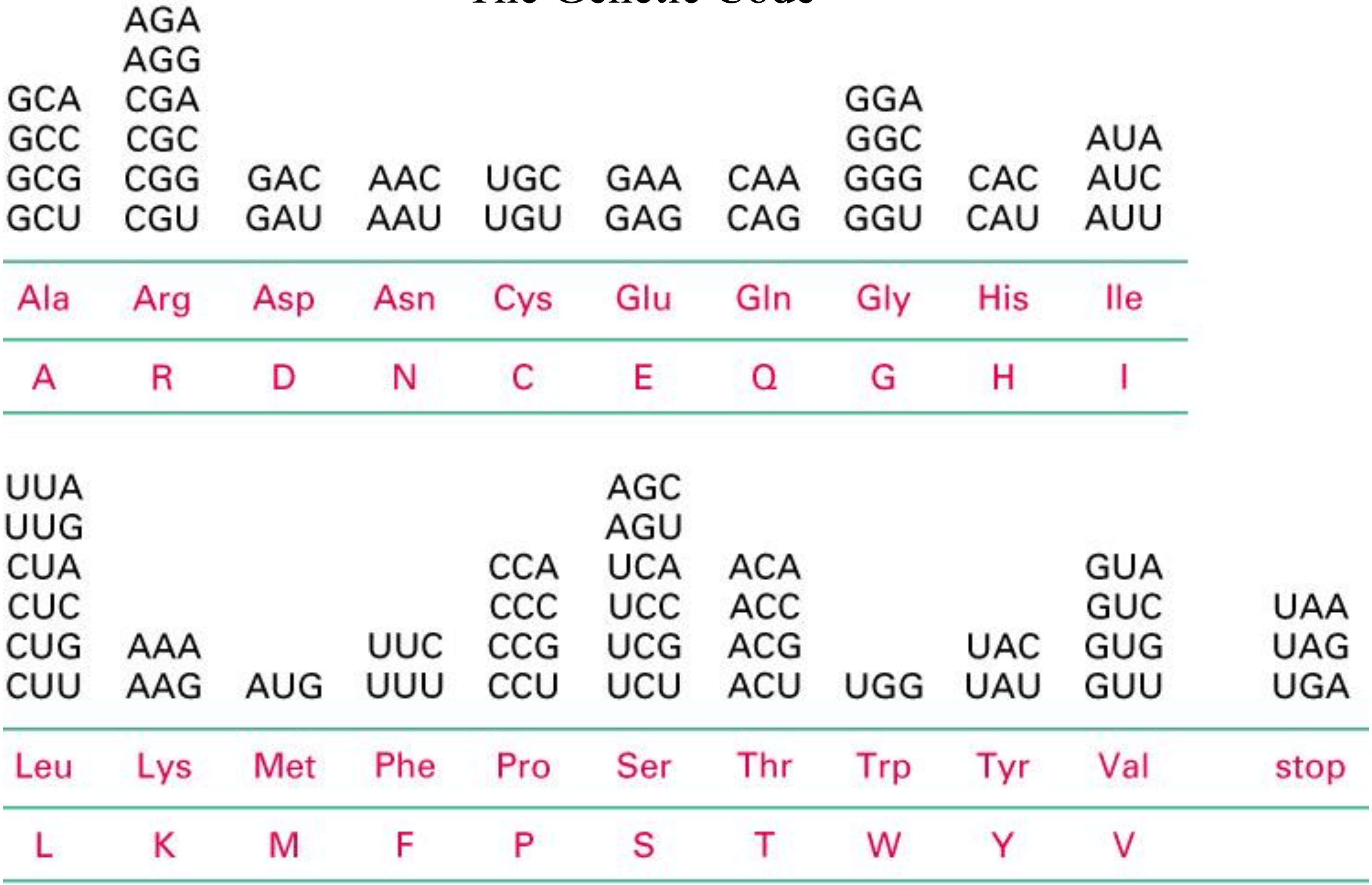
# The Genetic Code



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| GCA<br>GCC<br>GCG<br>GCU | AGA<br>AGG<br>CGA<br>CGC<br>CGG<br>CGU | GAC<br>GAU | AAC<br>AAU | UGC<br>UGU | GAA<br>GAG | CAA<br>CAG | GGA<br>GGC<br>GGG<br>GGU | CAC<br>CAU | AUA<br>AUC<br>AUU |
| Ala | Arg | Asp | Asn | Cys | Glu | Gln | Gly | His | Ile |
| A | R | D | N | C | E | Q | G | H | I |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| UUA<br>UUG<br>CUA<br>CUC<br>CUG<br>CUU | AAA<br>AAG | AUG | UUC<br>UUU | CCA<br>CCC<br>CCG<br>CCU | AGC<br>AGU<br>UCA<br>UCC<br>UCG<br>UCU | ACA<br>ACC<br>ACG<br>ACU | UGG | UAC<br>UAU | GUA<br>GUC<br>GUG<br>GUU | UAA<br>UAG<br>UGA |
| Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val | stop |
| L | K | M | F | P | S | T | W | Y | V | |

Figure 6–50. Molecular Biology of the Cell, 4th Edition.

