



The PG-BSM Concept

Summer 2020

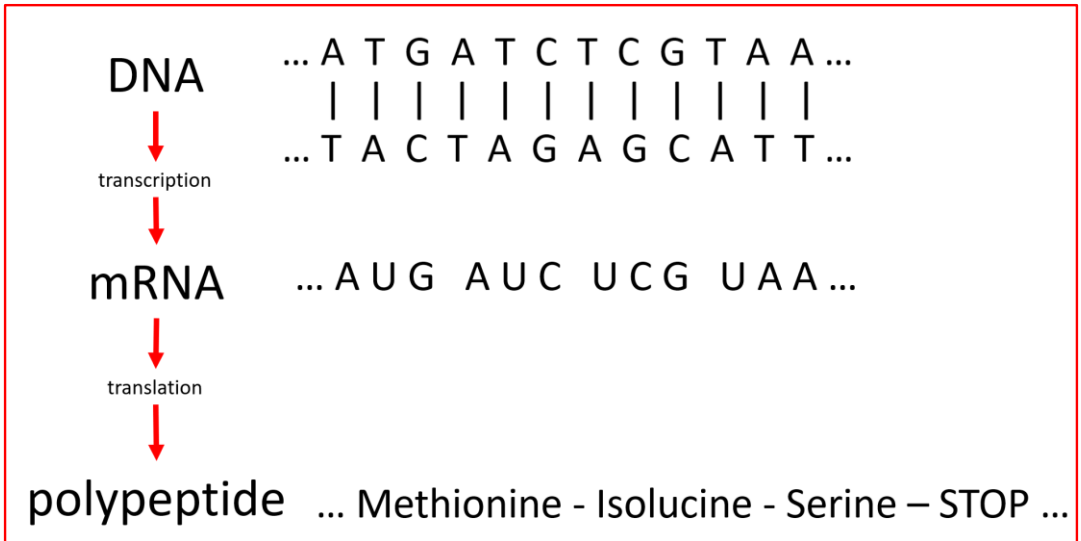
The purpose of this document is to explain the ideas behind the Phenotype-Genotype Branch-Site Model. **Introduction** provides a brief explanation of the purpose of codon substitution models in general. **The Problem** describes issues with standard codon substitution models and includes our way of thinking about a site-specific fitness landscape. **The PG-BSM** illustrates how the new model mitigates the issues with the standard approach. **An Example** provides an illustration of the PG-BSM applied to real data.



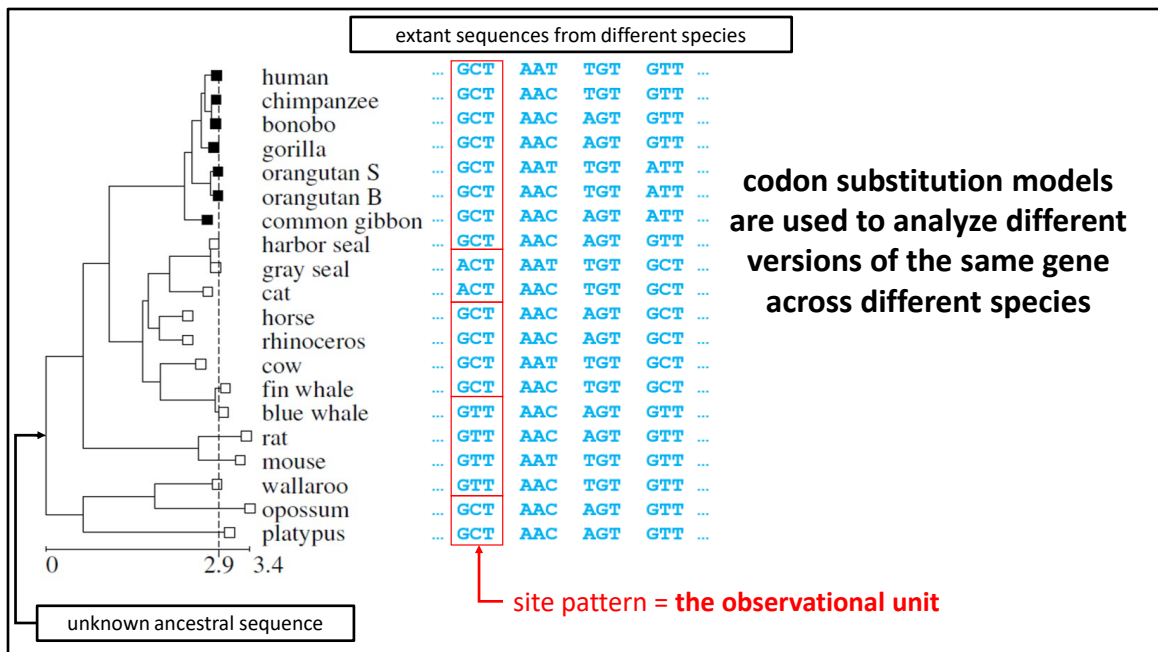
Introduction

Introduction provides a brief explanation of the purpose of codon substitution models.

Processes of change in protein-coding DNA



A polypeptide (i.e., a protein molecule) is encoded in DNA as a double-stranded sequence of nucleotides (T, G, C, and A). Protein machines in a cell **transcribe** DNA into a corresponding sequence of nucleotide triplets or codons (mRNA). Other machines then assemble the encoded protein by **translating** each codon to its amino acid. Mutations in DNA (when one nucleotide is replaced by another) can alter the amino acid sequence and so can change the properties of the translated protein. Any such mutation has two possible fates: it can be eliminated, especially if it substantially reduces the fitness of the organism (eliminated by negative selection); or it can be fixed, especially if it improves the fitness of the organism (fixed by positive selection). The fixation of a single nucleotide mutation is referred to as a codon substitution. Changes in the codon sequence of a given protein can be inferred by comparing homologues, sequences that all code for the same protein each taken from different species. Codon substitution models are used to infer from an alignment of homologous sequences those sites in the protein where codon substitutions can be attributed to positive selection.

