measures selective pressure at the protein level.
- **Indicative meaning** :
  - ω > 1 -> positive selection
  - ω = 1 -> neural selection
  - ω < 1 -> negative selection
- **Example :**
  - MHC ω have a higher value, structural protein gene have a smaller value

Input files->Nucleotide file

- Sequence file （suffix： .nuc .txt)



Nucleotide text file content

lysozymeSmall.nuc

```
    5   15

Hsa_Human
AAG CCT CCT CCT CCT

Hla_gibbon
AAG CAT CAG CCG CAG

Mouse
AAG CAT CCT CCT CCT

Chimpanzee
AAG CCT CCG CCT CCT

Outgroup
AAG CAT CAG CAG CCT

```

Species name

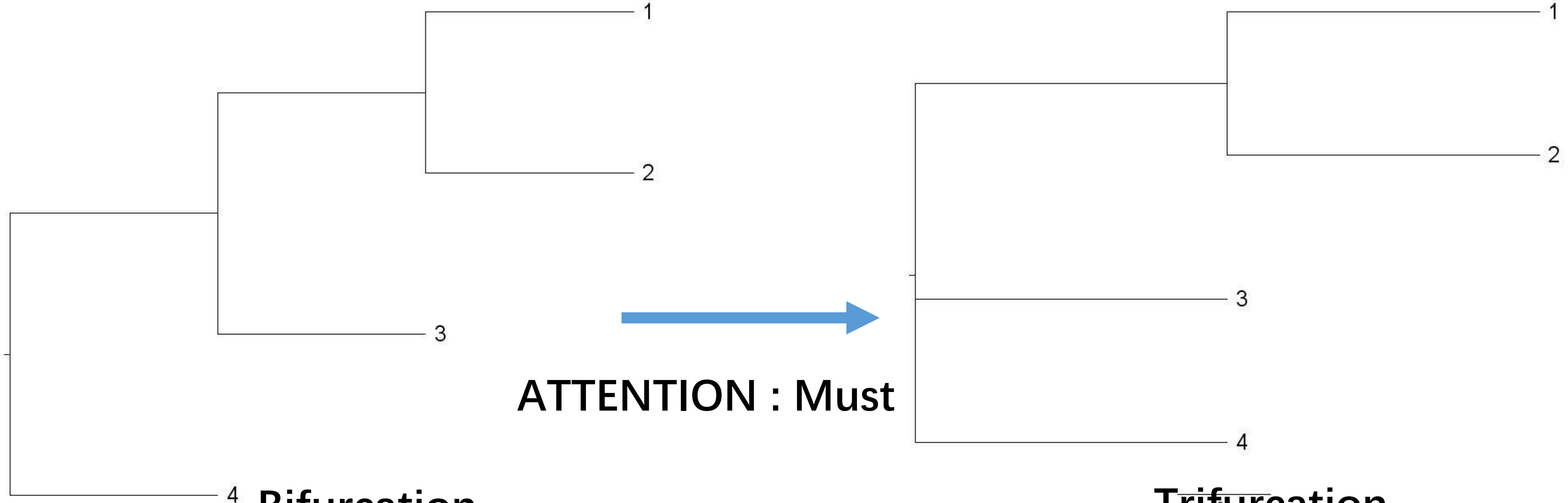
Sequence

Num of samples, length of Sequence

Attention : No need of stop codon

Input files->Tree files

- Un rooted tree is NEEDED
- (((1,2),3),4)-------->((1,2),3,4)



