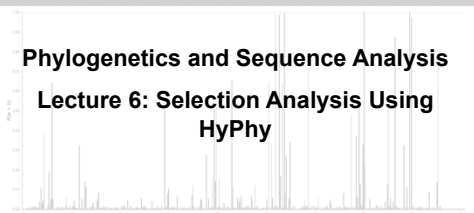National Institute of Allergy and Infectious Diseases

Phylogenetics and Sequence Analysis

Lecture 6: Selection Analysis Using HyPhy

Kurt Wollenberg, PhD
Phylogenetics Specialist
Bioinformatics and Computational Biology Branch (BCBB)
Fall 2015

NIAID

$$0 \longleftarrow 1 \longrightarrow \infty$$
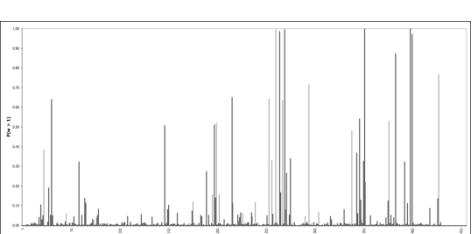
NIAID

Course Organization

- Building a clean sequence
- Collecting homologs
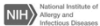- Aligning your sequences
- Building trees
- **Further Analysis**

NIAID

Lecture Organization

- What is selection?
- What does selection look like?
- HyPhy: using maximum likelihood to quantify selection
- HyPhy input
- Interpreting your output

NIAID

What is selection?

- Mutation – the source of all variation
- Substitution – changing one nucleotide to another
  - Synonymous substitution
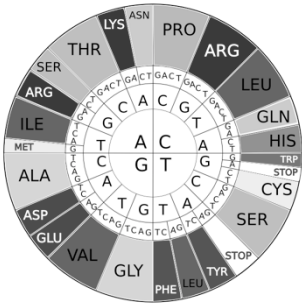 - Non-synonymous substitution
    - Conservative substitution

NIAID

Synonymous/non-synonymous Conservative/non-conservative

Adapted from Miller and Levine

Substitution

...GGC AGT TGG...   ...GGG TGT TGT...

Time

Translation   *Translation*

...G–S–W...   ...G–C–C...

conservative
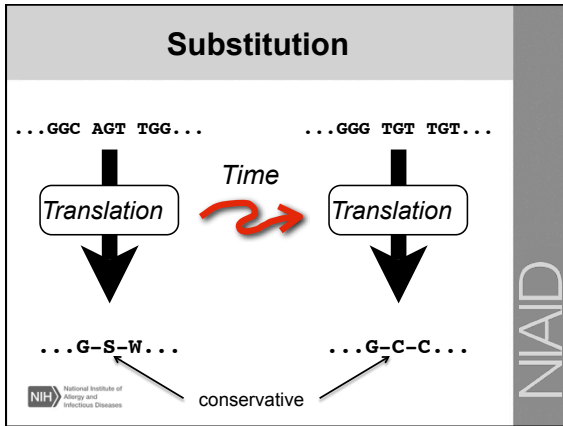
What is selection?

- Differential survival of genotypes
- Positive selection - diversifying
  - Maintain variation
  - Variation can also be maintained by genetic drift
- Negative selection – purifying
  - Remove variation

Why is selection important?

- Pathogens
  - ➢ Surviving the immune system
- Patients
  - ➢ Surviving infection
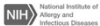- Identification of the molecular basis for survival

What is selection analysis?

- Determining the dN/dS ratio
  - ➢ Inference of ancestral states
  - ➢ Determination of the effect of inferred nucleotide changes on amino acids
- Is dN/dS significantly different from 1?

NIAID

NIH National Institute of Allergy and Infectious Diseases

What is selection analysis?

- Positive (Diversifying) Selection
  dN/dS > 1

- Negative (Purifying) Selection
  dN/dS < 1

NIAID

NIH National Institute of Allergy and Infectious Diseases

Two types of selection analysis

- Selection at sites
  - → Are there significant levels of non-synonymous substitution at specific codons?
  - → Averaged across lineages.

- Selection along lineages
  - → Are there significant levels of non-synonymous substitution in specific groups?
  - → Averaged across sequences.