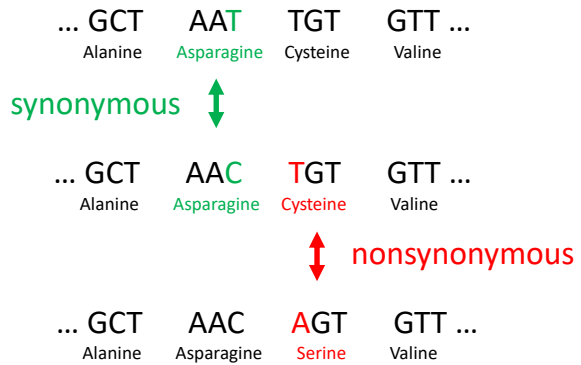


Aligned homologous protein-coding sequences are assumed to have evolved from a common unknown ancestral sequence. Each row in an alignment corresponds to a sequence taken from one species. Each column corresponds to a homologous codon site. The pattern of amino acids at a site (aka the site pattern) is the observational unit for any statistical analysis. Site patterns contain information about the processes of mutation, selection, and drift that gave rise to the alignment.

The Standard Genetic Code

Amino Acid	Codon Aliases
Alanine	GCT, GCC, GCA, GCG
Arginine	CGT, CGC, CGA, CGG, AGA, AGG
Asparagine	AAT, AAC
Aspartic acid	GAT, GAC
Cysteine	TGT, TGC
Glutamine	CAA, CAG
Glutamic acid	GAA, GAG
Glycine	GGT, GGC, GGA, GGG
Histidine	CAT, CAC
Isoleucine	ATT, ATC, ATA
Methionine	ATG (Start)
Leucine	TTA, TTG, CTT, CTC, CTA, CTG
Lysine	AAA, AAG
Phenylalanine	TTT, TTC
Proline	CCT, CCC, CCA, CCG
Serine	TCT, TCC, TCA, TCG, AGT, AGC
Threonine	ACT, ACC, ACA, ACG
Tryptophan	TGG
Tyrosine	TAT, TAC
Valine	GTT, GTC, GTA, GTG
Stop	TAA, TAG, TGA

Two kinds of mutations:



The genetic code is degenerate, meaning that most amino acids are encoded by more than one codon triplet. It follows that two kinds of mutations are possible.

Synonymous mutations are those that do not change the amino acid sequence.

These are selectively neutral because they have no impact on the translated protein.

Nonsynonymous mutations are those that do change the amino acid. These can alter the fitness of the organism by changing the protein for better or worse. Synonymous mutations, being selectively neutral, are fixed at the same rate at which they arise.

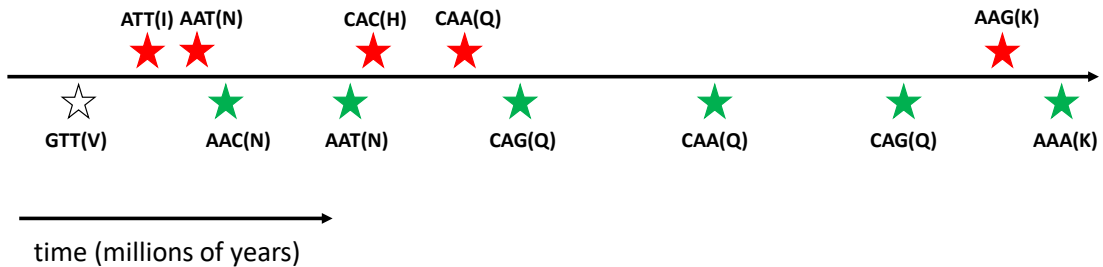
The inferred rate of synonymous substitutions therefore serves as a sort of molecular clock provided synonymous mutations can be assumed to arise at a constant rate.

Nonsynonymous mutations, by contrast, are fixed at a variable rate, since most are deleterious and so are not fixed and remain unobserved while some might occasionally improve the fitness of the organism and be fixed.

Molecular Evolution as a Markov Chain

★ synonymous substitutions, rate = dS

★ nonsynonymous substitutions, rate = dN

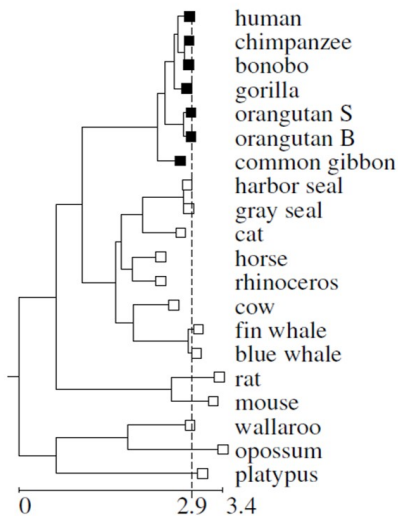


heterotachy: variation in the nonsynonymous substitution rate dN over time

Metaphorically speaking, changes at a codon site can be thought of as a random walk through the space of all possible codons via a chain of substitution events. Suppose a given site is occupied by valine (V, GTT, white star). In the above diagram the first step or substitution was to isoleucine (I, ATT). This changed the amino acid from V to I and so was nonsynonymous (indicated by a red star). There followed a nonsynonymous substitution to asparagine (N, AAT). The next substitution then changed the codon sequence from AAT to AAC but did not change the amino acid, which remained asparagine. Hence that substitution was synonymous (indicated by a green star).

Notice that the synonymous substitution rate (dS , inversely proportional to the spaces between adjacent green stars) is approximately constant over time. By contrast, the nonsynonymous substitution rate (dN , inversely proportional the spaces between adjacent red stars) varies over time, sometimes with $dN > dS$ but most of the time with $dN < dS$. Variation in dN at a codon site over time is commonly referred to as **heterotachy**.

Codon Substitution Models



The key measure:

$$\omega = dN/dS$$

A proxy for selection pressure.