ATTENTION : Must

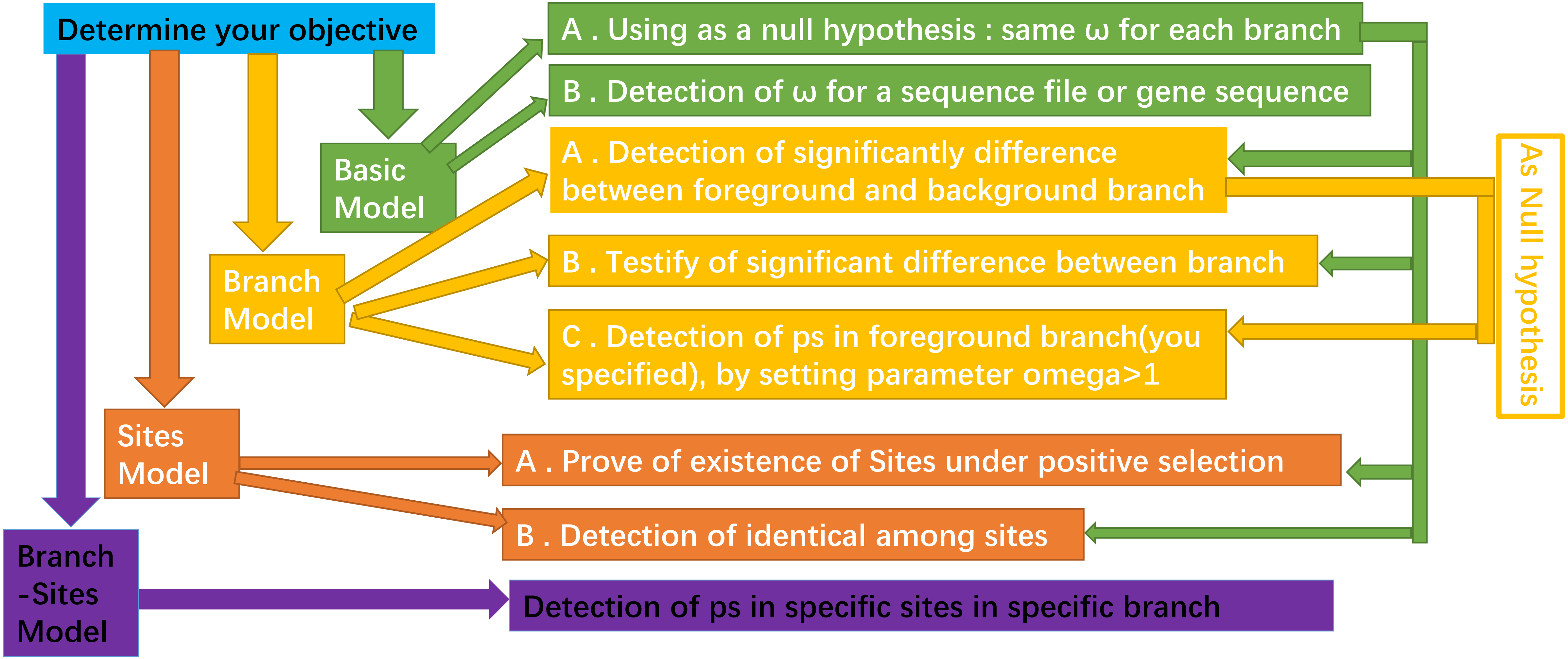
Bifurcation

Trifurcation

Input files－>Configuration file(.ctl)

- seqfile = lysozymeSmall.nuc ＊sequence data file name
- treefile = lysozymeSmall.trees ＊tree structure file name
- outfile = result.txt ＊main result file name
- seqtype = 1
- CodonFreq = 2 ＊0:1/61 each, 1:F1X4, 2:F3X4, 3:codon table
- **model** = 2, choose your model for branch.
- **NSsites** = 0, choose your model for sites/codon.
- fix_omega=0/1, 0 meaning ω value estimate in program, 1 meaning use the value you assign to parameter “omega”.
- omega=you setting
- Other parameter suggest using default.

Functionality and Determine your objective



Determine your objective

Basic Model

A . Using as a null hypothesis : same ω for each branch

B . Detection of ω for a sequence file or gene sequence

Branch Model

A . Detection of significantly difference between foreground and background branch

B . Testify of significant difference between branch

C . Detection of ps in foreground branch(you specified), by setting parameter omega>1

Sites Model

A . Prove of existence of Sites under positive selection

B . Detection of identical among sites

Branch -Sites Model

Detection of ps in specific sites in specific branch

As Null hypothesis

Note : ps meaning Positive selection

Running a Codeml

- Command Line
- Graphics User Interface
- Multiple or Batch running
- Simplify way:

  Four files in same folder:

  codeml.exe

  codeml.ctl

  xx.nuc or xx.txt

  xx.trees

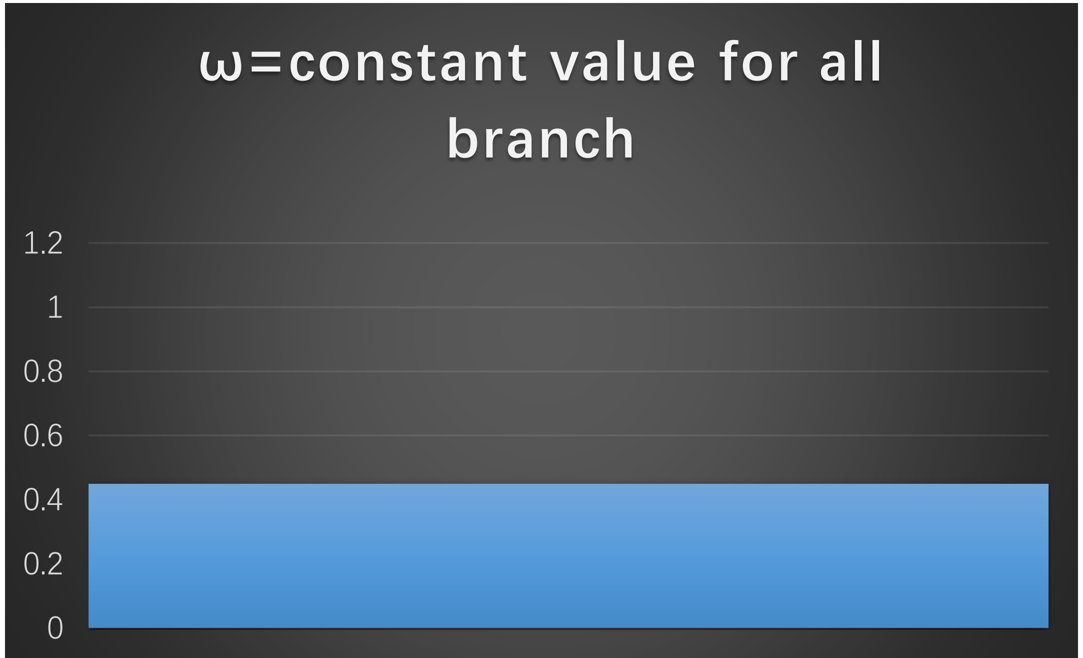
Introduction to Model and realistic usage

Basic Model

- **Assumption** : One ω=dN/dS ratio(omega) for each branches, usually as a null hypothesis for Likelihood ratio tests.
- **Setting** :Model=0 Nssites=0
- **Advice**： As a Null hypothesis； and omega is almost always smaller than 1 when you want

  use Basic as a detection of positive

  selection for sequence or gene.



