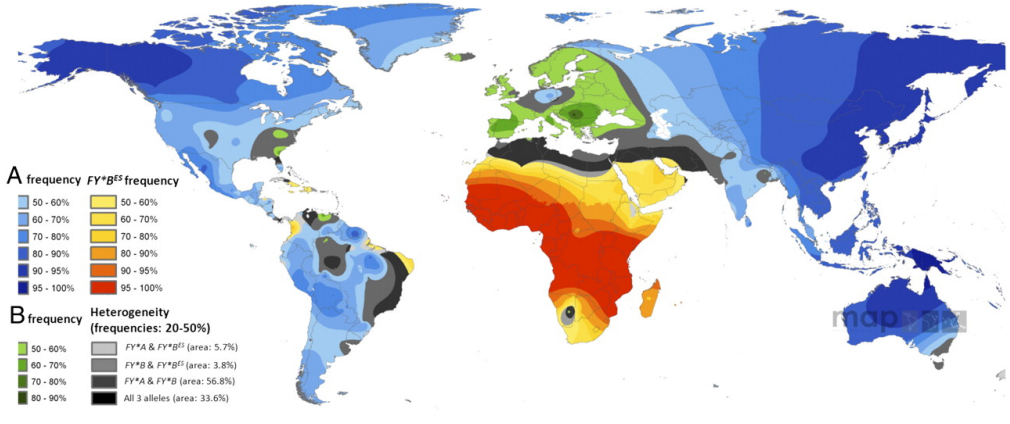
MIMM4750G

Detecting selection

Natural selection

- Variation in fitness that is associated with variation in the environment.
- Selection is responsible for the spread of drug resistance in pathogens.
- Host adaptation: a mutant allele of the Duffy blood-group antigen (Fy) that reduces risk of infection by *Plasmodium vivax* (vivax malaria) is near fixation in sub-Saharan Africa.



Types of selection

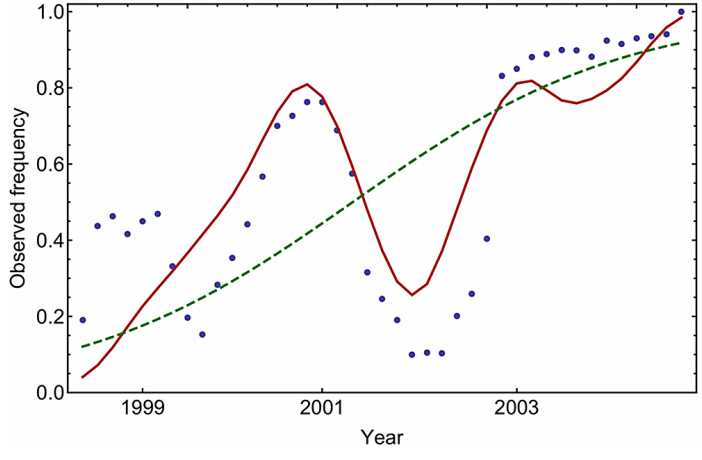
- **Purifying selection:** once the favored genetic variant has been "fixed" in the population, selection continues to remove all other variants.
- Most sites are under purifying selection ("if it ain't broke..")
- **Directional selection:** a specific genetic variant has a selective advantage and increases in frequency.
- Purifying selection is the end result of directional selection.
- Directional selection is difficult to observe; you must be at the right place at the right time.

Diversifying selection

- **Diversifying selection:** different genetic variants are favored in different environments.
- Directional selection depletes genetic variation; diversifying selection promotes variation.
- What causes diversifying selection in pathogens?
- Transmission of an infection from one host environment to another (host-specific immune responses).
- *e.g.*, strong div. selection at cytotoxic T-lymphocyte epitopes in virus proteins.

How do detect selection?

- Longitudinal data: track the frequencies of an allele over time.



Trajectory of `I161N` mutation in influenza A virus HA. Figure from Illingworth and Mustonen (2012) PLOS Pathog 8: e1003091.

- What if we don't know which allele is under selection? What if we don't have longitudinal data?
- Cross-sectional (comparative) methods: infer selection by comparing genetic sequences sampled at about the same time.

Protein evolution

- Infer selection by comparting relative rates of evolution.
- Requires a baseline/point of reference, *e.g.*, "neutrally evolving" pseudogenes.
- A popular contrast is non-synonymous versus synonymous variation within protein-coding sequence.
- Non-synonymous = nucleotide substitution alters the amino acid encoded by the codon.
- Such approaches are generally called "dN/dS" methods.

dN/dS

- There are 9 possible nucleotide substitutions in a 3-nt codon.
- We assume that "nonsense" substitutions to a stop codon don't persist.
- The genetic code determines how many of these 9 nt changes would result in a non-synonymous change — this is the number of NS sites.
- dN is the ratio between the numbers of observed NS substitutions and of NS sites.
- dS is the same ratio for synonymous substitutions and sites.