$\omega < 1 \rightarrow$ stringent

$\omega = 1 \rightarrow$ neutral

$\omega > 1 \rightarrow$ positive

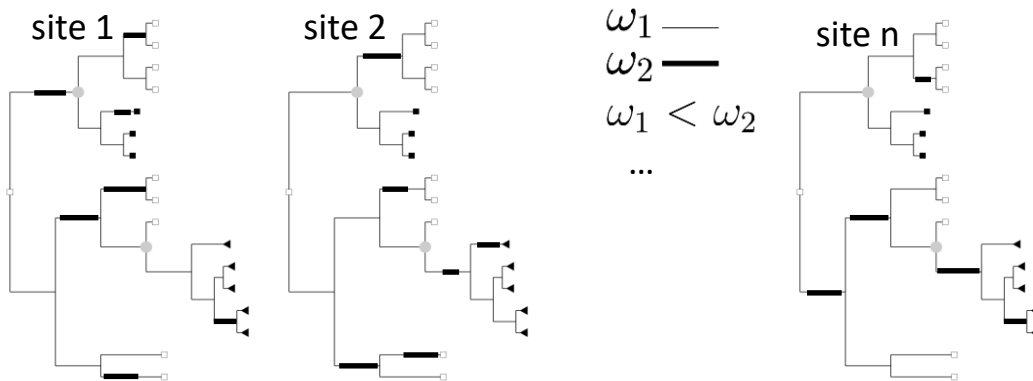
A codon substitution model is fitted to an alignment to estimate the dN/dS ratio. This ratio, commonly represented by the symbol ω , is used to infer the nature of the selective regime that acted on the protein sequence or at a site within the sequence. A stringent change-resistant regime (aka **negative selection**) is inferred when $dN < dS$ or when $\omega < 1$. A **selectively neutral** regime is inferred when $dN = dS$ or $\omega = 1$. And a change-promoting regime (aka **positive selection**) is inferred when $dN > dS$ or $\omega > 1$. Under the standard paradigm, evidence of **positive selection** is equated to evidence of **adaptive evolution**. The two are not equivalent, however, as will be demonstrated in **The Problem**.



The Problem

The Problem describes issues with standard codon substitution models and includes our way of thinking about a site-specific fitness landscape.

Heterotachy is the Key Signature (not $\omega > 1$)



modeled as phenomenological switching between $\omega_1 < \omega_2$ (**covarion-like model**)

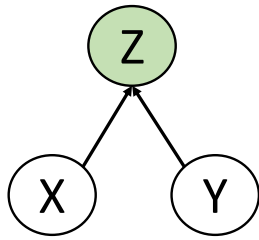
Fitch and Markowitz 1970, Fitch 1971, Galtier 2001, Guindon et al.2004

Sophisticated codon substitution models (e.g., covarion-like models) have been formulated to estimate $\omega = dN/dS$ at any one site over any one branch of the phylogenetic tree. These are meant to detect subtle signatures in an alignment that suggest site-specific variations in ω commonly or heterotachy. Following tradition, evidence for $\omega > 1$ at a site over a branch (i.e., evidence for positive selection) is assumed to indicate an episode of adaptive evolution. The tradition of equating positive selection with adaptation has recently been challenged by the recognition that $\omega > 1$ can be caused by non-adaptive processes (Jones et al 2017). In what follows it will be shown that $\omega > 1$ should no longer be taken as the key signature of change. An alternative approach is to make use of patterns of heterotachy to infer adaptation.

Adaptive and Non-Adaptive Processes are Confounded

(a challenge to the standard $\omega > 1$ paradigm)

one observed effect (heterotachy), two possible causes



Z = heterotachy with episodic $\omega > 1$

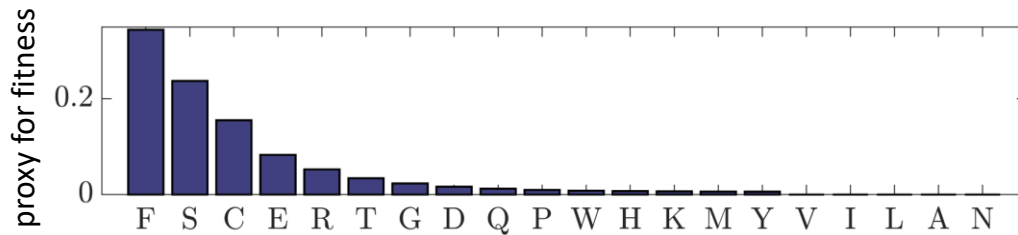
X = adaptive process

Y = non-adaptive process

The issue in broad strokes is that there are adaptive processes (X) and non-adaptive processes (Y) that can both generate patterns in an alignment (Z) consistent with episodic elevations in dN/dS to values $\omega > 1$. The two processes X and Y are said to be **confounded** in the data (Jones et al 2018), meaning that its not possible to discern which process (or what combination of the two) generated Z. The mechanisms that give rise to the confounding of adaptive and non-adaptive processes are explained in the remainder of this section. The explanation starts with a site-specific fitness landscape for the 20 amino acids.

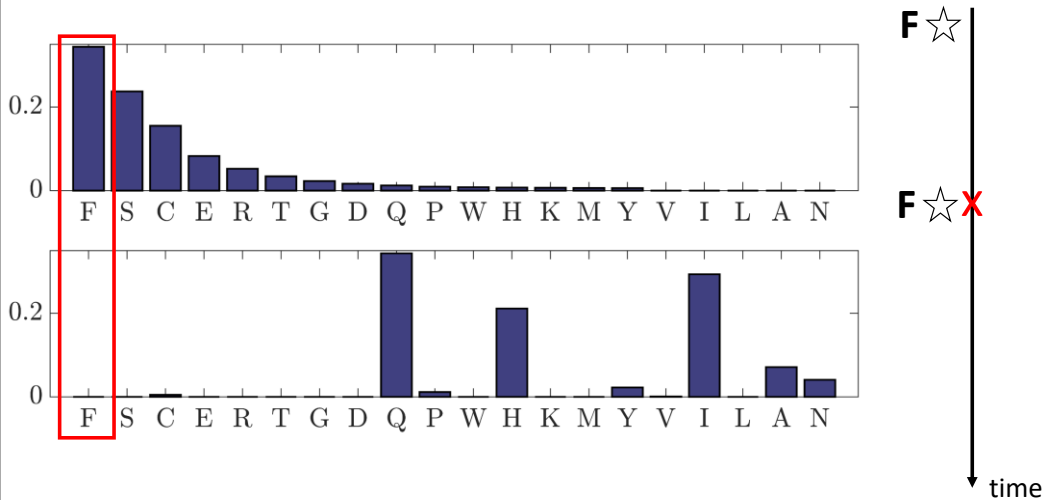
A Site-Specific Fitness Landscape

(a tool to explain confounding)