# The Reading Frames

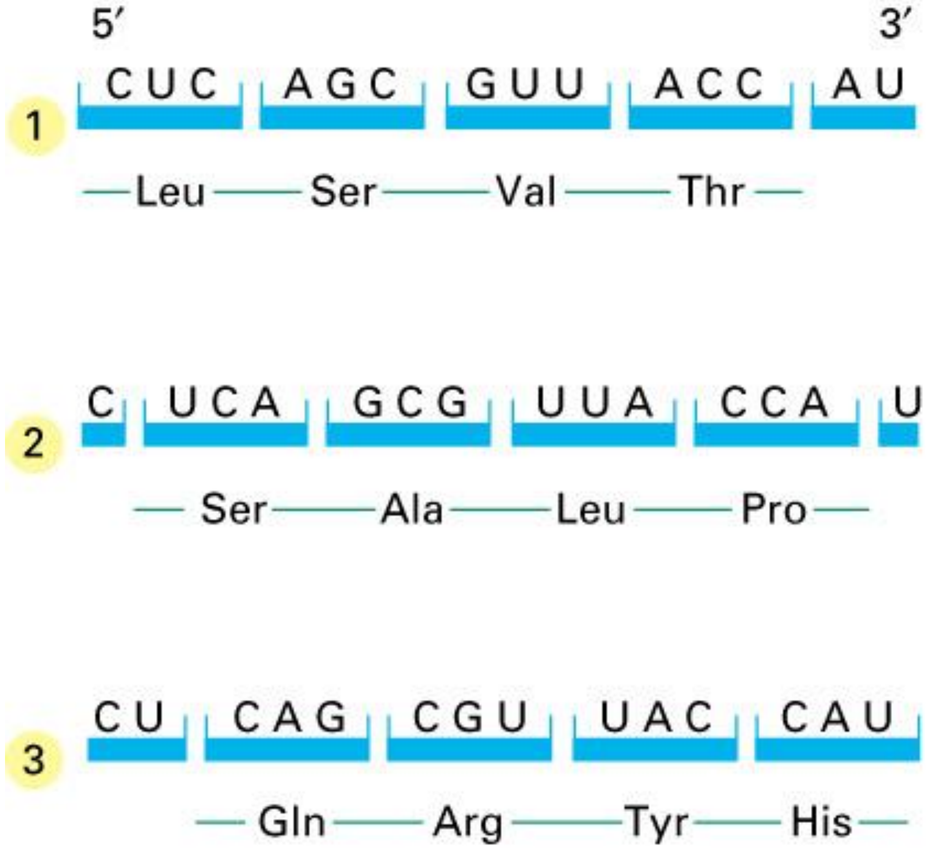


Figure 6–51. Molecular Biology of the Cell, 4th Edition.

# tRNA (clover leaf shape with four strands folded, finally L-shape)

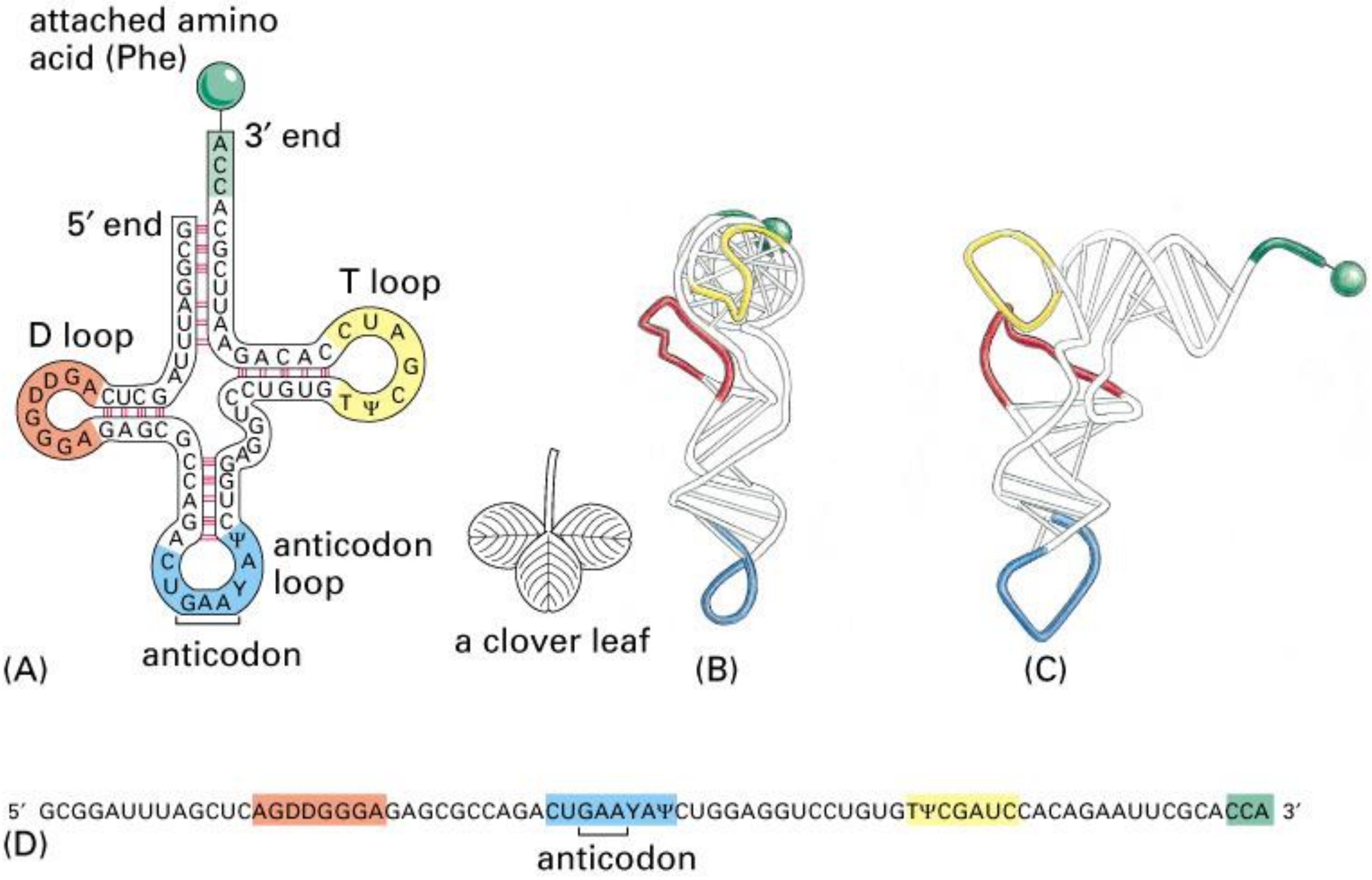
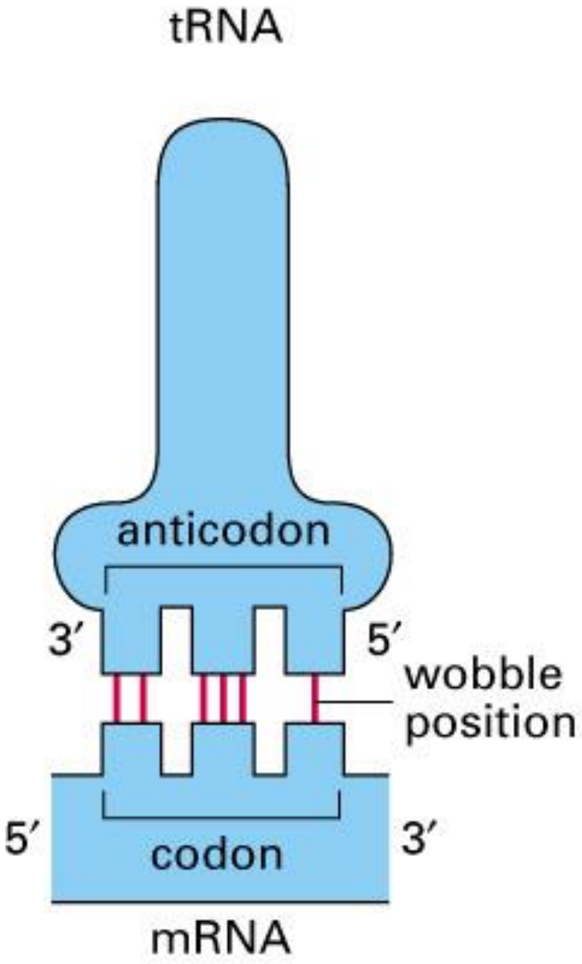

attached amino acid (Phe)
3′ end
5′ end
T loop
D loop
anticodon loop
anticodon
(A)
a clover leaf
(B)
(C)
5′ GCGGAUUUAGCUCAGDDGGGAGAGCGCCAGACUGAAYAΨCUGGAGGUCCUGUGTΨCGAUCCACAGAAUUCGCACCA 3′
(D)
anticodon

Figure 6–52. Molecular Biology of the Cell, 4th Edition.