A simple example

- The codon `ACG` encodes threonine (T). It has 3 synonymous sites and 6 nonsynymous sites (*e.g.*, `ATG` for methionine).
- Suppose we count 8 non-synonymous and 4 synonymous substitutions in the tree.
- We observed over twice as many non-synonymous substitutions! Is this evidence of strong diversifying selection at this codon?
- The ratio dN/dS is $\frac{8}{6} / \frac{4}{3} = 1$. This looks like neutral evolution.

Table 1
The Gene Groups on Which Positive Selection May Operate

Gene Group	Representative Species
Merozoite surface antigen (*MSA2*) gene	Malaria *Plasmodium falciparum*
Major surface protein (*msp1* α) gene	Rickettsia *Anaplasma marginale*
Outer membrane protein (*omp*) gene	*Chlamydia*
env	Equine infectious anemia virus
Glycoprotein *gH* gene	Pseudorabies virus
E gene	Phages *G4*, *ϕX174* and *S13*
Sigma-1 protein gene	Reovirus
Invasion plasmid antigen gene (*ipaC*)	*Shigella*
Invasion plasmid antigen gene (*ipaD*)	*Shigella*
Egg-laying hormone	*Aplysia californica*
Egg-laying hormone A peptide	*Aplysia californica*
ATP synthase F_0 subunit (*atp*-2) gene	*Escherichia coli*
Neomycin resistance protein gene	*Escherichia coli*
Virulence determinant gene (*yadA*)	*Yersinia*
Prostatic steroid binding protein	Rat
Neurotoxin	Snake
CDC6	*Saccharomyces cerevisiae*

Table from Endo, Ikeo and Gojobori. 1996, Mol Biol Evol 13: 685.

Using likelihood

- If we can model codon evolution like we did for nucleotides, then we can estimate dN/dS by maximum likelihood.
- There are a lot of parameters! (61 non-stop codons, $61 \times 60/2 = 1830$ rates.)
- Assumptions!:

  1. There is never more than one nucleotide substitution within a codon at a time (*e.g.*, no simultaneous mutations).
  2. The codon context has no effect on nucleotide substitution rates.
  3. The dN/dS ratio does not care what amino acid is encoded by the codon, only whether the amino acid is *changed*.

The Goldman-Yang / Muse-Gaut models

- In 1994, very similar models were proposed in two ground-breaking papers (in the same journal, next to each other *in the same issue*).
- Using these assumptions, the GY and MG models enable us to specify a codon model using as few as *two* parameters.
- The main parameter is called ω or R, depending who you ask. It is simply the ratio of non-syn. and syn. rates.
- The minimal second parameter is the synonymous rate (Jukes-Cantor model).

Maximum likelihood (ML)

- We can use ML to reconstruct the tree best supported by the data.
- We can also use ML to fit one of these codon models to the tree.
- It is *possible* to simultaneously estimate both the tree and the codon model, but
- It is simpler to "fix" our analysis to a single tree when fitting the model.
- Methods to simultaneously fit the tree and codon model have only recently been developed, *e.g.*, CodonPhyML.

Site-specific selection

- In the same year, Ziheng Yang described methods to allow dN/dS to change at different codon sites of a protein-coding gene.
- This was a critical improvement because it is often not the *entire gene* that is under diversifying selection.
- This allows us to pick out individual amino acids under strong selection, even if the rest of the gene is highly conserved.
- Implemented by Yang in 1997 into the software package PAML (Phylogenetic Analysis by Maximum Likelihood).

dN/dS in influenza A virus

- Influenza A virus (IAV) hemagglutinin (HA) is responsible for binding host cell receptor.
- HA binds sialic acids in upper respiratory tract - name stems from clumping of blood cells.
- Major target for antibody-mediated immune response.
- Specific amino acids in HA protein under strong selection.