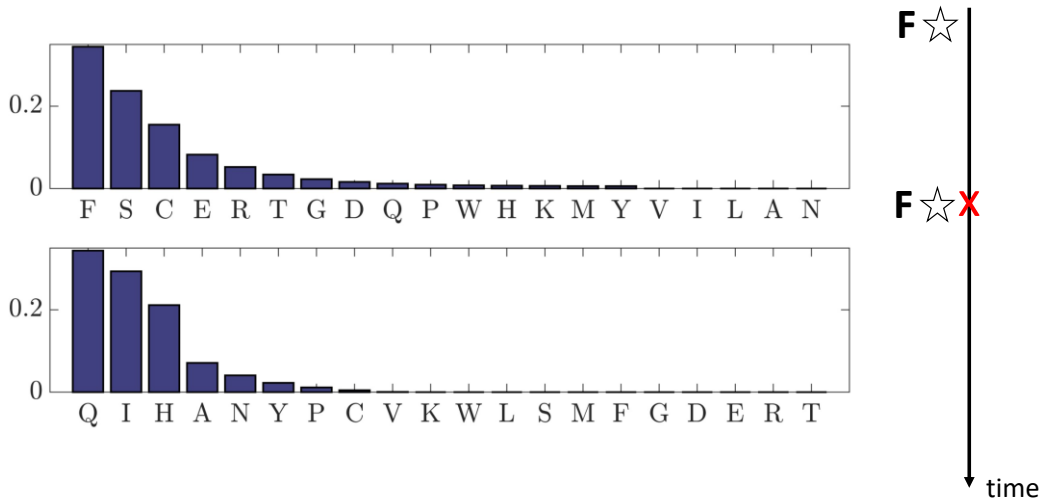
The bar plot illustrates a metaphorical **site-specific fitness landscape** for a codon site in a protein-coding sequence. Each bar corresponds to one of the 20 amino acids with height serving as a proxy for relative fitness. In this case, the organism is fittest when the amino acid at the site is phenylalanine (F). The highest bar corresponds to the fitness **peak** of the landscape. Down slope to the right is the **tail** of the landscape corresponding to amino acids that confer lower fitness.

An Adaptive Peak Shift = a change in amino acid fitnesses



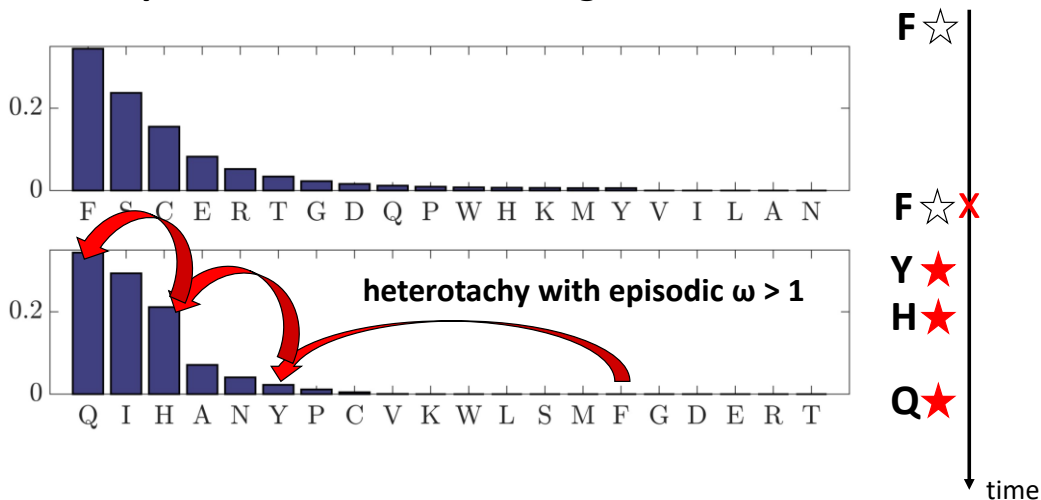
Let us define an adaptive peak shift to be a change in the location of the peak of the site-specific fitness landscape. In this case, the peak of the landscape was once at phenylalanine F. But at the time indicated by the red X the relative fitnesses of the amino acids changed to make glutamine Q the peak. For convenience, let us sort the amino acids in the new landscape according to height (next slide).

An Adaptive Peak Shift = a change in amino acid fitnesses



Sorting makes it easy to see that the new peak is at Q and that F is in the tail of the new landscape. The site was occupied by F when the landscape changed. Following that event, any mutation that changes the amino acid to something closer to the new fitness peak at Q would have a good chance of being fixed. Hence, we might expect the peak shift from F to Q (a change in the site-specific landscape) to be followed by a rapid series of nonsynonymous substitutions from F toward Q.