Branch Model Type

Branch Model->Introduction

- **Assumption** : The value ω is same value or vary between different branches.
- **Three Model for Branch model**

  A. Model=0 : Basic Model

     Setting : model=0 NSsites=0

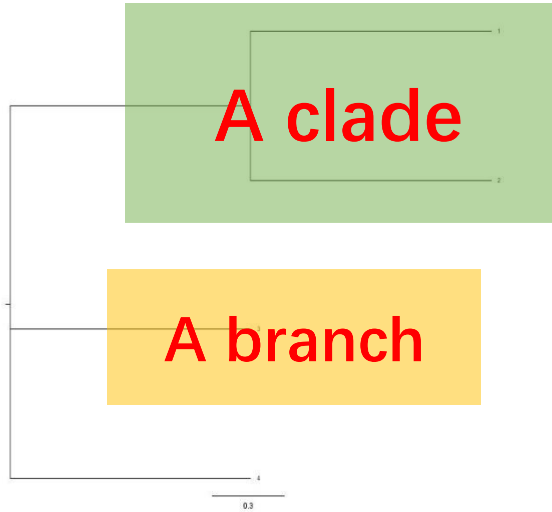
  B. Model=1 : All of branches have their unique ω value

     Setting : model=1 NSsites=0

  C. Model=2 : By label sign in your tree file that you have some clade have their own omega value different with the rest branch

     Setting : model=2 NSsites=0

Branch Model－>How to label in your tree file



- **Label sign : #** for clade, **$** for branch
- **No labeled :**
  - ((Hsa_Human ,Hla_gibbon),(Mouse,Chimpanzee),Outgroup);
- **Labeled :**
  - ((Hsa_Human ,Hla_gibbon)**#1**,(Mouse,Chimpanzee),Outgroup);

    ω_1 ω_0
  - ((Hsa_Human ,Hla_gibbon)**#1**,(Mouse,Chimpanzee)**#2**,Outgroup);

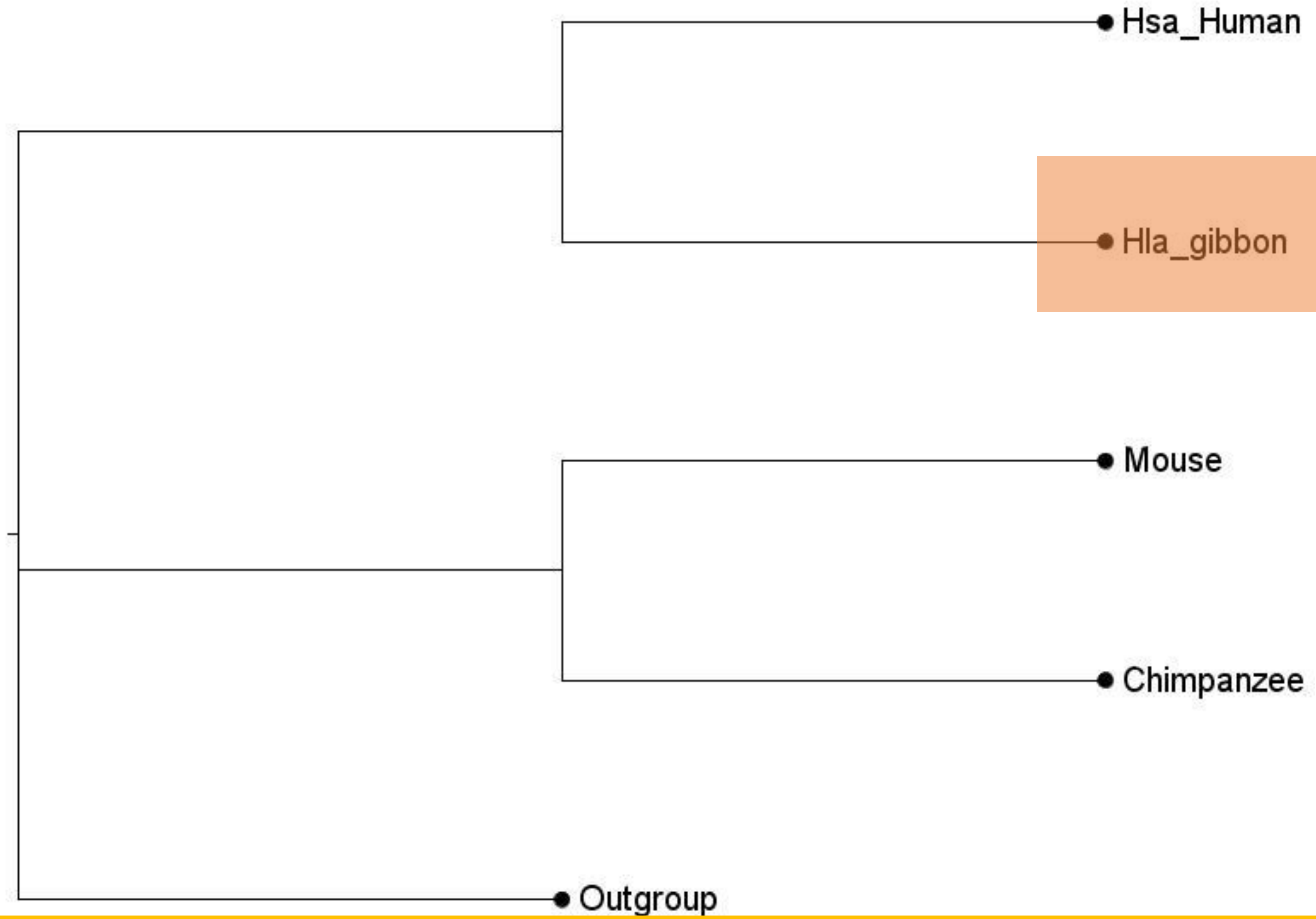
    ω_1 ω_2 ω_0
- **Precedence ：** $>#, sign close to root < sign far from root