# tRNA and mRNA pairing



bacteria

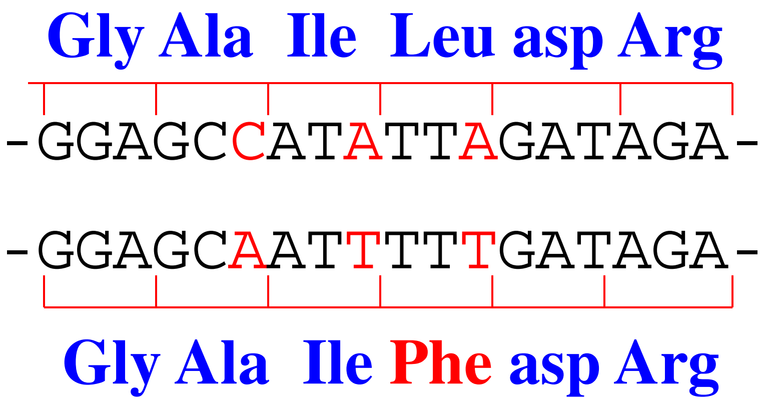
| wobble codon base | possible anticodon bases |
|---|---|
| U | A, G, or I |
| C | G or I |
| A | U or I |
| G | C or U |

eucaryotes

| wobble codon base | possible anticodon bases |
|---|---|
| U | G or I |
| C | G or I |
| A | U |
| G | C |

Figure 6–53. Molecular Biology of the Cell, 4th Edition.

# Nucleotide, amino-acid sequences



Gly Ala Ile Leu asp Arg

-GGAGCCATATTAGATAGA-

-GGAGCAATTTTTGATAGA-

Gly Ala Ile Phe asp Arg

-> gene

-> protein

- **3 different DNA positions but only one different amino acid position:**

**2 of the nucleotide substitutions are therefore synonymous and one is non-synonymous.**

DNA yields more phylogenetic information than proteins. **The nucleotide sequences of a pair of homologous genes have a higher information content than the amino acid sequences of the corresponding proteins**, because mutations that result in synonymous changes alter the DNA sequence but do not affect the amino acid sequence.

# Standard genetic code

- The genetic code specifies how a combination of any of the four bases (A,G,C,T) produces each of the 20 amino acids.

- The triplets of bases are called codons and with four bases, there are 64 possible codons:

($4^3$) possible codons that code for 20 amino acids (and stop signals).

# Standard genetic code

- Because there are only 20 amino acids, but 64 possible codons, the same amino acid is often encoded by a number of different codons, which usually differ in the third base of the triplet.

- Because of this repetition the genetic code is said to be **degenerate** and codons which produce the same amino acid are called **synonymous codons**.

# Important properties inherent to the standard genetic code

# Synonymous vs nonsynonymous substitutions

- **Nondegenerate sites: are codon position where mutations always result in amino acid substitutions.**

**(exp. TTT (Phenylalanyne, CTT (leucine), ATT (Isoleucine), and GTT (Valine)).**

- **Twofold degenerate sites: are codon positions where 2 different nucleotides result in the translation of the same aa, but the 2 others code for a different aa.**

**(exp. GAT and GAC code for Aspartic acid (asp, D),**

**whereas GAA and GAG both code for Glutamic acid (glu, E)).**

- **Threefold degenerate site: are codon positions where changing 3 of the 4 nucleotides has no effect on the aa, while changing the fourth possible nucleotide results in a different aa.**

**There is only 1 threefold degenerate site: the 3rd position of an isoleucine codon. ATT, ATC, or ATA all encode isoleucine, but ATG encodes methionine.**

# Standard genetic code

- **Fourfold degenerate sites: are codon positions where changing a nucleotide in any of the 3 alternatives has no effect on the aa.**

**exp. GGT, GGC, GGA, GGG(Glycine);**

**CCT,CCC,CCA,CCG(Proline)**

- **Three amino acids: Arginine, Leucine and Serine are encoded by 6 different codons:**

- **Five amino-acids are encoded by 4 codons which differ only in the third position. These sites are called "fourfold degenerate" sites**

# Standard genetic code

- Nine amino acids are encoded by a pair of codons which differ by a transition substitution at the third position. These sites are called "twofold degenerate" sites.

Transition:

A/G; C/T

- Isoleucine is encoded by three different codons

- Methionine and Triptophan are encoded by single codon

- Three stop codons: TAA, TAG and TGA

# Standard Genetic Code

**Nucleotide substitutions in protein coding genes can be divided into :**

**• *synonymous* (or silent) substitutions i.e. nucleotide substitutions that do not result in amino acid changes.**

**• *non synonymous* substitutions i.e. nucleotide substitutions that change amino acids.**

**• nonsense mutations, mutations that result in stop codons.**

**exp: Gly: any changes in 3rd position of codon results in Gly; any changes in second position results in amino acid changes; and so is the first position.**

↓ ↓

**exp:** **AGC** **Ser**

# Nonsynonymous/synonymous substitutions

- Estimation of synonymous and nonsynonymous substitution rates is important in understanding the dynamics of molecular sequence evolution.

- As synonymous (silent) mutations are largely invisible to natural selection, while nonsynonymous (amino-acid replacing) mutations may be under strong selective pressure, comparison of the rates of fixation of those two types of mutations provides a powerful tool for understanding the mechanisms of DNA sequence evolution.

- For example, variable nonsynonymous/synonymous rate ratios among lineages may indicate adaptative evolution or relaxed selective constraints along certain lineages.

- Likewise, models of variable nonsynonymous/synonymous rate ratios among sites may provide important insights into functional constraints at different amino acid sites and may be used to detect sites under positive selection.

# Codon usage

- There are 64 ($4^3$) possible codons that code for 20 amino acids (and stop signals).

- If nucleotide substitution occurs at random at each nucleotide site, every nucleotide site is expected to have one of the 4 nucleotides, A, T, C and G, with equal probability.

- Therefore, if there is no selection and no mutation bias, one would expect that the codons encoding the same amino acid are on average in equal frequencies in protein coding regions of DNA.