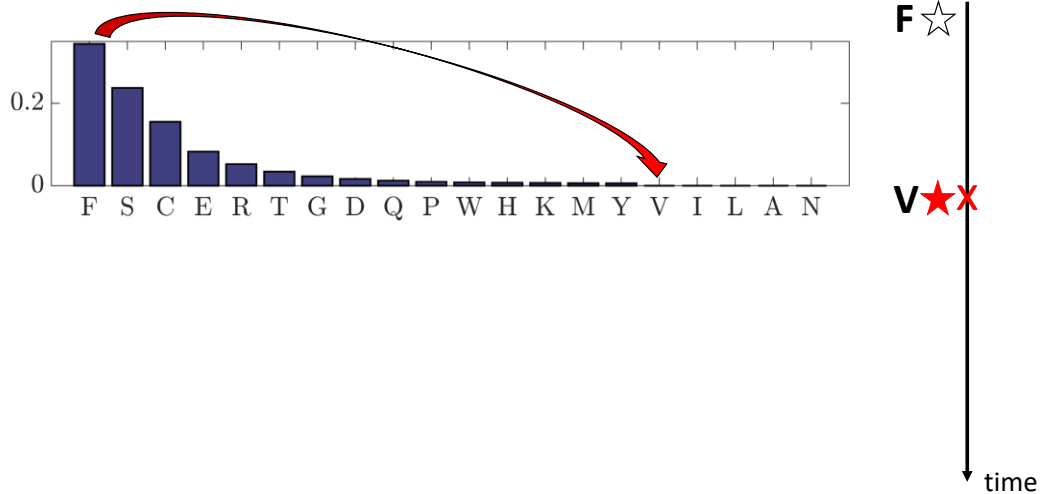
An Adaptive Peak Shift = a change in amino acid fitnesses



Here, for example, we image a rapid sequence of nonsynonymous substitutions from F (TTT) to Y(TAT) to H(CAT) to Q (CAG)*. Whatever the sequence of events, evidence for a change from F to Q at the site over a relatively short branch of the phylogenetic tree might easily result in inference of $\omega > 1$ at that site over that branch. **In this instance, equating positive selection to adaptation would be correct.**

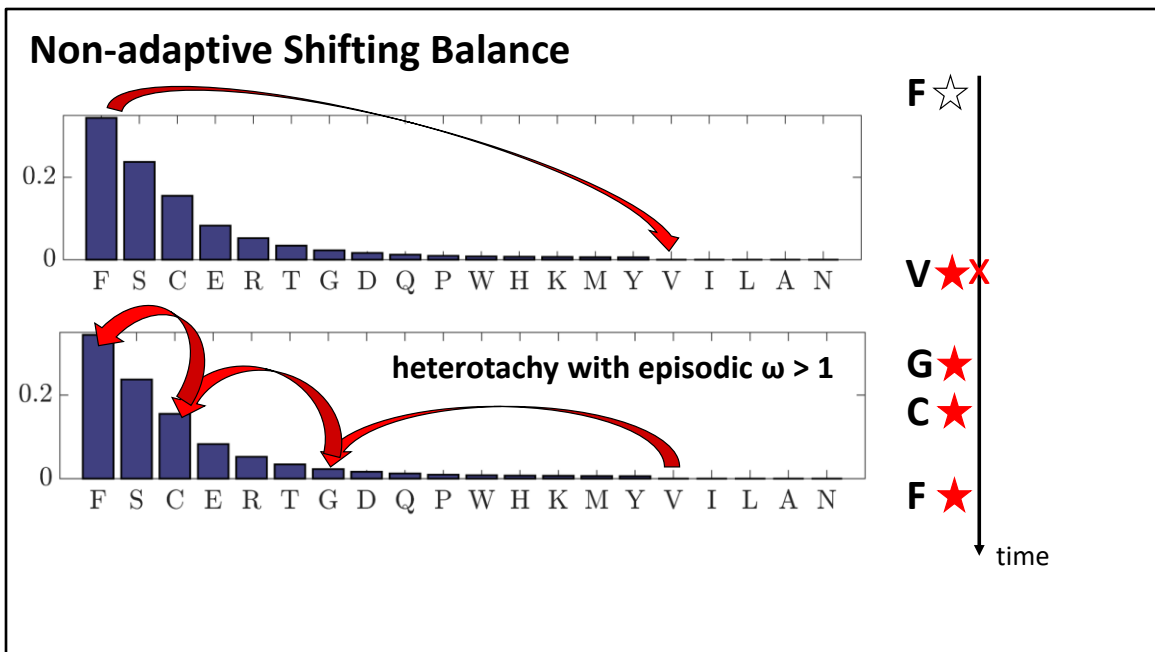
*Most codon substitution models assume codons evolve in single nucleotide steps. However, double and even triple nucleotide mutations can sometime occur and be fixed. It has been suggested that codon substitution models should be modified to account for such events but see Jones et al. 2018 for a discussion of the potential problem with this suggestion.

Non-adaptive Shifting Balance



Returning to the original site-specific fitness landscape, imagine that a mutation from the peak at F (TTT) to somewhere in the tail, valine V (GTT) say, was fixed by chance.

Genetic drift of this kind would be rare depending on the relative fitness of V compared to F but can sometimes happen. Since V corresponds to low fitness, any mutation that changes the amino acid to something closer to the fitness peak would have a good chance of being fixed.



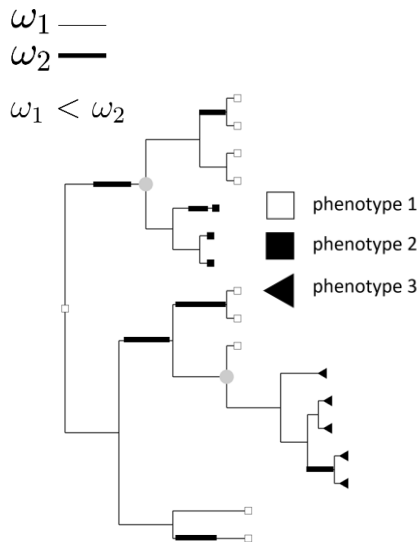
Here, for example, we image a rapid sequence of nonsynonymous substitutions from V (GTT) to G (GGT) to C (TGT) and then back to F (TTT). Whatever the sequence of events, evidence for a change from V to F at the site over a relatively short branch of the phylogenetic tree might easily result in inference of $\omega > 1$ at that site over that branch. **In this instance, equating positive selection to adaptation would be incorrect.** We call the process whereby a site drifts away from its fitness peak and then returns by a combination of positive selection and drift “non-adaptive shifting balance” (Jones et al 2017). Codon substitution models designed to detect episodic positive selection at individual sites over single branches of the tree cannot distinguish between $\omega > 1$ caused by adaptive peak shifts from $\omega > 1$ caused by non-adaptive shifting balance. The adaptive and non-adaptive processes are therefore confounded with respect to those models (Jones et al 2018).



