Models of amino acid and codon substitution

Comparative Genomic Analyses

Rui Borges
Vetmeduni Vienna

Phylogenetic models

Models of sequence substitution can be built for different evolutionary units

- nucleotides: 4 states $\{A, C, G, T\}$
- amino acids: 20 states $\{Phe, Leu, Ile, ...\}$
- codons: 61 states $\{AAA, AAC, AAT, ...\}$

Models of amino acid and codon substitution

Substitutions between amino acids in proteins or between codons in protein-coding genes can be very informative

- natural selection operates mainly at the protein level
- **synonymous or silent substitutions**: nucleotide substitutions that do not change the encoded amino acid
- **nonsynonymous or replacement substitutions**: those that change the amino acid

Amino acid and codon substitutions

	U		C		A		G		
	Codon	Amino acid	Codon	Amino acid	Codon	Amino acid	Codon	Amino acid	
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	**U**
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	**C**
	UUA	Leu	UCA	Ser	UAA	STOP	UGA	STOP	**A**
	UUG	Leu	UCG	Ser	UAG	STOP	UGG	Trp	**G**
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	**U**
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	**C**
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	**A**
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	**G**
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	**U**
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	**C**
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	**A**
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg	**G**
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	**U**
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	**C**
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	**A**
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	**G**

Types of phylogenetic models

Empirical models

- describe the relative rates of substitution
- do not consider explicitly factors that influence the evolutionary process
- large quantities of sequence data

Mechanistic models

- consider the biological process involved: mutational biases, natural selection...
- more interpretative power
- particularly useful for studying the evolutionary forces and mechanisms

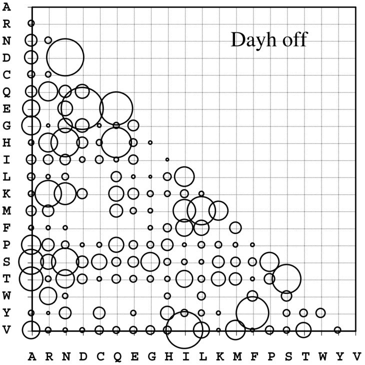
Models of amino acid replacement

First empirical amino acid substitution matrix

DAYHOFF78

Dayhoff et al. (1978)

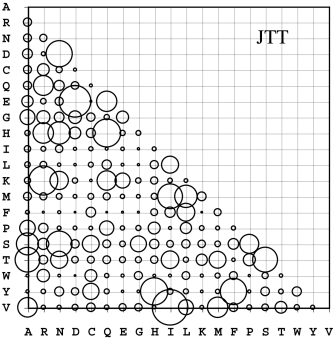
- protein sequences available at the time
- parsimony argument was used to reconstruct ancestral protein sequences and transitions



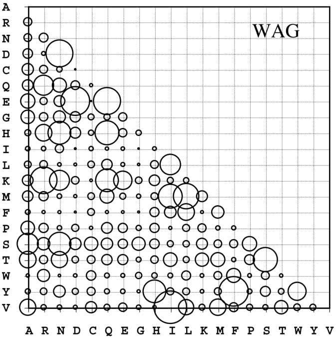
Models of amino acid replacement

Other empirical substitution matrices

JTT92



WAG01



Models of amino acid replacement

Empirical substitution matrices:

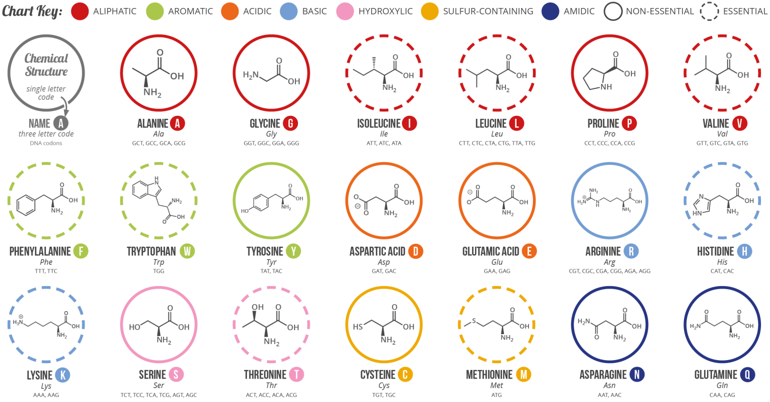
- nuclear proteins: DAYHOFF, JTT, WAG
  Whelan and Goldman (2001)
- mitochondrial proteins: mtMAM, mtREV
  Adachi and Hasegawa (1966)
- chloroplast proteins: cpREV
  Adachi et al. (2000)