NIAID

NIH National Institute of Allergy and Infectious Diseases

Two types of selection analysis

Selection at sites (codons)

```
             5 100 I
Selection  → MC1_01B4fs  TGCACTAATA ATCTGATT-- -------AAT ATCACTGAGA ATACTAATAA
along      → MC1_01A10   TGCACT---A ATCTGACAAA GGCTATTAAG ACCAATGGGA ATGCTAATAA
lineages   → MC1_01C1    TGCACTAATA ATCTGATT-- -------AAT ATCACTGAGA ATACTAATAA
           → MC1_01A20   TGCACTAATA ATCTGACAAA GGCTAGTAAT GCCACTGAGA AGGCTAATAA
           → MC1_01TA1   TGCACTAATA ATCTGATT-- -------AAT ATCACTGAGA ATACTAATAA

             TACCATTACT AAT------- --ACCACTAT TAATAGCGGG GGAGAAATAA
             TACCAGTACT AATGCCACTG AGAGTACCAT TACTAATACC ACTGAAATAA
             TACCATTACT AAT------- --ACCACTAT TAATAGCGGG GGAGAAATAA
             TACCATTACT AAT------- --ACCACTAT TGATAGCGGG GGAGAAATAA
             TACCATTACT AAT------- --ACCACTAT TGATAGCGGG GGAGAAATAA
```

MC1_01B4fs
MC1_01A10
MC1_01C1
MC1_01A20
MC1_01TA1

NIAID
NIH National Institute of Allergy and Infectious Diseases

Calculating dN/dS

- Estimate transition/transversion rate ratio
  → Transition: A↔G C↔T   A ↔ G purines
  → Transversion: purine↔purimidine C ↔ T pyrimidines
  → Based on four-fold degenerate sites at third codon positions and nondegenerate sites.
- Count synonymous and nonsynonymous substitutions for each codon site or along each lineage.
  → Use ML estimates of ancestral codon states.
- Correct counts for multiple hits.
- Fitting the data to a distribution of dN/dS ratios.
  → Calculate the probability of dN/dS falling into a particular class.

NIAID

HyPhy: the software

http://datamonkey.org

- Input data format
  → Sequences must be aligned.
  → Fasta, PHYLIP, or Nexus format.
  → Include the tree after the alignment.
- Program options
  → Analysis scripts.

NIAID
NIH National Institute of Allergy and Infectious Diseases

HyPhy: the software

Standard Analyses

Select a standard analysis to run

- Basic Analysis
- Codon Selection Analyses
- Compartmentalization
- Data File Tools
- Miscellaneous
- Model Comparison
- Kernel Analysis Tools
- Molecular Clock
- Phylogeny Reconstruction
- **Positive Selection**
- Recombination
- Selection/Recombination
- Relative Rate
- Relative Ratio
- Substitution Rates

Item Description

Perform codon-based positive selection tests.

OK Cancel

NIAID

NIH National Institute of Allergy and Infectious Diseases

HyPhy: the software

Positive Selection Analyses

Select a standard analysis to run

- Phylogeny Reconstruction
- **Positive Selection**
  - BranchSiteREL.bf
  - YangNielsenBranchSite2005.bf
  - TestBranchDNDS.bf
  - BUSTED.bf
  - CompareSelectivePressure.bf
  - CompareSelectivePressureIVL.bf
  - NielsenYang.bf
  - MFPositiveSelection.bf
  - QuickSelectionDetection.bf
  - SubtreeSelectionComparison.bf
  - DirectionalREL.bf
  - MEDS.bf
- Recombination

Item Description

Perform codon-based positive selection tests.

OK Cancel

NIAID

NIH National Institute of Allergy and Infectious Diseases

HyPhy: the software

Quick Selection Detection

- Single Ancestor Counting
- Weighted Ancestor Counting
- Sample Ancestral States
- Process Sampled Ancestral States
- One rate FEL
- Two rate FEL
- Rate Distribution
- Full site-by-site LRT
- Multirate FEL
- BGM co-evolution
- MEME

Item Description

Fixed effects site-by-site likelihood estimation. dS adjusted across sites.