Background branch

Branch Model->Three Applications

- Application 1- Testify of significant difference between branch
  - Null hypothesis : Basic Model Alternative hypothesis: model=0, Nssites=0
- Application 2- Detection of significantly difference between foreground and background branch
  - Null hypothesis : Basic Model Alternative hypothesis : model=2, Nssites=0
  - Note : # or $ label is needed.
- **Application 3- Detection of positive selection in foreground branch(you specified)**
  - Null hypothesis : model=2 NSsites=0 Alternative hypothesis : model=2, Nssites=0 fix_omega=1 setting omega > 1
  - Note : The omega set >1 meaning set the **last ω**>1, which is meaning only last class of branch is under positive selection for this model comparison.

Branch Model->Function of Application 3



AAG CCT CCT CCT CCT

AAG CAT CAG CCG CAG

AAG CAT CCT CCT CCT

AAG CCT CCG CCT CCT

AAG CAT CAG CAG CCT

Find out a branch is under positive selection, but only can be foreground branch

Branch Model->Analysis for three Applications

- Test → Get Ln(likelihood) from running-out result

  Example: Null hypothesis result: lnL(ntime: 7 np: 9):   -31.366310   +0.000000

  Alternative hypothesis result: lnL(ntime: 7 np: 10):   -31.366306   +0.000000

- Likelihood ratio tests(LRTs)

  Statistics value: q=2(ln L1-ln L2) → fit to Chi-Square Distribution → Get p value

  Method: R package → function : pchisq(statistic value, df, lower.tail=FALSE) , df=np1- np0 → p-value= 0.9984042

  Meaning : In a situation ,rejecting Null hypothesis is wrong with possibility of   p=0.9984042

Sites Model Type

Sites Model->Introduction

- **Assumption** : Allowing the DN/DS ratio to vary among sites (among codons or amino acids in the protein)
- **Setting**: Model=0   NSsites=number(0..13,22)

  Tips: NSsites = 0 1 2 3 4 meaning you can run several model in one running.
- **Function** : Fitting the Evolution Rate of Different sequence position with Different Models then to prove the positive selection in sites.

Sites Model->Important model and model information

- **Nssites=0,one ω**: one-ratio model, all of position have identical ω ratio
- **NSsites=1, neutral** : ω<=1
- **NSsites=2, selection** : some position : ω>1 ω<=1
- **NSsites=3, discrete** : Discrete Distribution, three ω : $\omega_{0,}$ $\omega_{1,}$ ω_2
- **NSsites=7, beta** : Distribution of ω along position is beta distribution
- **NSsites=8, beta&ω** : based on model 7, but some position ω>1