Amino acid replacements

Several features of these matrices are worth noting:

- the genetic code has a major impact on the interchange rates
- acids with similar physicochemical properties interchange more than dissimilar amino acids

Amino acid replacements



Models of amino acid replacement

Models of amino acid replacement have several applications:

- phylogeny reconstruction
- alignment of protein sequences: can be used as cost matrices to penalize mismatches (heavier penalties applied to rarer changes)

Codon evolution and selection

Understanding the selective pressures underlying genetic variation is a central goal in evolutionary biology

- nonsynonymous mutations can directly affect protein function
- nonsynonymous mutations are more likely to influence the fitness of an organism than synonymous mutations

Codon evolution and selection

Comparing the relative rates of non-synonymous and synonymous substitutions became a standard measure of selective pressure

Miyata and Yasunaga (1980)

$$\omega = \frac{dN}{dS}$$

- $\omega \approx 1$: signifies **neutral evolution**
- $\omega < 1$: **negative selection**
- $\omega > 1$: **positive selection**

Models of codon substitution

The models of codon evolution describe substitution from one codon to another

- codon triplet is the unit of evolution
- the state space includes only the sense codons (stop codons are ignored)
- the genetic code is not universal

Models of codon substitution

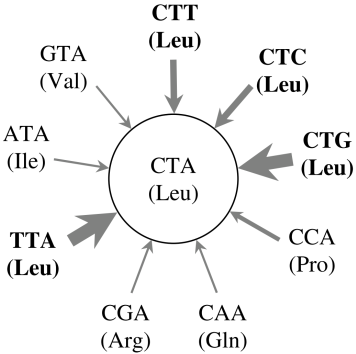
$Q = \{q_{ij}\}$: the instantaneous rate from codon i to j

Nielsen & Yang (1998)

$$q_{ij} = \begin{cases} \pi_j & i \text{ and } j \text{ synonymous transition} \\ \kappa\pi_j & i \text{ and } j \text{ synonymous transversion} \\ \omega\pi_j & i \text{ and } j \text{ nonsynonymous transition} \\ \omega\kappa\pi_j & i \text{ and } j \text{ nonsynonymous transversion} \\ 0 & i \text{ and } j \text{ more than one DNA substitution} \end{cases}$$

- κ: transition/transversion rate ratio,
- ω: nonsynonymous/synonymous rate ratio
- π_j: equilibrium frequency of codon j

Models of codon substitution



- synonymous transversion:
  CTG → CTA
- synonymous transition:
  CTT → CTA
- nonsynonymous transversion:
  CCA → CTA
- nonsynonymous transition:
  CAA → CTA

Codon evolution and selection

Condon models help to answer several questions:

- Is there evidence of selection operating on a gene?
- Where did selection happen?
- When did selection happen?

Testing for positive selection

To test for positive selection, the null hypothesis is often the neutral scenario and the alternative allowing for positive selection.

- null hypothesis: ω is constrained to be smaller than 1
- alternative hypothesis: ω higher than 1 (i.e., allowing for diversifying selection)

Testing for positive selection

There are several different tests for positive selection: the M7 vs. M8 is a widely used model comparison.

- the null model M7 (beta) assumes a beta distribution for ω
- the alternative model M8 (beta&ω) adds an extra class of sites under positive selection with $\omega > 1$.
- LRT$\rightarrow \chi^2$ with two degrees of freedom

Model	**Description**	**Free parameters**
M7: Beta	All sites are from $B(\alpha, \beta)$	α and β
M8: Beta & ω	p_0 sites from $B(\alpha, \beta)$, $p_1 = 1 - p_0$ sites with $\omega > 1$	α, β, p_0 and ω

Literature

Computational Molecular Evolution by Yang (2006)
Oxford University Press

- Chapter 2: sections 2.1, 2.2 and 2.4