- In practice, the frequencies of different codons for the same amino acid are usually different, and some codons are used more often than others. This codon usage bias is often observed.

- Codon usage bias is controlled by both mutation pressure and purifying selection.

# Codon Adaptation Index (CAI)

**In recognition of the role of selection in producing high codon bias, a statistic called Codon Adaptation Index (or CAI) is calculated.**

**Pattern of codon usage in very highly expressed genes can reveal:**

**(i) which of the alternative synonymous codons for an amino acid is the most efficient for translation;**

**(ii) the relative extent to which other codons are disandvantageous**

**Sharp, PM & Li WH (1987). NAR 15:p.1281-1295.**

# RSCU

- **Relative Synonymous Codon Usage :**

**a statistical measure of codon usage bias**

**RSCU = $X_{ij}$ /(1/$n_i$ *Σ{$X_{ij}$; j=1, $n_i$ })**

**where $X_{ij}$ is the number of occurrences of the j^th codon for the i^th amino acid, and $n_i$ is the number of (from 1 to 6) of alternative codons for the i^th amino acid.**

**i.e. the observed number of the j^th codon for the amino-acid i normalized by the average number of all codons coding the same amino-acid i.**

# Relative adaptiveness of a codon

$$w_{ij} = RSCU_{ij}/RSCI_{imax} = X_{ij}/X_{imax}$$

where $RSCU_{imax}$ and $X_{imax}$ are RSCU and X values for the most frequently used codon for the ith amino acid.

# Codon Adaptation Index

The CAI for a gene is calculated as the geometric mean of the RSCU values corresponding to each of the codons used in that gene, divided by the maximum possible CAI for a gene of the same amino acid composition:

$$\mathrm{CAI} = \mathrm{CAI}_{\mathrm{obs}} / \mathrm{CAI}_{\mathrm{max}}$$

where $\mathrm{CAI}_{\mathrm{obs}} = (\pi \mathrm{RSCU}_k;\ k=1,L)^{1/L}$

$$\mathrm{CAI}_{\mathrm{max}} = (\pi \mathrm{RSCU}_{\mathrm{kmax}};\ k=1,L)^{1/L}$$

where $\mathrm{RSCU}_k$ is the RSCU value for the kth codon in the gene, $\mathrm{RSCU}_{\mathrm{kmax}}$ is the maximum RSCU value for the amino acid encoded by the kth codon in the gene, and L is the number of codons in the gene.

# Estimating synonymous and nonsynonymous differences

- **For a pair of homologous codons presenting only one nucleotide difference, the number of synonymous and nonsynonymous substitutions may be obtained by simple counting of silent versus non silent amino acid changes;**

- **For a pair of codons presenting more than one nucleotide difference, distinction between synonymous and nonsynonymous substitutions is not easy to calculate and statistical estimation methods are needed;**

- **For example, when there are 3 nucleotide differences between codons, there are 6 different possible pathways between these codons. In each path there are 3 mutational steps.**

- **More generally there can be many possible pathways between codons that differ at all three positions sites; each pathway has its own probability.**

# Estimating synonymous and nonsynonymous differences

- Observed nucleotide differences between 2 homologous sequences are classified into 4 categories: synonymous transitions, synonymous transversions, nonsynonymous transitions and nonsynonymous transversions.

- When the 2 compared codons differ at one position, the classification is obvious.

- When they differ at 2 or 3 positions, there will be 2 of 6 parsimonious pathways along which one codon could change into the other, and all of them should be considered.

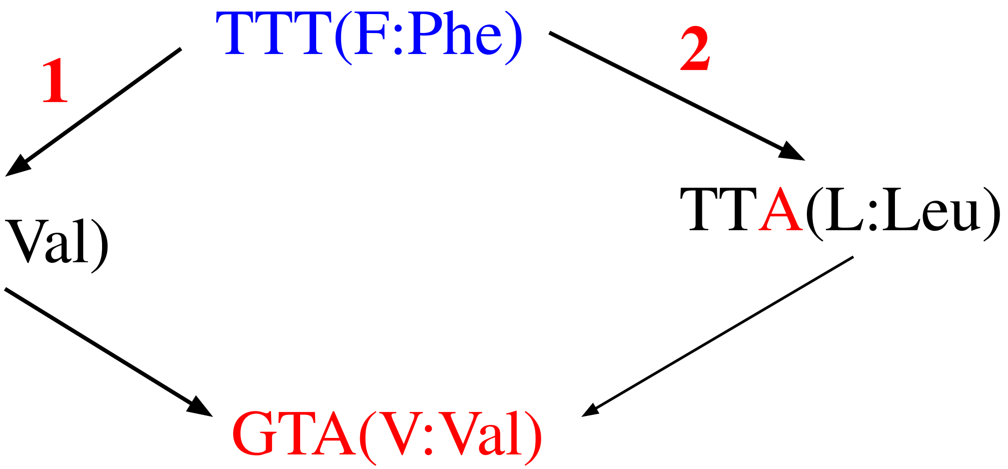
- Since different pathways may involve different numbers of synonymous and nonsynonymous changes, they should be weighted differently.

# Example: 2 homologous sequences

```
         Glu  Val  Phe
SEQ.1    GAA  GTT  TTT

SEQ.2    GAC  GTC  GTA
          Asp  Val Val
```

- Codon 1: GAA --> GAC ;1 nuc. diff., 1 nonsynonymous difference;
- Codon 2: GTT --> GTC ;1 nuc. diff., 1 synonymous difference;
- Codon 3: counting is less straightforward:



TTT(F:Phe)

1

2

GTT(V:Val)

TTA(L:Leu)

GTA(V:Val)

**Path 1 : implies 1 non-synonymous and 1 synonymous substitutions;**

**Path 2 : implies 2 non synonymous substitutions;**

# Evolutionary Distance estimation between 2 sequences

**The simplest problem is the estimation of the number of synonymous ($d_S$) and nonsynonymous ($d_N$) substitutions per site between 2 sequences:**

- **the number of synonymous (S) and nonsynonymous (N) sites in the sequences are counted;**
- **the number of synonymous and nonsynonymous differences between the 2 sequences are counted;**
- **a correction for multiple substitutions at the same site is applied to calculate the numbers of synonymous ($d_S$) and nonsynonymous ($d_N$) substitutions per site between the 2 sequences.**

**==> many estimation Methods**

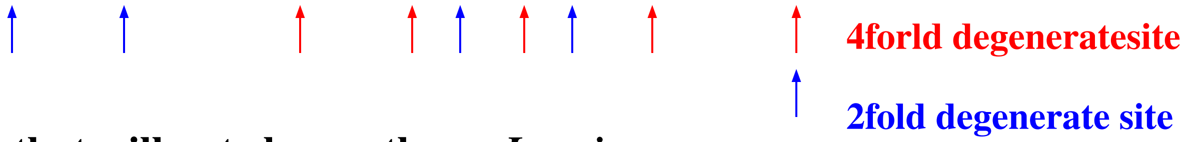
# Evolutionary Distance estimation

**In general the genetic code affords fewer opportunities for nonsynonymous changes than for synonymous changes.**

**rate of synonymous >> rate of nonsynonymous substitutions.**

**Furthermore, the likelihood of either type of mutation is highly dependent on amino acid composition.**

**For example: a protein containing a large number of *leucines* will contain many more opportunities for synonymous change than will a protein with a high number of *lysines*.**



**Several possible substitutions that will not change the aa *Leucine***

**Only one possible mutation at 3rd position that will not change *Lysine***

# Evolutionary Distance estimation

- Fundamental for the study of protein evolution and useful for constructing phylogenetic trees and estimation of divergence time.