1 required (1 chosen).   OK Cancel

NIAID

NIH National Institute of Allergy and Infectious Diseases

HyPhy: the software

Quick Selection Detection

- Single Ancestor Counting
- Weighted Ancestor Counting
- Sample Ancestral States
- Process Sampled Ancestral States
- One rate FEL
- Two rate FEL
- Rate Distribution
- Full site-by-site LRT
- Multirate FEL
- BGM co-evolution
- MEME

Item Description

Mixed effects model of evolution to search for evidence of episodic selection at individual sites [EXPERIMENTAL, March 2011]

1 required (1 chosen).   OK   Cancel

HyPhy: the software

The Algorithms

- **REL** – relative effects model – very similar to PAML codeml F61 analysis
- **2pFEL** –2-parameter fixed-effects model
- **MEME** – mixed-effects model

HyPhy: the software

The MEME Algorithm

Mixed Effects Model of Evolution

- Analyzes levels of positive selection at individual sites while allowing level of positive selection to vary from branch to branch.
- Other models (FEL, REL, etc.) assume level of positive selection to be uniform across all branches.
- Includes site-to-site variation in synonymous substitution rate, just like FEL.

HyPhy: Input file formats

Fasta

No stop codons!

Make sure these match!

```
>MC1_01B4fs
TGCACTAATAATCTGATT---------AATATCACTGAGAATACTAATAATACCATTACT
>MC1_01A10
TGCACT---AATCTGACAAAGGCTATTAAGACCAATGGGAATGCTAATAATACCAGTACT
>MC1_01C1
TGCACTAATAATCTGATT---------AATATCACTGAGAATACTAATAATACCATTACT
>MC1_01A20
TGCACTAATAATCTGACAAAGGCTAGTAATGCCACTGAGAAGGCTAATAATACCATTACT
>MC1_01TA1
TGCACTAATAATCTGATT---------AATATCACTGAGAATACTAATAATACCATTACT

(((MC1_01B4fs,MC1_01A10),MC1_01C1),(MC1_01A20,MC1_01TA1));
```

NIAID

NIH National Institute of Allergy and Infectious Diseases

HyPhy: Input file formats

Fasta - interleaved

Make sure these match!

No stop codons!

```
>MC1_01B4fs
>MC1_01A10
>MC1_01C1
>MC1_01A20
>MC1_01TA1

TGCACTAATAATCTGATT---------AATATCACTGAGAATACTAATAATACCATTACT
TGCACT---AATCTGACAAAGGCTATTAAGACCAATGGGAATGCTAATAATACCAGTACT
TGCACTAATAATCTGATT---------AATATCACTGAGAATACTAATAATACCATTACT
TGCACTAATAATCTGACAAAGGCTAGTAATGCCACTGAGAAGGCTAATAATACCATTACT
TGCACTAATAATCTGATT---------AATATCACTGAGAATACTAATAATACCATTACT

AAT---------ACCACTATTAATAGCGGGGAGAAATAA
AATGCCACTGAGAGTACCATTACTAATACCACTGAAATAA
AAT---------ACCACTATTAATAGCGGGGAGAAATA
AAT---------ACCACTATTGATAGCGGGGAGAAAT
AAT---------ACCACTATTGATAGCGGGGAGAAATAA

(((MC1_01B4fs,MC1_01A10),MC1_01C1),(MC1_01A20,MC1_01TA1));
```

NIAID

NIH National Institute of Allergy and Infectious Diseases

HyPhy: Input file formats

PHLYIP

No stop codons!

Make sure these match!

```
  5 100 I
MC1_01B4fs  TGCACTAATA ATCTGATT-- -------AAT ATCACTGAGA ATACTAATAA
MC1_01A10   TGCACT---A ATCTGACAAA GGCTATTAAG ACCAATGGGA ATGCTAATAA
MC1_01C1    TGCACTAATA ATCTGATT-- -------AAT ATCACTGAGA ATACTAATAA
MC1_01A20   TGCACTAATA ATCTGACAAA GGCTAGTAAT GCCACTGAGA AGGCTAATAA
MC1_01TA1   TGCACTAATA ATCTGATT-- -------AAT ATCACTGAGA ATACTAATAA

            TACCATTACT AAT------- --ACCACTAT TAATAGCGGG GGAGAAATAA
            TACCAGTACT AATGCCACTG AGAGTACCAT TACTAATACC ACTGAAATAA
            TACCATTACT AAT------- --ACCACTAT TAATAGCGGG GGAGAAATAA
            TACCATTACT AAT------- --ACCACTAT TGATAGCGGG GGAGAAATAA
            TACCATTACT AAT------- --ACCACTAT TGATAGCGGG GGAGAAATAA

1
(((MC1_01B4fs,MC1_01A10),MC1_01C1),(MC1_01A20,MC1_01TA1));
```