Sites Model->Important model information

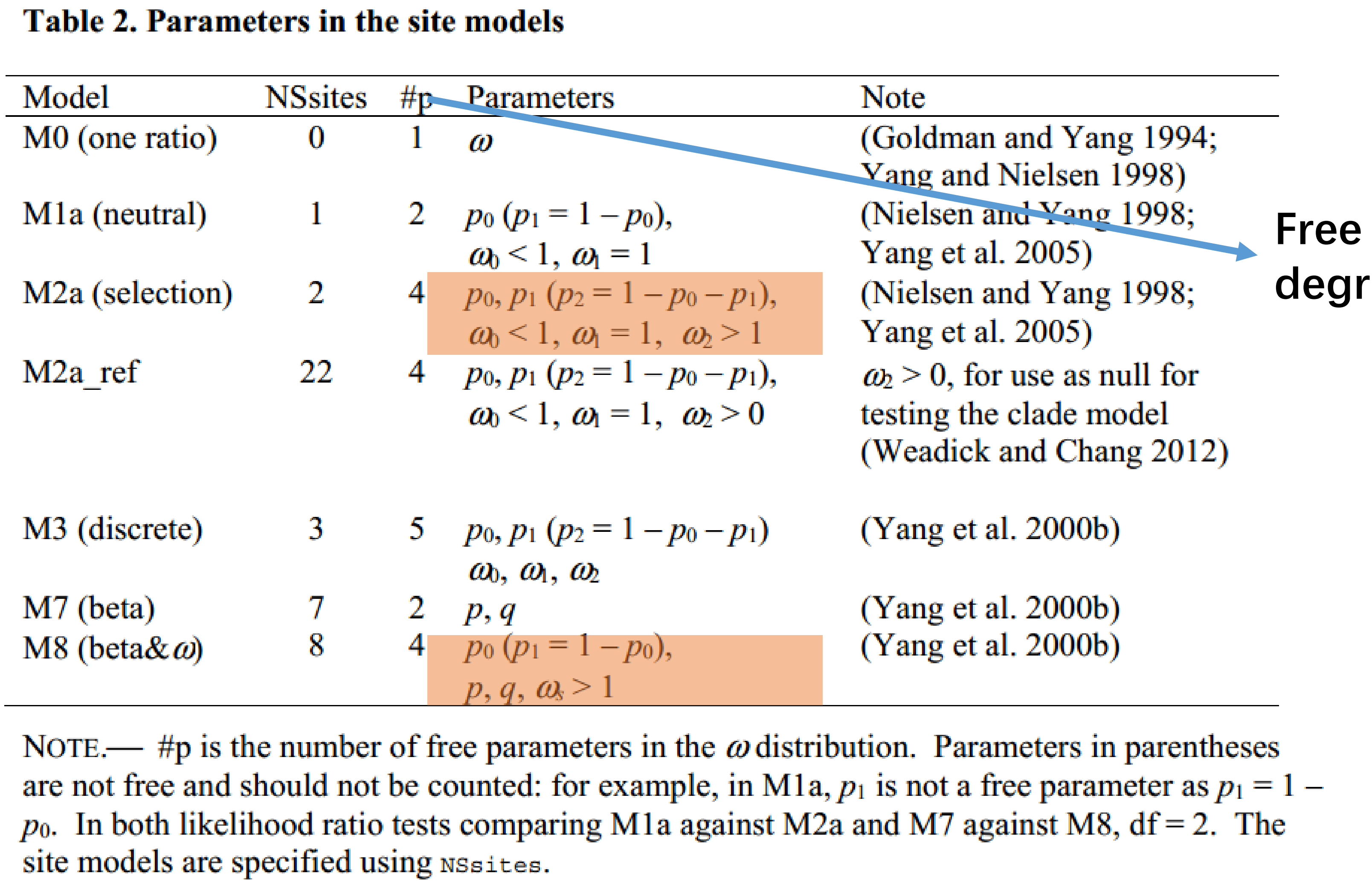


Table 2. Parameters in the site models

Model	NSsites	#p	Parameters	Note
M0 (one ratio)	0	1	ω	(Goldman and Yang 1994; Yang and Nielsen 1998)
M1a (neutral)	1	2	p_0 ($p_1 = 1 - p_0$), $\omega_0 < 1$, $\omega_1 = 1$	(Nielsen and Yang 1998; Yang et al. 2005)
M2a (selection)	2	4	p_0, p_1 ($p_2 = 1 - p_0 - p_1$), $\omega_0 < 1$, $\omega_1 = 1$, $\omega_2 > 1$	(Nielsen and Yang 1998; Yang et al. 2005)
M2a_ref	22	4	p_0, p_1 ($p_2 = 1 - p_0 - p_1$), $\omega_0 < 1$, $\omega_1 = 1$, $\omega_2 > 0$	$\omega_2 > 0$, for use as null for testing the clade model (Weadick and Chang 2012)
M3 (discrete)	3	5	p_0, p_1 ($p_2 = 1 - p_0 - p_1$) ω_0, ω_1, ω_2	(Yang et al. 2000b)
M7 (beta)	7	2	p, q	(Yang et al. 2000b)
M8 (beta&ω)	8	4	p_0 ($p_1 = 1 - p_0$), p, q, $\omega_s > 1$	(Yang et al. 2000b)

NOTE.— #p is the number of free parameters in the ω distribution. Parameters in parentheses are not free and should not be counted: for example, in M1a, p_1 is not a free parameter as $p_1 = 1 - p_0$. In both likelihood ratio tests comparing M1a against M2a and M7 against M8, df = 2. The site models are specified using `NSsites`.

Free parameter or degree of freedom

(Yang PAMl software description file)

Sites Model->Detection of identical of sites

- **Null hypothesis** : All branch or sites ω are identical

  Model=0 and NSsites = 0

- **Alternative hypothesis** : Distribution of ω along sites are Discrete Distribution