# Estimating synonymous and nonsynonymous substitution rates

- **Ziheng Yang & Rasmus Nielsen (2000)**

**Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol.* 17:32-43.**

## Purifying selection:

Most of the time selection eliminates deleterious mutations, keeping the protein as it is.

## Positive selection:

In few instances we find that $d_N$ (also denoted $K_a$) is much greater than $d_S$ (also denoted $K_s$) (i.e. $d_N/d_S >> 1$ ($K_a/K_s >>1$ )). This is strong evidence that selection has acted to change the protein.

Positive selection was tested for by comparing the number of nonsynonymous substitutions per nonsynonymous site ($d_N$) to the number of synonymous substitutions per synonymous site ($d_S$). Because these numbers are normalized to the number of sites, if selection were neutral (i.e., as for a pseudogene) the $d_N/d_S$ ratio would be equal to 1. An unequivocal sign of positive selection is a $d_N/d_S$ ratio significantly exceeding 1, indicating a functional benefit to diversify the amino acid sequence.

$d_N/d_S < 0.25$ indicates purifying selection;

$d_N/d_S = 1$ suggests neutral evolution;

$d_N/d_S >> 1$ indicates positive selection.

**Negative (purifying) selection eliminates disadvantageous mutations i.e. inhibits protein evolution.**

**(explains why $d_N < d_S$ in most protein coding regions)**

**Positive selection is very important for evolution of new functions especially for duplicated genes.**

**(must occur early after duplication otherwise null mutations and will be fixed producing pseudogenes).**

- **$d_N/d_S$ (or $K_a/K_s$) measures selection pressure**

# Mutational saturation

**Mutational saturation in DNA and protein sequences occurs when sites have undergone multiple mutations causing sequence dissimilarity (the observed differences) to no longer accurately reflect the "true" evolutionary distance i.e. the number of substitutions that have actually occurred since the divergence of two sequences.**

**Correct estimation of the evolutionary distance is crucial.**

**Generally: sequences where $d_S > 2$ are excluded to avoid the saturation effect of nucleotide substitution.**

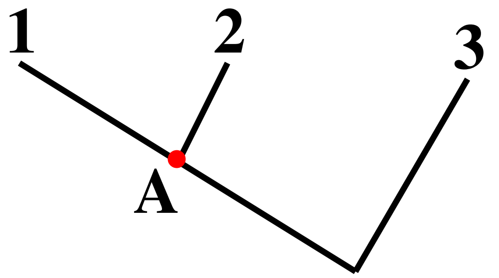
-> yn00   similar results than ML (**Yang & Nielsen (2000)**)

-> advantage : easy automation for large scale comparisons;

- **PAML: Phylogenetic Analysis by Maximum Likelihood (PAML)**

**http://abacus.gene.ucl.ac.uk/software/paml.html**

# Relative Rate Test



For determining the relative rate of substitution in species 1 and 2, we need and outgroup (species 3).

The point in time when 1 and 2 diverged is marked A (common ancestor of 1 and 2).

The number of substitutions between any two species is assumed to be the sum of the number of substitutions along the branches of the tree connecting them:

$$\begin{cases} d_{13}=d_{A1}+d_{A3} \\ d_{23}=d_{A2}+d_{A3} \\ d_{12}=d_{A1}+d_{A2} \end{cases}$$

$d_{13}$, $d_{23}$ and $d_{12}$ are measures of the differences between 1 and 3, 2 and 3 and 1 and 2 respectively.

$$\begin{cases} d_{A1}=(d_{12}+d_{13}-d_{23})/2 \\ d_{A2}=(d_{12}+d_{23}-d_{13})/2 \end{cases}$$

$d_{A1}$ and $d_{A2}$ should be the same (A common ancestor of 1 and 2).

# Reference

**Yang & Nielsen,**

**Esimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models**

***Mol. Biol. Evol*. 2000, 17:32-43**

**=>Other estimation Models**

# Evolutionary Distance estimation between 2 sequences

- **Under certain conditions, however, nonsynonymous substitution may be accelerated by positive Darwinian selection. It is therefore interesting to examine the number of synonymous differences per synonymous site and the number of nonsynonymous differences per nonsynonymous site.**

## p-distance:

- $p_s = S_d/S$ proportion of synonymous differences ;

$var(p_s) = p_s(1-p_s)/S.$

- $p_n = N_d/N$ proportion of non synonymous differences;

$var(p_n) = p_n(1-p_n)/S.$

$S_d$ and $N_d$ are respectively the total number of synonymous and non synonymous differences calculated over all codons. S and N are the numbers of synonymous and nonsynonymous substitutions.

S+N=n total number of nucleotides and N >> S.

$p_s$ is often denoted $K_s$ and $p_n$ is denoted $K_a$.

# Substitutions between protein sequences

$p = n_d/n$

$V(p)=p(1-p)/n$

$n_d$ and n are the number of amino acid differences and the total number of amino acids compared.

However, refining estimates of the number of substitutions that have occurred between the amino acid sequences of 2 or more proteins is generally more difficult than the equivalent task for coding sequences (see paths above).

One solution is to weight each amino acid substitution differently by using empirical data from a variety of different protein comparisons to generate a matrix as the PAM matrix for example.

# Number of synonymous ($d_s$) and non synonymous ($d_n$) substitutions per site

1) **Jukes and Cantor**, “one-parameter method” denoted “**1-p**” :

This model assumes that the rate of nucleotide substitution is the same for all pairs of the four nucleotides A, T, C and G (generally not true!).

**$d = -(3/4)_{*}Ln(1-(4/3)_{*}p)$ where p is either $p_s$ or $p_n$.**

2) **Kimura's 2-parameter, denoted “2-p” :**

The rate of transitional nucleotide substitution is often higher than that of transversional substitution.