NIAID

NIH National Institute of Allergy and Infectious Diseases

## HyPhy: Input file formats

Nexus

```
#NEXUS
Begin data;
        Dimensions ntax=5 nchar=100;
        Format datatype=DNA gap=- interleave;          No stop codons!
MATRIX
MC1_01B4fs    TGCACTAATA ATCTGATT-- -------AAT ATCACTGAGA ATACTAATAA
MC1_01A10     TGCACT---A ATCTGACAAA GGCTATTAAG ACCAATGGGA ATGCTAATAA
MC1_01C1      TGCACTAATA ATCTGATT-- -------AAT ATCACTGAGA ATACTAATAA
MC1_01A20     TGCACTAATA ATCTGACAAA GGCTAGTAAT GCCACTGAGA AGGCTAATAA
MC1_01TA1     TGCACTAATA ATCTGATT-- -------AAT ATCACTGAGA ATACTAATAA

              TACCATTACT AAT------- --ACCACTAT TAATAGCGGG GGAGAAATAA
              TACCAGTACT AATGCCACTG AGAGTACCAT TACTAATACC ACTGAAATAA
              TACCATTACT AAT------- --ACCACTAT TAATAGCGGG GGAGAAATAA
              TACCATTACT AAT------- --ACCACTAT TGATAGCGGG GGAGAAATAA
              TACCATTACT AAT------- --ACCACTAT TGATAGCGGG GGAGAAATAA
;
End;
Begin trees;
        TREE tree=(((MC1_01B4fs,MC1_01A10),MC1_01C1),(MC1_01A20,MC1_01TA1));
End;
```

Make sure these match!

HyPhy: Input file formats

GAPS!

Sequence data

```
  5 100 I
MC1_01B4fs    TGCACTAATA ATCTGATT-- -------AAT ATCACTGAGA ATACTAATAA
MC1_01A10     TGCACT---A ATCTGACAAA GGCTATTAAG ACCAATGGGA ATGCTAATAA
MC1_01C1      TGCACTAATA ATCTGATT-- -------AAT ATCACTGAGA ATACTAATAA
MC1_01A20     TGCACTAATA ATCTGACAAA GGCTAGTAAT GCCACTGAGA AGGCTAATAA
MC1_01TA1     TGCACTAATA ATCTGATT-- -------AAT ATCACTGAGA ATACTAATAA

              TACCATTACT AAT------- --ACCACTAT TAATAGCGGG GGAGAAATAA
              TACCAGTACT AATGCCACTG AGAGTACCAT TACTAATACC ACTGAAATAA
              TACCATTACT AAT------- --ACCACTAT TAATAGCGGG GGAGAAATAA
              TACCATTACT AAT------- --ACCACTAT TGATAGCGGG GGAGAAATAA
              TACCATTACT AAT------- --ACCACTAT TGATAGCGGG GGAGAAATAA
```

- Gaps must be in multiples of three <u>and</u> preserve the reading frame!
- ML models cannot analyze gapped codon sites – they will be ignored for the analysis.

HyPhy: data input

Codon frequency models

0: 1/61 – All codons equally probable

1: F1x4 - Codon frequencies based on overall nucleotide frequencies (f_A, f_C, f_G, f_T)

2: F3x4 - Codon frequencies based on nucleotide frequencies at each codon site (f_{A1}, f_{C1}, f_{G1}, f_{T1}, f_{A2}, f_{C2}, f_{G2}, f_{T2}, f_{A3}, f_{C3}, f_{G3}, f_{T3})

3: F61 - Codon frequencies based on frequency of each codon in the data (f_{aaa}, f_{aac}, f_{aag}, f_{aat}, f_{aca}, f_{acc}, f_{acg}, f_{act}, f_{aga}, f_{agc}, f_{agg}, f_{agt}, ...)