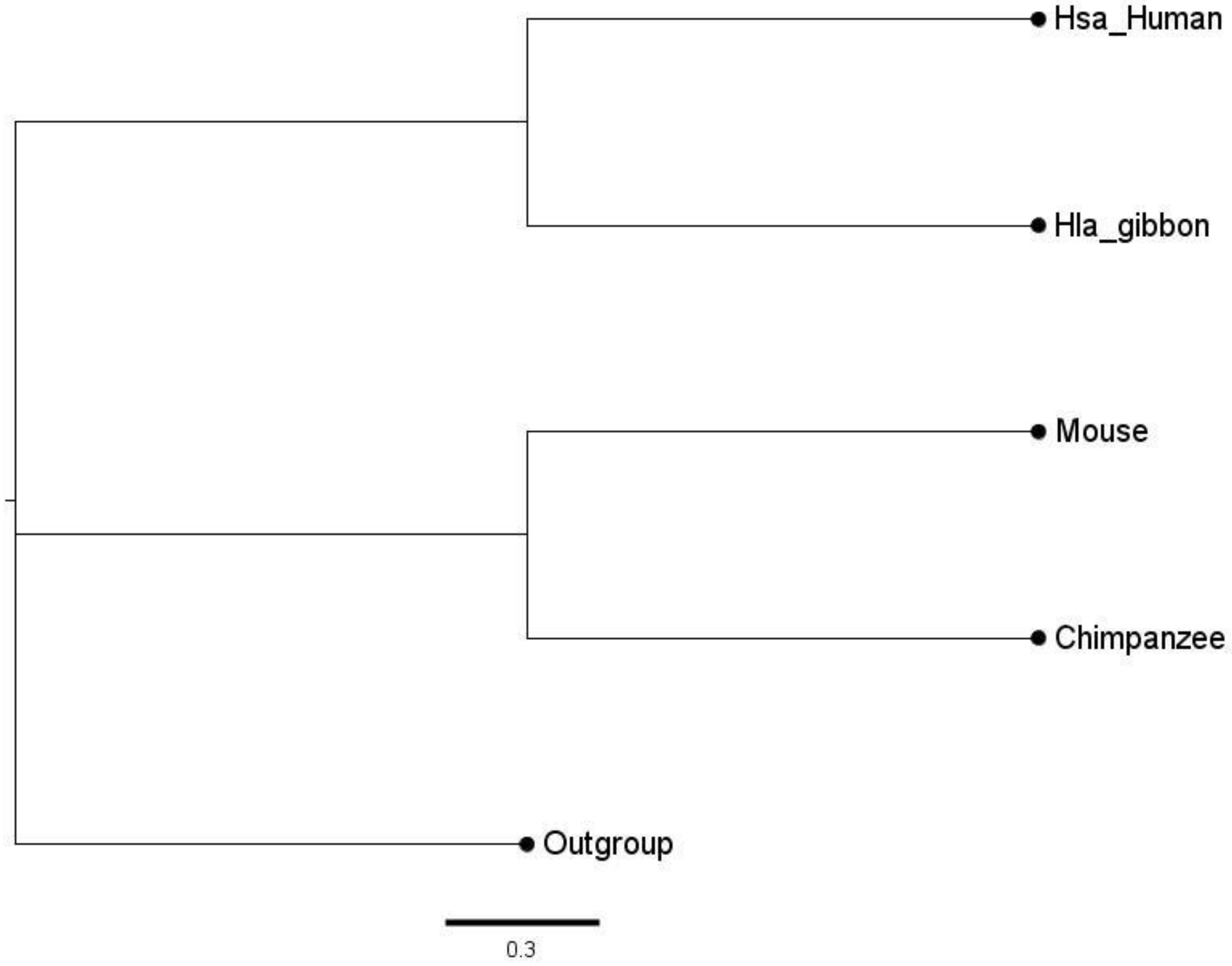
  Model=0 Nssites = 3

- **Likelihood Ratio Tests(LRTs)** :

  Ln(likelihood value) get from result file

  Like LRTS in comparison between Basic Model and Branch Model above.

Sites Model->Prove of existence of Sites under positive selection



AAG CCT CCT CCT CCT

AAG CAT CAG CCG CAG

AAG CAT CCT CCT CCT

AAG CCT CCG CCT CCT

AAG CAT CAG CAG CCT

Tips : But we can not find out which sites are under positive selection, only can prove they exist.

Sites Model->Prove of existence of Sites under positive selection

- **Suggest by Yang: two model comparison use for detection.**

A. Null hypothesis : Model=0 NSsites=1

Alternative hypothesis : Model=0 NSsites=2

B. Null hypothesis : Model=0 NSsites=7

Alternative hypothesis : Model=0 NSsites=8

- **Likelihood Ratio Tests(LRTs)** : Same with above procedure.

