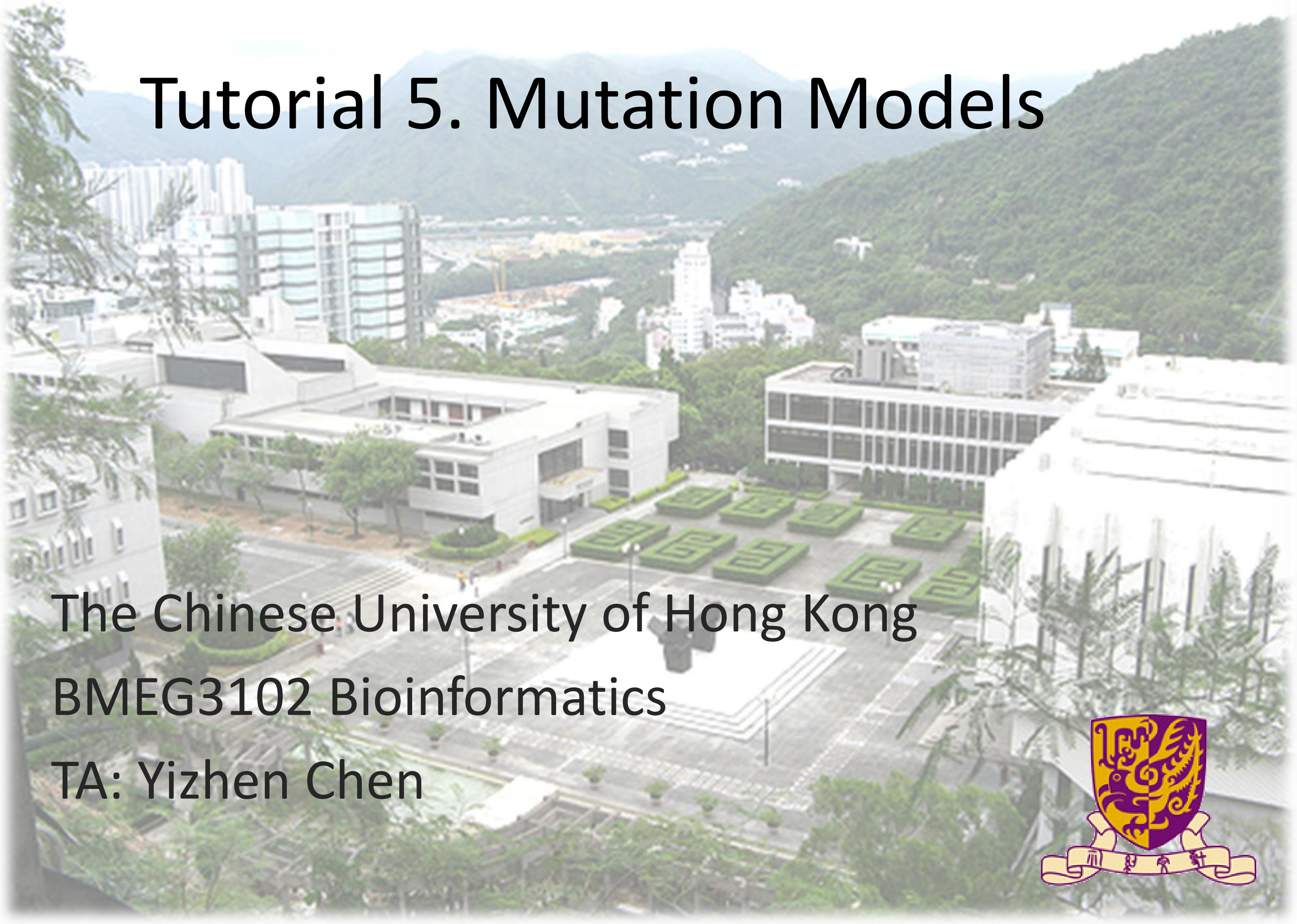
Tutorial 5. Mutation Models



The Chinese University of Hong Kong

BMEG3102 Bioinformatics

TA: Yizhen Chen

Agenda

- DNA Mutation Models
  - Jukes-Cantor Model
  - Kimura Two-parameter Model
- Protein Mutation Models
  - PAM
  - BLOSUM
- Phylogenetic Tree