F3x4 is used in the MEME analysis

NIAID

HyPhy: data output

NIAID

HyPhy: data output

NIAID

HyPhy: data output

Codon	alpha	beta1	p1	beta2	p2	LRT	p-value	q-value	Log(L)
125	0.0000	0.0000	0.7253	13.6010	0.2747	8.0176	0.0081	0.3258	-30.6957
153	0.9139	0.0000	0.8927	49.5335	0.1073	8.7797	0.0055	0.2580	-22.6063
163	0.6646	0.0000	0.8478	20.0270	0.1522	5.4309	0.0304	0.8559	-25.8135
188	0.0000	0.0000	0.9673	1883.3800	0.0327	9.7907	0.0033	0.1853	-12.7060
211	0.7245	0.0000	0.9660	4586.8100	0.0340	11.9000	0.0011	0.1062	-22.5928
218	0.8903	0.0000	0.9662	304.8150	0.0338	13.3192	0.0006	0.1555	-17.2640
234	0.0000	0.0000	0.8160	263.5910	0.1840	10.1891	0.0027	0.1893	-33.7409
237	0.0969	0.0969	0.7191	17.6831	0.2809	12.3769	0.0009	0.1252	-35.2664
239	0.6719	0.0000	0.8254	11.6736	0.1746	5.9541	0.0232	0.7269	-22.8152
264	0.0000	0.0000	0.9370	8.8091	0.0630	6.6746	0.0160	0.5655	-11.4430

Column Headings

Codon	The index of the codon in the alignment
α	The maximum likelihood estimate (MLE) of the synonymous substitution rate α
β^-	The maximum likelihood estimate (MLE) of the non-synonymous rate for the branch class with $\beta \le \alpha$
$Pr[\omega=\alpha]$	The MLE of the proportion of sites evolving at β^-.
β^+	The MLE of the unconstrained β non-synonymous rate
$Pr[\beta=\beta^+]$	The MLE of the proportion of sites evolving at β^+.
LRT	Likelihood ratio test statistic for $\beta^- = \alpha$ (null) versus β^+ unrestricted (alternative)
p-value	The p-value for the LRT test, using the mixture distribution : $0.33\ \chi^2_0 + 0.30\ \chi^2_1 + 0.37\ \chi^2_2$
q-value	The q-value for independent tests (upper bound on the false discovery rate), derived from the corresponding p-value using Simes' procedure

---

HyPhy: further analyses

Selection Analysis

Site vs. Lineage Selection

Generally, if significant positive (or negative) selection occurs at only a few sites, averaging over the length of the sequences will result in no significant positive (or negative) selection being detected.
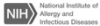
Selection Analysis

Significance of Results

Selection analysis only determines if there is a significant **excess** or **lack** of non-synonymous substitution. The relative level of physiochemical similarity among the two amino acids is beyond the capabilities of selection analysis software.

NIAID

NIH National Institute of Allergy and Infectious Diseases

References

❖ Fundamentals of Molecular Evolution. 2nd edition. Grauer and Li. 2000
   A good introduction to the major concepts in molecular evolution.

❖ The Phylogenetics Handbook. P. Lemey, editor. 2005.
   Chapter 14: Theory of and practice of diversifying selection analysis.

❖ MEME reference:
   Murrell B, et al. (2012) Detecting individual sites subject to episodic diversifying selection. *PLoS Genetics*, 8(7):e1002764.

NIAID

NIH National Institute of Allergy and Infectious Diseases
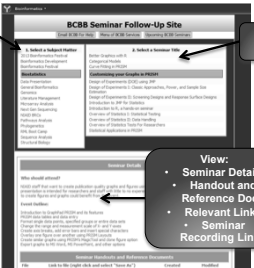
Seminar Follow-Up Site

- For access to past recordings, handouts, slides visit this site from the NIH network: http://collab.niaid.nih.gov/sites/research/SIG/Bioinformatics/

Recommended Browsers:
- IE for Windows,
- Safari for Mac (Firefox on a Mac is incompatible with NIH Authentication technology)

Login
- If prompted to log in use "NIH\" in front of your username

NIH National Institute of Allergy and Infectious Diseases

NIAID

Retrieving Slides/Handouts

Retrieving Slides/Handouts

Questions?

ScienceApps@niaid.nih.gov

Next

Thursday, 10 December at 1300

Molecular Evolutionary Analysis of Pathogens Using BEAST