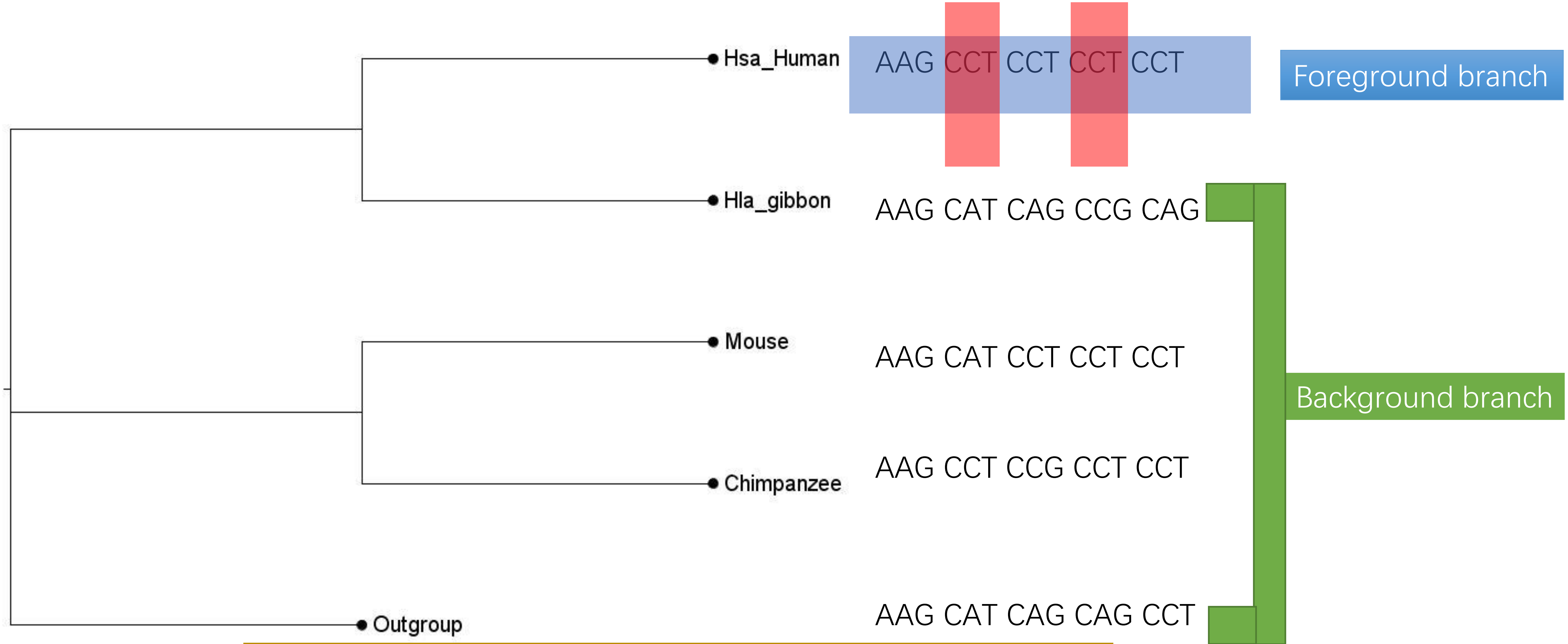
Tips : Suggesting by Yang, The M1-M2 comparison appears to be more robust (or less powerful) than the M7- M8 comparson.

Branch-Sites Model Type

Branch-Sites Model->Introduction

- **Assumption** : Allowing ω to vary among sites in protein and across branches on the tree.
- **Function** : Aiming to detect positive selection affecting a few sites only exist in **particular lineages/foreground branches**.
- **Key point** : Branch-Sites Model is combination of branch model and sites model, branch label is needed.

Branch-Sites Model->Demonstration



Hsa_Human	AAG CCT CCT CCT CCT
Hla_gibbon	AAG CAT CAG CCG CAG
Mouse	AAG CAT CCT CCT CCT
Chimpanzee	AAG CCT CCG CCT CCT
Outgroup	AAG CAT CAG CAG CCT

Overlapped area:
Sites/Codon under positive selection only in foreground branch

Branch-Sites Model->Model description

- Alternative hypothesis : Model A
  - Setting : Model=2 NSsites=2 fix_omega=0
- Null hypothesis Model:
  - Setting: Model=2 NSsites=2 fix_omga=1 omega=1
- Detail of Model A

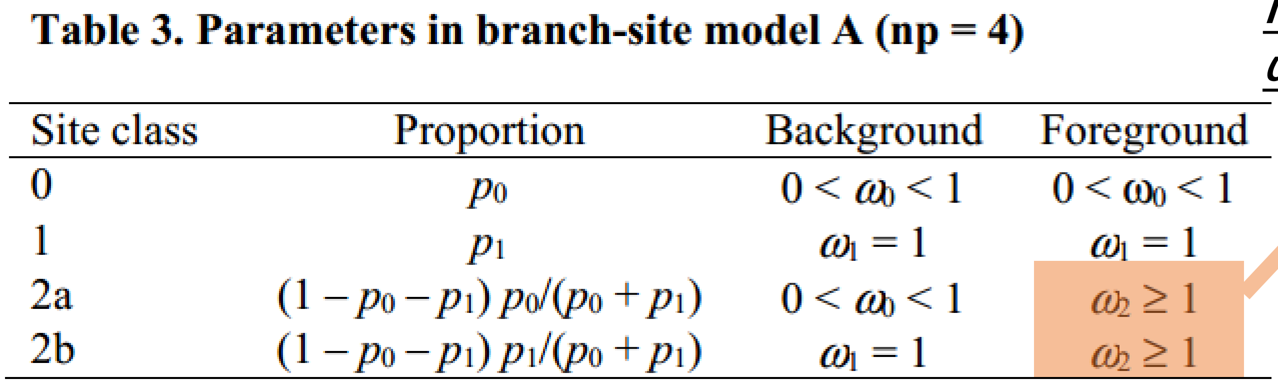


Table 3. Parameters in branch-site model A (np = 4)