Mutation

- Defined as the permanent change in the DNA
- Different types of mutation:
  - Small scale: affect one or several nucleotides
    - Point mutation: substitution, insertion, deletion
  - Large scale: usually affect the chromosomal structure
    - E.g., translocations, inversion

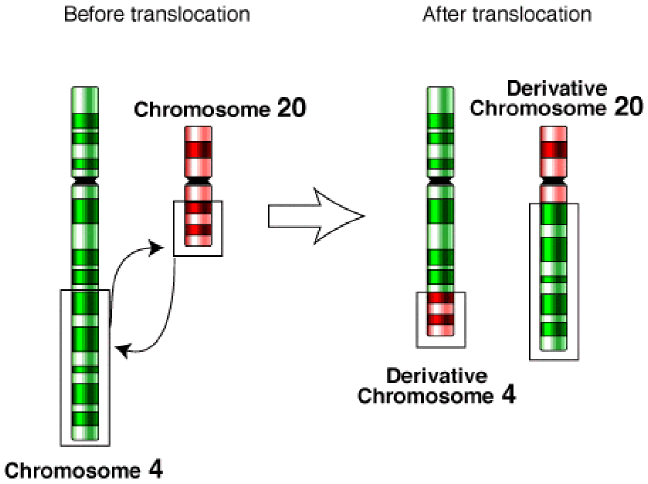
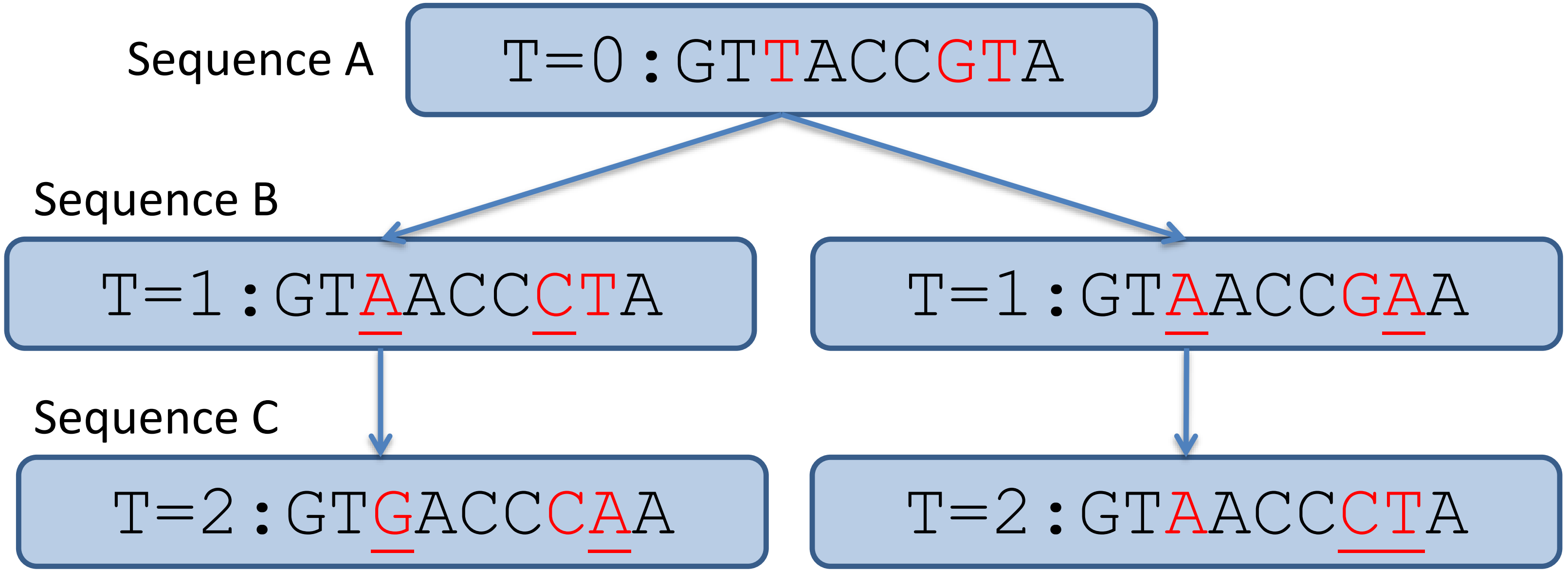