The PG-BSM

The PG-BSM illustrates how the new model mitigates the problem of confounding by introducing additional information into the analysis.

Breaking confounding using a covarion-like model with phenotype.

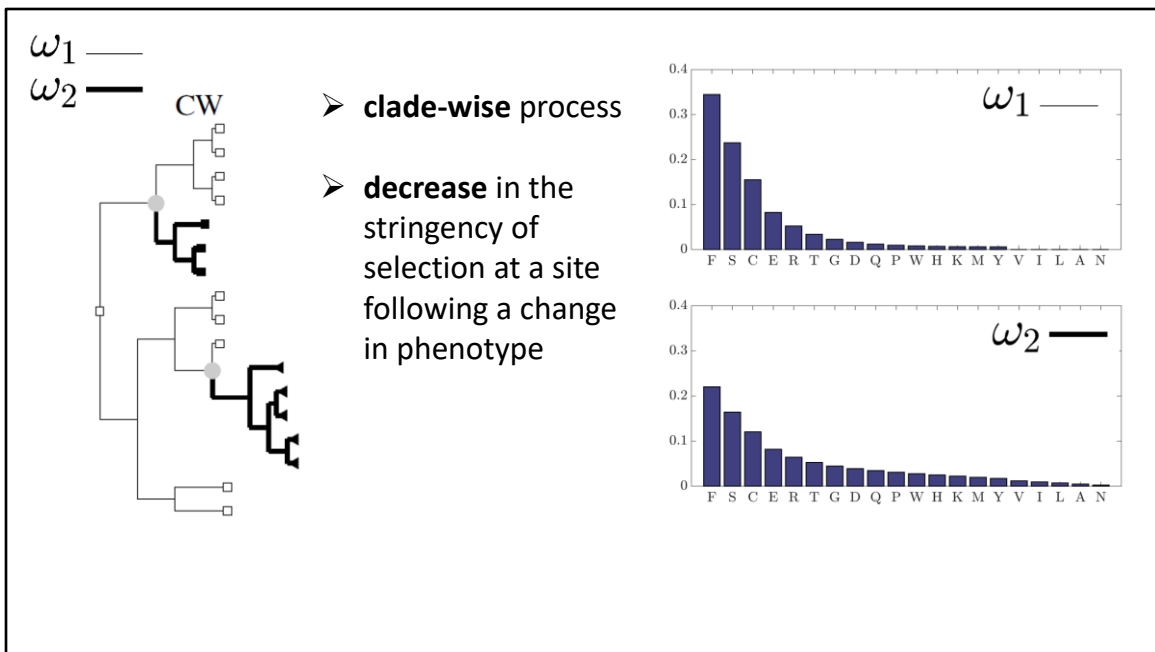


- The phenotype-genotype branch-site model combines a covarion-like model with a model for the evolution of a discrete phenotype.
- The **covarion-like model** does not assign a cause to heterotachy (e.g., dN/dS is not assumed to indicate adaptation) but only accounts for **heterotachy-by-any-cause**.
- Under the **null model** the genotype and phenotype are assumed to have evolved independently.
- The **alternate model** seeks to detect dependencies between genes and phenotype (PG-association) by identifying modes of heterotachy consistent with specific mechanisms of adaptation that co-occur with changes in phenotype

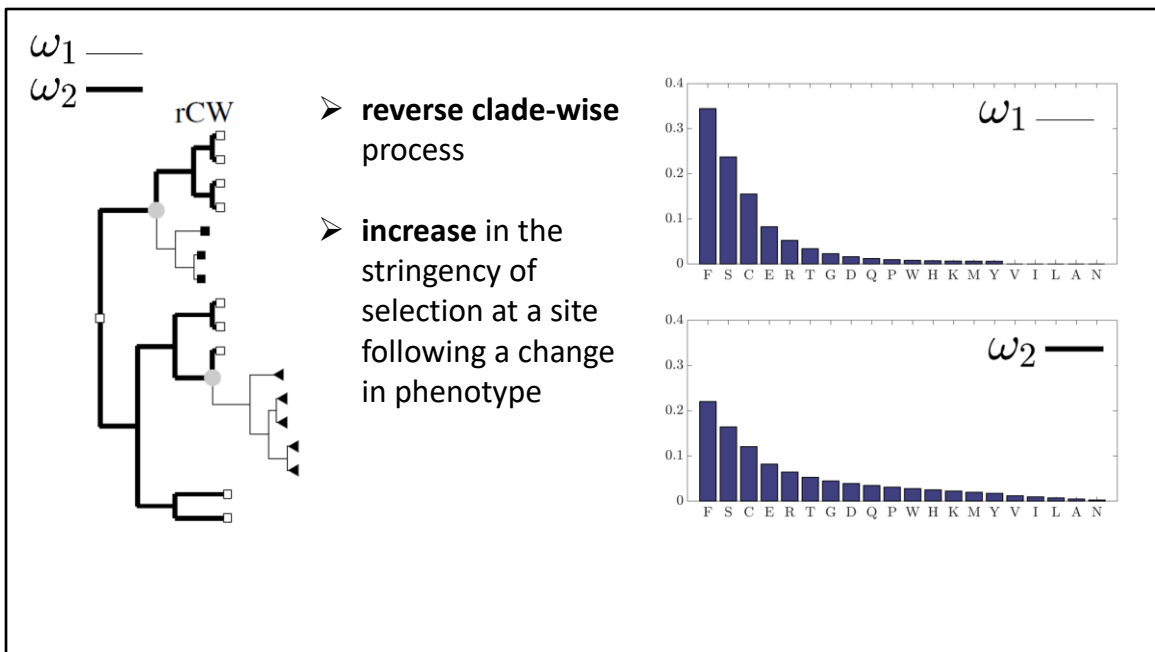
The **null PG-BSM** has two model components. The first is a covarion-like model for detecting heterotachy in the form of random changes between $\omega_1 < \omega_2$ over time that might include $\omega_2 > 1$. The covarion-like model is protected against confounding because it does not attempt to assign a cause to $\omega_2 > 1$. Specifically, the model does not equate evidence of $\omega_2 > 1$