**$d = -(1/2)_{*}Ln(1 -2_{*}P -Q) -(1/4)_{*}Log(1 -2_{*}Q)$**

P is the proportion of transitional differences,

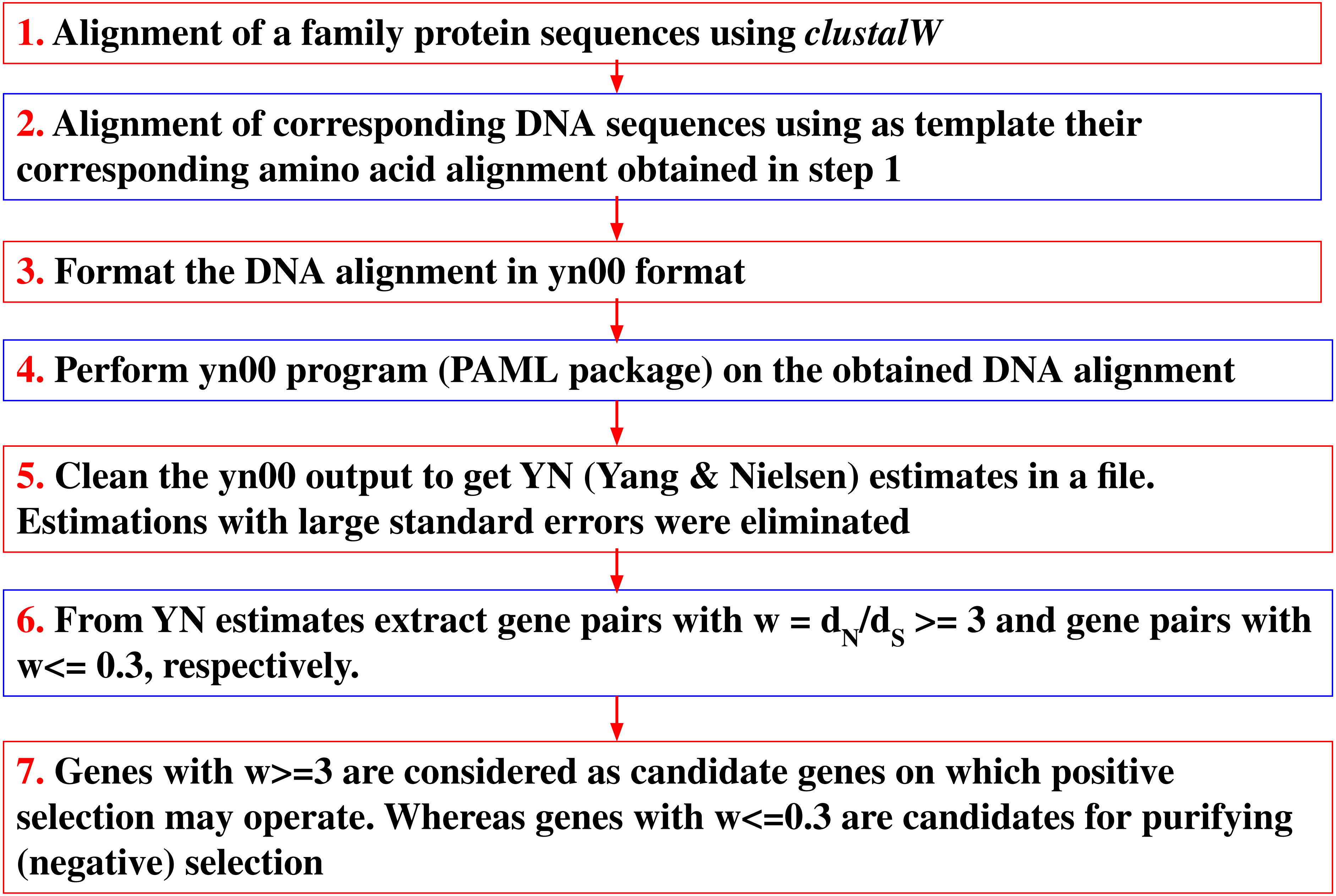
Q is the proportion of transversional differences

P and Q are respectively calculated over synonymous and non synonymous differences.

- **Example: yn00 in PAML.**

- **Protein sequences in a family and corresponding DNA sequences**

# Procedure



**1. Alignment of a family protein sequences using *clustalW***

**2. Alignment of corresponding DNA sequences using as template their corresponding amino acid alignment obtained in step 1**

**3. Format the DNA alignment in yn00 format**

**4. Perform yn00 program (PAML package) on the obtained DNA alignment**

**5. Clean the yn00 output to get YN (Yang & Nielsen) estimates in a file. Estimations with large standard errors were eliminated**

**6. From YN estimates extract gene pairs with w = $d_N/d_S$ >= 3 and gene pairs with w<= 0.3, respectively.**

**7. Genes with w>=3 are considered as candidate genes on which positive selection may operate. Whereas genes with w<=0.3 are candidates for purifying (negative) selection**

- **Most of the genes are under purifying selection**

- **Only few genes might be under positive selection**

- **Codon volatility**

# A new concept: codons volatility

(Plotkin et al. 2004. nature 428. p.942-945).

- **New method recently introduced, the utility of which is still under debate;**
- **has interresting consequences on the study of codon variability;**

# Detecting Selection

- If a protein coding region of a nucleotide sequence has undergone an excess number of amino-acid substitutions, then the region will on average contain an overabundance of "volatile" codons, compared with the genome as a whole.