Image credit: wikipedia

Consequences of point mutations

- Occurs in the coding region of a gene:
  - Synonymous (silent) mutation – no change in protein sequence
  - Missense mutation – substitution of one amino acid for another
    - Conservative – function of protein is not affected
    - Non-conservative – function of protein altered
  - Nonsense mutation – substitution of one amino acid for a stop codon, leading to the truncated protein
- Occurs in the noncoding region:
  - Affect gene regulation, e.g., mutation in promoter or enhancer
  - Affect protein binding on DNA
  - Affect post-translational processing (e.g., defective splicing)

Introduction to Evolutionary Distance



Sequence A

T=0:GTTACCGTA

Sequence B

T=1:GTAACCCTA

T=1:GTAACCGAA

Sequence C

T=2:GTGACCCAA

T=2:GTAACCCTA

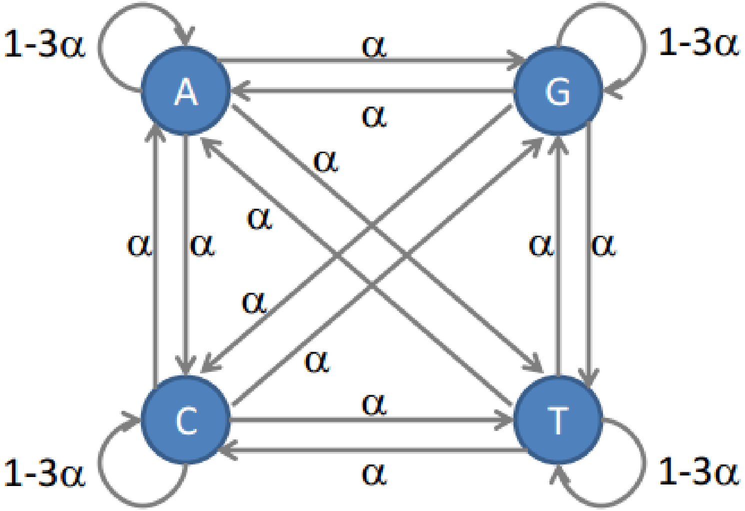
- At one nucleotide, you can see:
  - There are no mutations, e.g. G→G→G
  - There is exactly one mutation, e.g. G→C→C
  - There are two mutations in which the latest mutation overrides an old mutation, e.g. T→A→G
  - There are two mutations in which the latest mutation changes the nucleotide back to the ancestor sequence (unobservable mutation), e.g. T→A→T

Introduction to Evolutionary Distance

- Evolutionary Distance
  - The number of observed and unobserved base substitutions per site that have occurred since the divergence of two sequences
- Problem difficulties
  - Some mutations are not observable
  - We don't have ancestor sequence (in most cases)
  - We don't know how long the sequences have diverged
- We will study
  - Expectation of Evolutionary Distance
    - The expectation is larger than the number of observed differences between sequences
    - The sequences are usually observed sequences at present, not including their ancestor sequence
    - We try to infer the expected changes after the divergence from T=0 time point
  - Variance of Evolutionary Distance