to adaptation. The second component of the null PG-BSM accounts for changes in a discrete phenotype over the tree. The phenotype is assumed to evolve independently of the genotype, meaning that there is no assumed association between the pattern of heterotachy at any site and the pattern of change in the phenotype. The **alternate PG-BSM** includes additional model components that can detect phenotype-genotype dependences in the form of modes of heterotachy at individual sites consistent with specific mechanisms of adaptation (or changes in site-specific fitness landscapes) that co-occur with changes in phenotype. The additional components of the alternate PG-BSM can include one or more of the following: a model for **clade-wise (CW)** change in rate ratio at a site, a model for **reverse clade-wise change**

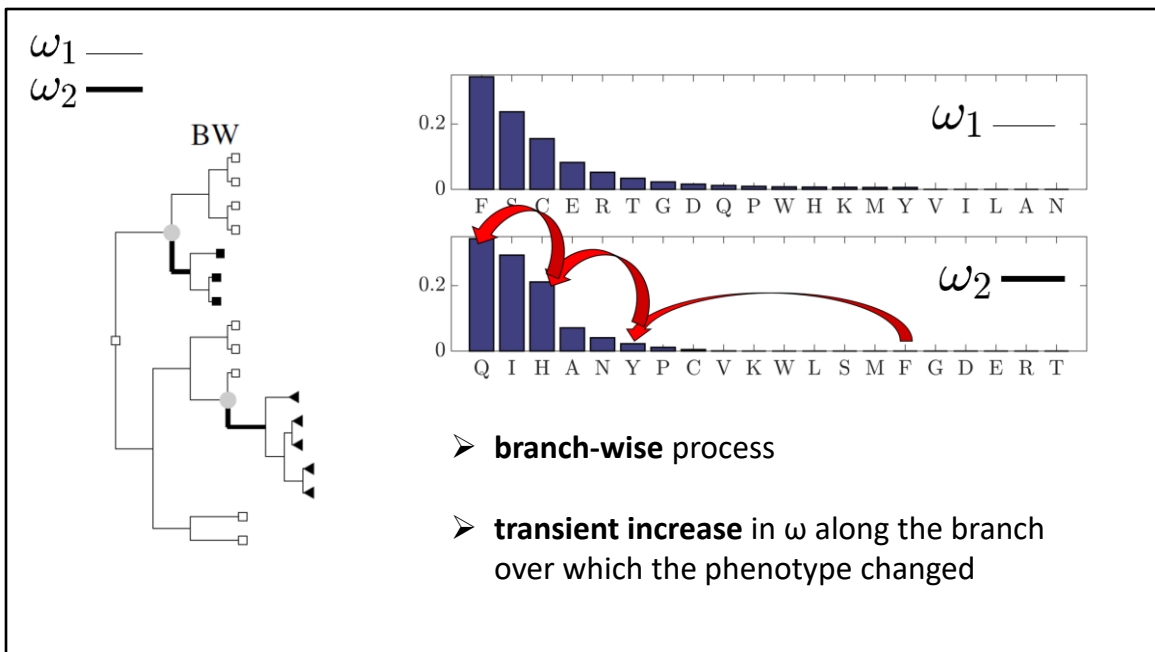
(rCW) in rate ratio at a site, and a model for **branch-wise change (BW)** in rate ratio at a site. Each respective change is assumed to co-occur with a change in phenotype, and if detected provides evidence of a change in a site-specific landscape consistent with adaptation.



This slide depicts the clade-wise process. The two gray circles on the phylogenetic tree indicate places where the phenotype was inferred to have changed from type 1 (white square) to types 2 (black square) and 3 (black triangle). The thickness of the branches indicate the inferred rate ratio at a specific site. In this case the rate ratio was always the smaller ω_1 before the phenotype changed and the larger ω_2 after the phenotype changed. This suggests a persistent relaxation in selection pressure in each of the clades descending from the branch over which the phenotype changed. The bar plots illustrate a hypothetical site-specific landscape before and after the change in phenotype. Notice that the order of amino acids does not change, but only the height of the fitness peak. Relaxation of selection pressure corresponds to a lowering of the peak and an increase in the rate ratio from ω_1 to ω_2 . Hence, a **mode of heterotachy** (a clade-wise switch to a higher rate ratio) that corresponds to a **mechanism of adaptation** (relaxation of selection pressure) and that **co-occurred with a change in the phenotype** can be used to infer adaptation.



This slide depicts the **reverse clade-wise process**. In this case the rate ratio was always the larger ω_2 before the phenotype changed and the smaller ω_1 after the phenotype changed. This suggests a persistent increase in selection pressure in each of the clades descending from the branch over which the phenotype changed. Here too the order of amino acids did not change on the site-specific landscape, but only the height of the fitness peak. An increase in selection pressure corresponds to an increase in the height of the fitness peak and a decrease in the rate ratio from ω_2 to ω_1 . Hence, a **mode of heterotachy** (a clade-wise switch to a lower rate ratio) that corresponds to a **mechanism of adaptation** (an increase in selection pressure) and that **co-occurred with a change in the phenotype** can be used to infer adaptation.



This slide depicts the branch-wise process. In this case the rate ratio at the site was the smaller ω_1 everywhere in the tree except for the branches over which the phenotype was inferred to have changed, where it was the larger ω_2 . This transient increase in rate ratio is consistent with an adaptive peak shift that co-occurred with the change in phenotype. Once again this illustrates how a **mode of heterotachy** (a transient increase in rate ratio) that corresponds to a **mechanism of adaptation** (a change in the relative fitnesses of the amino acids) and that **co-occurred with a change in the phenotype** can be used to infer adaptation (in this case in the form of an adaptive peak shift).