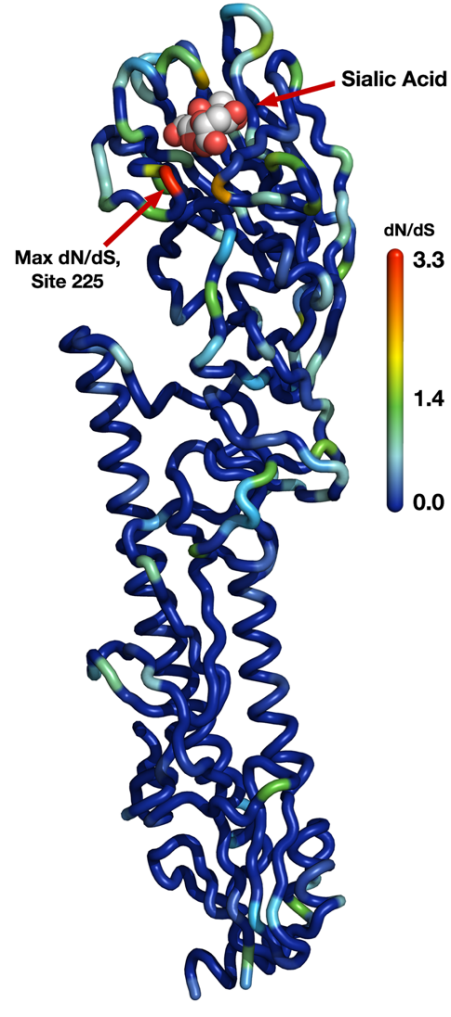


Figure from Meyer and Wilke (2015) PLOS Pathog 11: e1004940.

Random-effects likelihood

- Yang allows dN/dS (ω) to vary among sites by assuming that these values followed a gamma distribution.
- The gamma distribution is a *continuous* probability density function for values greater than zero, ideal for rates.
- To make easier to compute, Yang split the gamma distribution up into 5 rate categories of equal area (probability).
- This approach is still used for many models, and gamma is represented by a G or the symbol Γ.

Fixed-effects likelihood

- Also proposed by Yang, a fixed-effect model attempts to assign each codon site to one of multiple categories, each with its own estimated dN/dS rate.
- Random-effects models tend to have greater power (fewer parameters) but less flexible.
- Fixed-effects models tend to model rate variation more accurately (more flexible) but may require more data than REL.
- *These methods basically give you the same results if you have enough data.*

Episodic selection

- The previous models assume that differences in dN/dS are constant over time.
- What if selection at a specific site is driven by a change in the environment?
- Yang (again) and Nielsen (2002) proposed the branch-site method that assigns branches of the tree into two categories.
- Difficult to work more complex models because of over-fitting!

Ziheng Yang.

Detecting directional selection

- Directional selection is transient, making it difficult to "see".
- When selection brings a specific variant from low to high frequency in the population, there is a local depletion of genetic variation.
- This is called a "selective sweep".

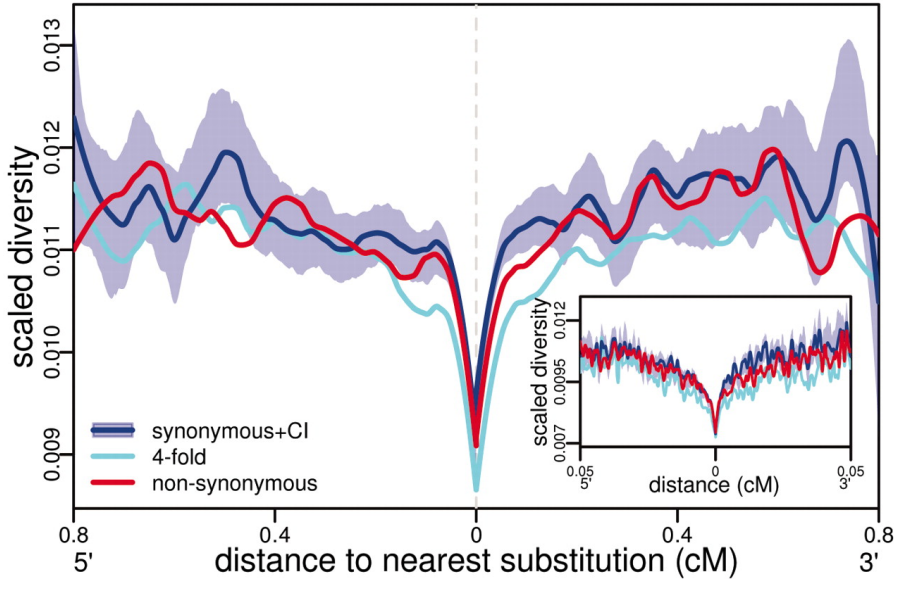


Figure from RD Hernandez *et al.* 2011, Science 331: 920.

Selective sweeps

- Presently a very active area of research.
- Methods look for parts of the genome with reduced variation, patterns in linkage disequilibrium or the allele frequency distribution.