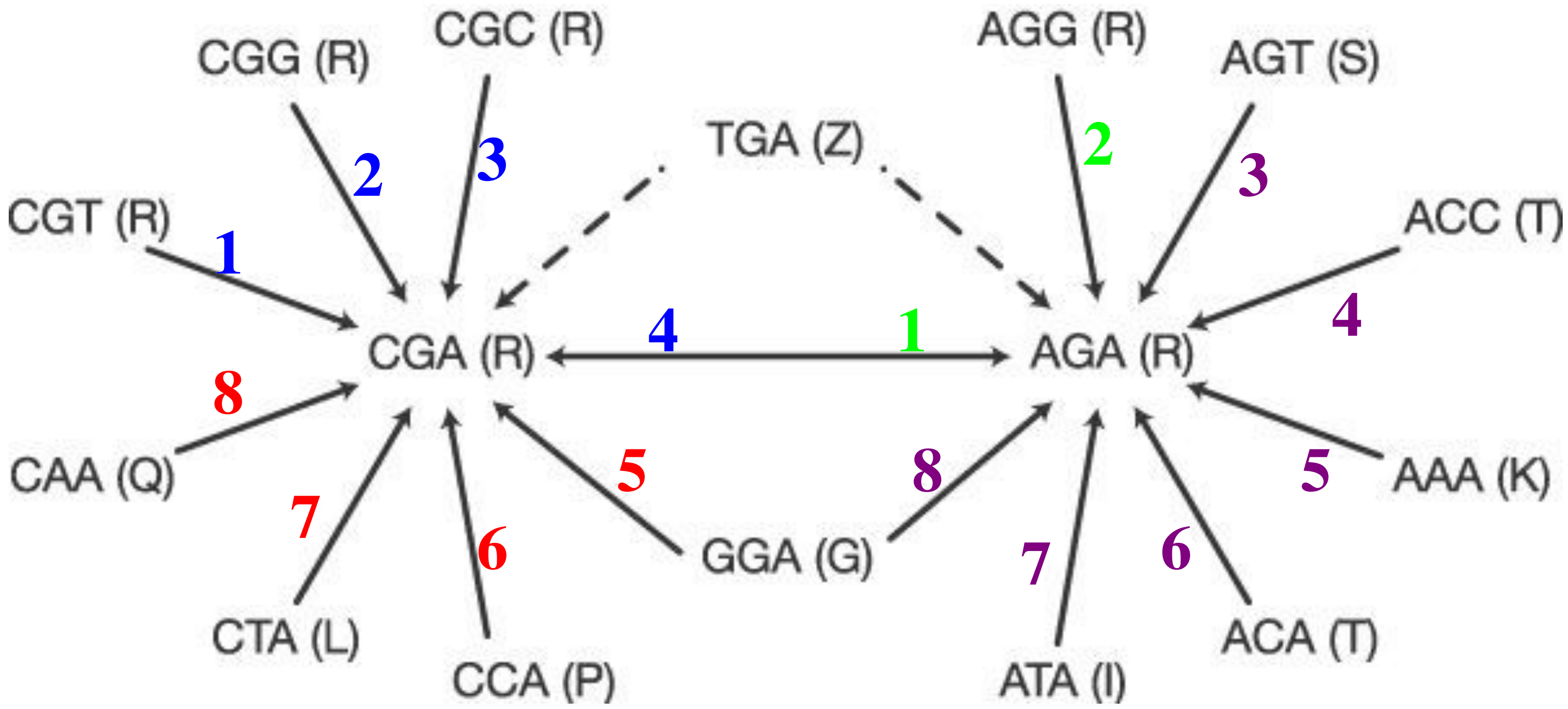
- Using the concept of codon volatility, we can scan an entire genome to find genes that show significantly more, or less, pressure for amino-acid substitutions than the genome as a whole.

- If a gene contains many residues under pressure for aa replacements, then the resulting codons in that gene will on average exhibit elevated volatility.

- If a gene is under purifying selection not to change its aa, then the resulting sequence will on average exhibit lower volatility.

Plotkin et al. *Nature* 428; 942-945

# Codons volatility



- The codon CGA encoding arginine (R), has 8 potential ancestor codons (i.e. non stop codon) that differ from CGA by one substitution.

- Volatility of a codon is defined as the proportion of nonsynonymous codons over the total neighbour sense codons obtained by a single substitution.

- The volatility of CGA = 4/8.

- The volatility of AGA also encodes an arginine = 6/8.

Plotkin et al. 2004.

*Nature* 428.
p.942-945

# Codons volatility

- 22 codons have at least one synonymous with a different volatility;

- Volatility of a codon c:

$v(c) = 1/n \sum\{D[aacid(c) - aacid(c_i)];i=1,n\};$

n is the number of neighbors (other than non-stop codons) that can mutate by a single substitution.

D is the Hamming distance = 0 if the 2 aa are identical;

=1 otherwise.

- Volatility of a gene G:

$v(G) = \sum\{v(c_k);k=1,l\};$ l is the number of codons in the gene G.

# Codons volatility

- **Volatility is used to quantify the probability that the most recent substitution of a site caused an amino-acid change.**
- **Each gene's observed volatility is compared with a bootstrap distribution of alternative synonymous sequences, drawn according to the background codon usage in the genome, and its significance statistically assessed.**
- **Randomization procedure controls for the gene's length and amino-acid composition.**
- **The volatility of a gene G is defined as the sum of the volatility of its codons.**

# Codons volatility

## Volatility p-value of G:

- The observed v(G) is compared with a bootstrap distribution of $10^6$ synonymous versions of the gene G.

- In each randomization sample, a nucleotide sequence G' is constructed so that it has the same translation as G but whose codons are drawn randomly according to the relative frequencies of synonymous codons in the whole genome.

- p-value for G = proportion of randomized samples;

  so that v(G') > v(G).

- 1-p is a p-value that tests whether a gene is significantly less volatile than the genome as a whole.

# Detecting Selection

- A p-value near zero indicates significantly elevated volatility, whereas a p-value near one indicates significantly depressed volatility.

- The probability that a site's most recent substitution caused a non-synonymous change is:

  - greater for a site under positive selection;
  - smaller for a site under negative (purifying) selection.

- http://www.cgr.harvard.edu/volatility

**1) Paul M. Sharp**
**Gene "volatility" is Most Unlikely to Reveal Adaptation**
*MBE* Advance Access published on December 22, 2004.
doi:10.1093/molbev/msi073

**2) Tal Dagan and Dan Graur**
**The Comparative Method Rules! Codon Volatility Cannot Detect Positive Darwinian Selection Using a Single Genome Sequence**
*MBE* Advance Access published on November 3, 2004.
doi:10.1093/molbev/msi033

**3) Robert Friedman and Austin L. Hughes**
**Codon Volatility as an Indicator of Positive Selection: Data**
*MBE* Advance Access originally published on November 3, 20
doi:10.1093/molbev/msi038

**4) Hahn MW, Mezey JG, Begun DJ, Gillespie JH, Kern AD**
**Evolutionary genomics: Codon bias and selection on single**
*Nature*. 2005 Jan 20;433(7023):E5-6.

**5) Nielsen R, Hubisz MJ.**
**Evolutionary genomics: Detecting selection needs compara**
*Nature*. 2005 Jan 20;433(7023):E6.

**6) Chen Y, Emerson JJ, Martin TM**
**Evolutionary genomics: Codon volatility does not detect se**
*Nature*. 2005 Jan 20;433(7023):E6-7.

**7) Zhang J, 2005.**
**On the evolution of codon volatility**
*Genetics* 169: 495-501.

**8) Plotkin JB, Dushoff J, Fraser HB.**
**Evolutionary genomics: Codon volatility does not detect se**
*Nature*. 2005 Jan 20;433(7023):E7-8.

**9) Plotkin JB, Dushoff J, Desai MM and Fraser HB**

**Synonymous codon and selection on proteins**

-> Volatility is not adequate for predicting selection;

-> Extreme volatility classes have interesting properties, in terms of aa composition or codon bias;

-> Volatility may be another measure of codon bias;

-> Authors : some genes are under more positive, or less negative, selection than others.