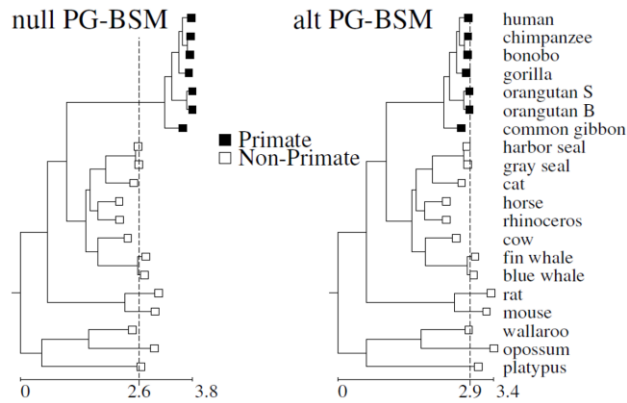
In all three cases, the CW, rCW and BW process can be inferred regardless of whether $\omega_2 > 1$. The PG-BSM therefore abandons the traditional method of inferring adaptation. Note however, that like all codon substitution models, the inference of adaptation is not to be taken as a conclusion, but a hypothesis. Sites identified as having undergone adaptation in concert with a change in phenotype should be considered candidates in a search for sites associated with that phenotype. Further in vitro or in vivo (rather than in silico) analyses would be required to provide more support for any conclusion of a phenotype-genotype association.



An Example

An Example provides an illustration of the PG-BSM applied to real data.

Analysis of mammalian mtDNA



Model	log-likelihood	(w_1, w_2)	(P_{CW}, P_{BW})
null PG-BSM	-88,723	(0.02, 0.35)	
alt PG-BSM	-88,685	(0.02, 0.31)	(0.06, 0.04)

(Jones et al. 2020)

The trees depict the phylogeny for the concatenation of 12 H-strand mitochondrial DNA sequences (3,331 codon sites) from 20 mammalian species distributed by the PAML software package (Yang ZH. 2007. PAML4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591). The topology is that reported in [Cao et al. \(1998\)](#) (Cao Y, Janke A, Waddell PJ, Westerman M, Takenaka O, Murata S, Okada N, Paabo S, Hasegawa M. 1998. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. *J Mol Evol.* 47:307–322). The same data set has been used in several studies to test new models (see Jones et al.

2020).

The scale on the horizontal axis in each tree is the expected number of single nucleotide substitution per codon site. Branch lengths on the left are those estimated by fitting the null PG-BSM to the alignment assuming the illustrated topology. Branch lengths on the right are those estimated using the alternate PG-BSM. The log-likelihood for the contrast $2(88,723 - 88,685) = 76$ on two degrees of freedom (for the proportion of CW and BW sites, P_{CW} and P_{BW} , estimated under the alternate model) is highly significant. Sites patterns consistent with the relaxation of selection pressure (CW sites) and others consistent with a peak shift (BW sites) that occurred over the branch leading to the primate clade were detected (next slide).

$0.87 \geq P(CW) \geq 0.77$												$0.75 \geq P(BW) \geq 0.67$						
Y	A	T	A	V	D	S	S	I	A	I	I	A	P	T	H	S	N	human
H	A	T	A	V	D	S	S	V	A	I	I	A	P	T	H	S	N	chimpanzee
H	A	T	A	V	D	S	S	V	A	I	I	A	P	T	H	S	N	bonobo
Y	A	T	A	V	D	S	S	V	A	I	I	A	P	T	H	S	N	gorilla
H	A	T	A	V	D	S	S	V	A	I	I	A	P	T	H	S	N	orangutan S
Y	A	T	A	V	D	S	S	V	A	I	I	A	P	T	H	S	N	orangutan B
Y	A	T	A	V	D	S	S	V	A	I	I	A	P	T	H	S	N	common gibbon
R	S	A	I	I	K	K	A	F	A	V	V	I	S	C	C	E	G	harbor seal
R	S	A	I	I	K	K	A	F	A	V	V	I	S	C	C	E	G	gray seal
R	S	A	I	I	K	K	A	F	A	V	V	I	S	C	C	E	G	cat
R	S	A	I	I	K	K	A	F	A	V	V	I	S	C	C	E	G	horse
R	S	A	I	I	K	K	A	F	A	V	V	I	S	C	C	E	G	rhinoceros
R	S	A	I	I	K	K	A	F	A	V	V	I	S	C	C	E	G	cow
R	S	A	I	I	K	K	A	F	A	V	V	I	S	C	C	E	G	fin whale
R	S	A	I	I	K	K	A	F	A	V	V	I	S	C	C	E	G	blue whale
R	S	A	I	I	K	K	A	F	A	V	V	I	S	C	C	E	G	rat
R	S	A	I	I	K	K	A	F	A	V	V	I	S	C	C	E	G	mouse
R	S	A	I	I	K	K	A	F	A	V	V	I	S	C	C	E	G	wallaroo
R	S	A	I	I	K	K	A	F	A	V	V	I	S	C	C	E	G	opossum
2	2	3	3	3	3	3	2	2	3	2		I	S	C	C	E	G	platypus

These are the site patterns identified by the PG-BSM to be consistent with the CW process (left) and the BW process (right). Amino acids are color coded according to their physicochemical properties. The indicated probabilities are posteriors (see Jones et al. 2020 and Jones et al. 2020 SI for details). The CW sites exhibit low variability among non-primates consistent with the smaller estimated rate ratio $\omega_1 = 0.02$ and higher variability among primates consistent with the larger estimated rate ratio $\omega_1 = 0.31$, patterns that suggest relaxation of selective pressure in the primate clade. The BW sites exhibit low variability in both clades, but different amino acids in each, consistent with sites that evolved under stringent selection but that were subject to a peak shift along the branch separating the two clades.