Site class	Proportion	Background	Foreground
0	p_0	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$
2a	$(1 - p_0 - p_1)\ p_0/(p_0 + p_1)$	$0 < \omega_0 < 1$	$\omega_2 \geq 1$
2b	$(1 - p_0 - p_1)\ p_1/(p_0 + p_1)$	$\omega_1 = 1$	$\omega_2 \geq 1$

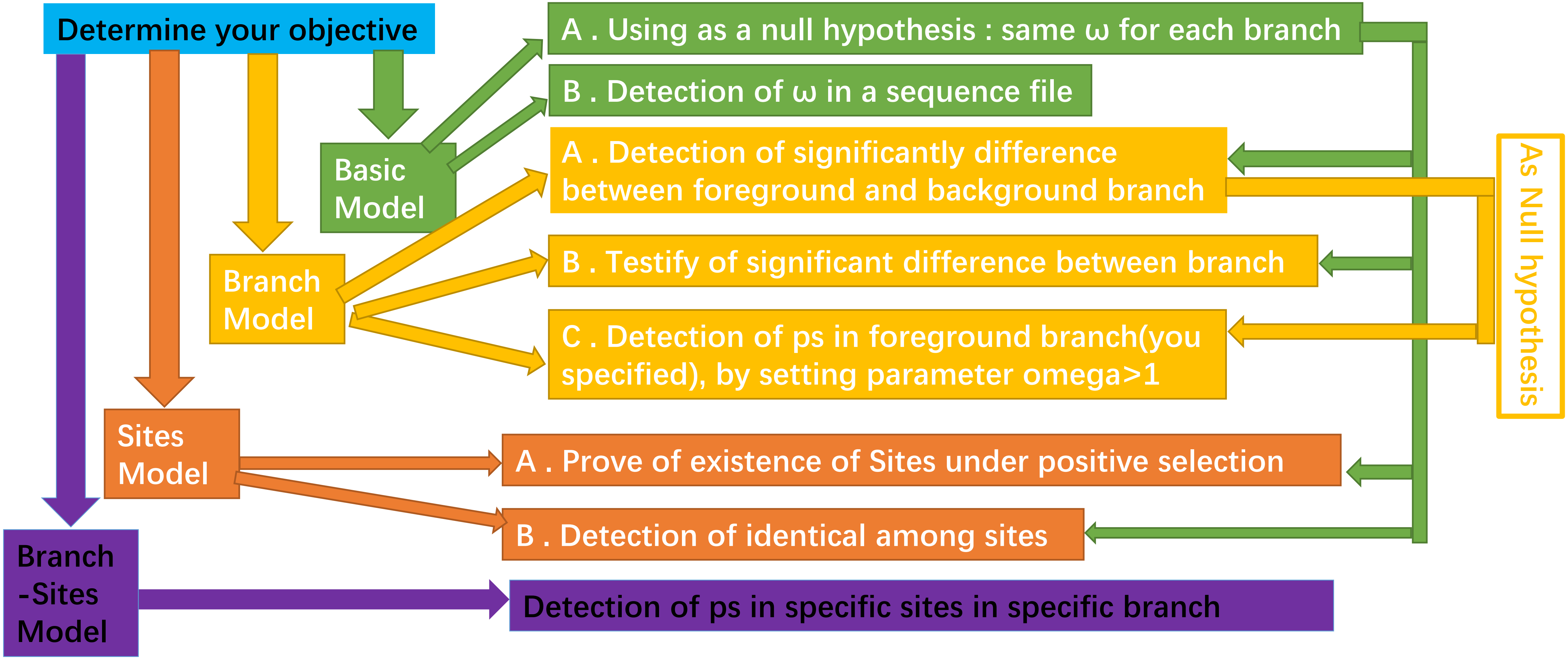
By setting omega=1 in Null hypothesis model. In fact assign ω_2 =1 compare with Model A ω_2>=1.

(Yang PAMI software description file)

Branch-Sites Model-> Detection

- **Application** : Detection for positive selection sites in specific lineage.
- **Null Hypothesis** :
  - Setting: Model=2 NSsites=2 fix_omga=1 omega=1
- **Alternative hypothesis** : use Model A
  - Setting : Model=2 NSsites=2 fix_omega=0
- Modified Likelihood Ratio Tests(Modified LRTs):
  - Assuming you calculate out **p-value**=0.22, the **real-p-value**=p-value/2=0.11, use **real-p-value** as test.

Functionality and Determine your objective



Determine your objective

Basic Model

A . Using as a null hypothesis : same ω for each branch

B . Detection of ω in a sequence file

Branch Model

A . Detection of significantly difference between foreground and background branch

B . Testify of significant difference between branch

C . Detection of ps in foreground branch(you specified), by setting parameter omega>1

Sites Model

A . Prove of existence of Sites under positive selection

B . Detection of identical among sites

Branch -Sites Model

Detection of ps in specific sites in specific branch

As Null hypothesis

Note : ps meaning Positive selection

Summary

- Functionality of codeml is to prove existence of positive selection.
- Know your purpose, choose suitable comparison of model
- Common hypothesis
  - Null hypothesis : Making sure no ω>1
  - Alternative hypothesis : Keeping a portion of omega distribution are larger than 1.
- Pay attention to other parameter when needed(such as “clock”)
- More details and comprehensive understanding need pay your attention to description file.

Thanks for your attention