# Codon Volatility (simple substitution model):

Codons and volatility under simple substitution model

# References:

- Ziheng Yang and Rasmus Nielsen (2000)

Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models.
*Mol Biol Evol.* 17:32-43.

- Yang Z. and Bielawski J.P. (2000)

Statistical methods for detecting molecular adaptation

*Trends Ecol Evol.* 15:496-503.

- Phylogenetic Analysis by Maximum Likelihood (PAML)

http://abacus.gene.ucl.ac.uk/software/paml.html

- Plotkin JB, Dushoff J, Fraser HB (2004)

Detecting selection using a single genome sequence of M. tuberculosis and P. falciparum. *Nature* 428:942-5.

- Molecular Evolution; A phylogenetic Approach

Page, RDM and Holmes, EC (Blackwell Science, 2004)

- Sharp, PM & Li WH (1987). NAR 15:p.1281-1295.

# References

- Phylogeny programs :

http://evolution.genetics.washington.edu/phylip/sftware.html

- MEGA: http://www.megasoftware.net/
- PAML: http://abacus.gene.ucl.ac.uk/software/paml.html

## Books:

- Fundamental concepts of Bioinformatics.

Dan E. Krane and Michael L. Raymer

- Genomes 2 edition. T.A. Brown
- Molecular Evolution; A phylogenetic Approach

Page, RDM and Holmes, EC

Blackwell Science

# Molecular evolution: Definitions

## Purifying (negative) selection

- A consequence of gene “drift” through random mutations, is that many mutations will have deleterious effects on fitness.

- “Purifying selective force” prevents accumulation of mutation at important functional sites, resulting in sequence conservation.

-> “Purifying selection” is a natural selection against deleterious mutations.

-> The term is used interchangeably with “negative selection” or “selection constraints”.

# Neutral theory

- Majority of evolution at the molecular level is caused by random genetic "drift" through mutations that are selectively neutral or nearly neutral.
- Describes cases in which selection (purifying or positive) is not strong enough to outweigh random events.
- Neutral mutation is an ongoing process which gives rise to genetic polymorphisms; changes in environment can select for certain of these alleles.

# Positive selection

- Positive selection is a darwinian selection fixing advantageous mutations.

The term is used interchangeably with “molecular adaptation” and “adaptive molecular evolution”.

- Positive selection can be shown to play a role in some evolutionary events

- This is demonstrated at the molecular level if the rate of nonsynonymous mutation at a site is greater than the rate of synonymous mutation

- Most substitution rates are determined by either neutral evolution of purifying selection against deleterious mutations

# Molecular evolution

- **We observe and try to decode the process of molecular evolution from the perspective of accumulated differences among related genes from one or diverse organisms.**

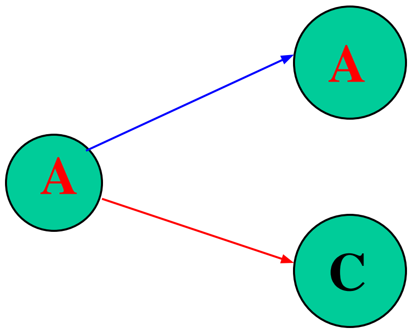
- **The number of mutations that have occurred can only be estimated.**

**Real individual events are blurred by a long history of changes.**

# Kinds of nucleotide substitutions

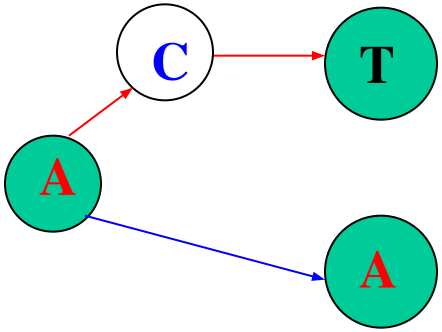
**Given 2 nucleotide sequences, we can ask how their similarities and differences arose from a common ancestor?**

### Single substitution



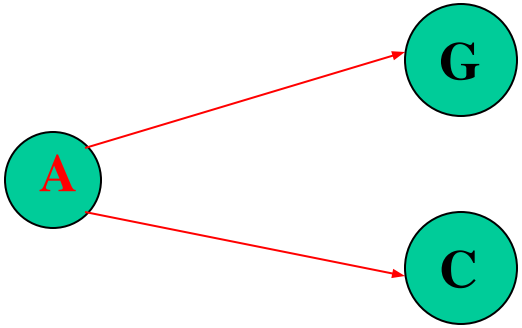
1 change, 1 difference

### Multiple substitution



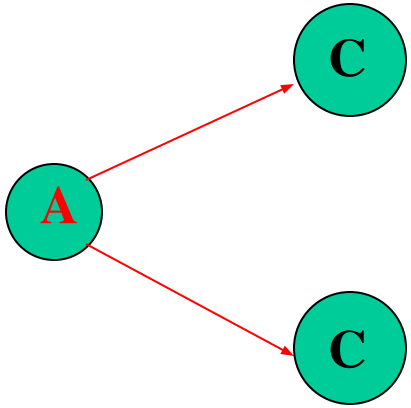
2 changes, 1 difference

### Coincidental substitution



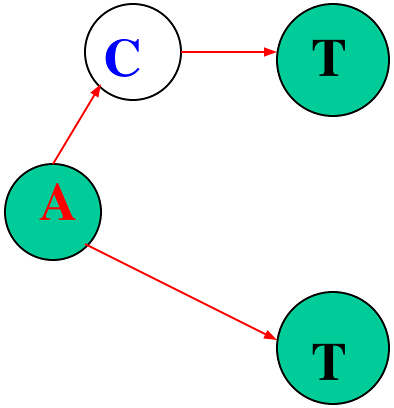
2 change, 1 difference

### Parallel substitution



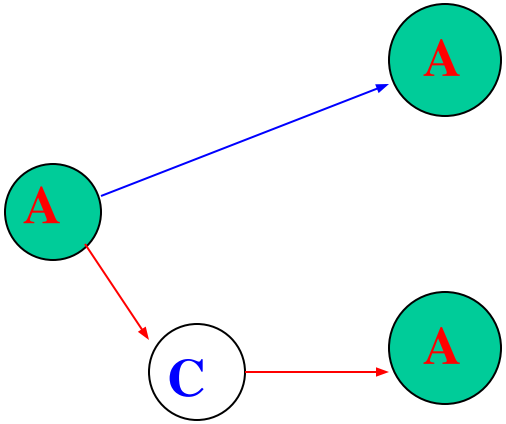
2 changes, no difference

### Convergent substitution



3 changes, no difference

### Back substitution



2 changes, no difference