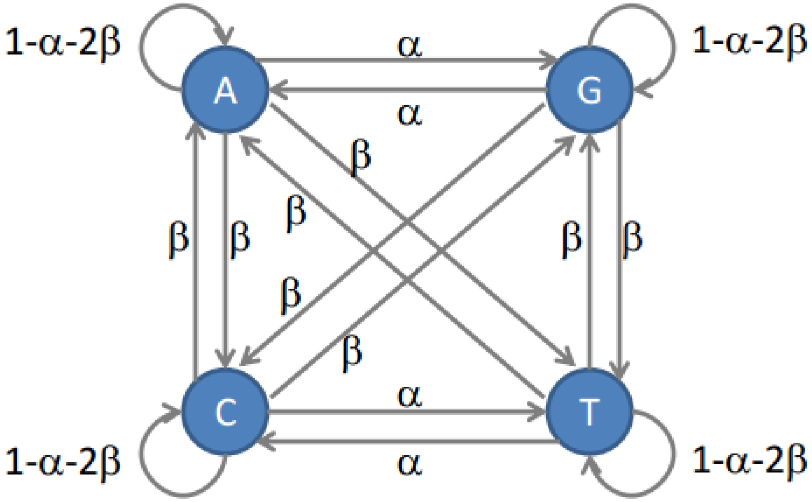
Jukes-Cantor model

- Parameter: rate of substitution α
- Notation: $\mathrm{P}_{\mathrm{X}\to\mathrm{Y}}(t)$ – the probability that for a base that was X at time 0, it is Y at time t for any X and Y
- Estimation formulas:
  - $P_{X\to X}(t) = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t}$
  - $P_{X\to Y}(t) = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t}$
  - $E[K_{\text{sup}}] = -\frac{3}{4}\ln(1 - \frac{4}{3}p_{\text{diff}})$
  - $Var[E[K_{\text{sup}}]] = \frac{p_{\text{diff}} - (p_{\text{diff}})^2}{n(1-\frac{4}{3}p_{\text{diff}})^2}$
  - $p_{\text{diff}} = \frac{x}{n} = \frac{\text{number of observed substitutions}}{\text{lenghth of sequence}}$



Kimura two-parameter model

- Parameters:
  - rate of transition α
  - rate of transversion β
  - β < α



- Estimation formulas:
  - $P_{X\to X}(t) = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} + \frac{1}{2}e^{-2(\alpha+\beta)t}$
  - $P_{\text{transition}}(t) = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} - \frac{1}{2}e^{-2(\alpha+\beta)t}$
  - $P_{\text{transversion}}(t) = \frac{1}{4} - \frac{1}{4}e^{-4\beta t}$

Kimura two-parameter model

- Estimation formulas (cont'):
  - $E\left[K_{\text{sup}}\right] = \frac{1}{2}\ln\left(\frac{1}{1-2p_{\text{diff1}}-p_{\text{diff2}}}\right) + \frac{1}{4}\ln\left(\frac{1}{1-2p_{\text{diff2}}}\right)$
  - $Var\left[E\left[K_{\text{sup}}\right]\right]$

$$= \frac{1}{n}\left[\mathrm{p}_{\text{diff 1}}\left(\frac{1}{1-2\mathrm{p}_{\text{diff 1}}-\mathrm{p}_{\text{diff 2}}}\right)^2 + \mathrm{p}_{\text{diff 2}}\left(\frac{1}{2-4\mathrm{p}_{\text{diff 1}}-2\mathrm{p}_{\text{diff 2}}}+\frac{1}{2-4\mathrm{p}_{\text{diff 2}}}\right)^2 - \left(\frac{\mathrm{p}_{\text{diff 1}}}{1-2\mathrm{p}_{\text{diff 1}}-\mathrm{p}_{\text{diff 2}}}+\frac{\mathrm{p}_{\text{diff 2}}}{2-4\mathrm{p}_{\text{diff 1}}-2\mathrm{p}_{\text{diff 2}}}+\frac{\mathrm{p}_{\text{diff 2}}}{2-4\mathrm{p}_{\text{diff 2}}}\right)^2\right]$$

  - $p_{\text{diff1}} = \frac{\text{number of transition}}{\text{lenghth sequence}}, p_{\text{diff2}} = \frac{\text{number of transverstion}}{\text{lenghth sequence}}$

Exercise 1: DNA mutation models

- You are given the current states of two sequences :
  - A: `GTGACCCAA`
  - B: `GTAACCCCA`

$$E[K_{\text{sup}}] = -\frac{3}{4}\ln(1 - \frac{4}{3}p_{\text{diff}})$$

$$E[K_{\text{sup}}] = \frac{1}{2}\ln\left(\frac{1}{1 - 2p_{\text{diff1}} - p_{\text{diff2}}}\right) + \frac{1}{4}\ln\left(\frac{1}{1 - 2p_{\text{diff2}}}\right)$$

- Using Jukes-Cantor model and Kimura model, respectively, calculate the followings:
  - Observed differences;
  - Probability of observing differences at a site;
  - Expected substitutions per site;
  - Expected substitutions in total.

Answer of Exercise 1

- For Jukes-Cantor model:
  - diff = 2
  - P_{diff} = 2/9
  - $E[K_{sup}]$ = 0.26355
  - Total = 0.26355 × 9 = 2.37195
- For Kimura model:
  - diff1 = 1, diff2 = 1
  - P_{diff1} = 1/9, P_{diff2} = 1/9
  - $E[K_{sup}]$ = 0.26556
  - Total = 0.26556 × 9 = 2.39004

Exercise 2: DNA mutation models

- Using both Jukes-Cantor model (rate of substitution = α) and Kimura model (rate of transition = α, rate of transversion = β) respectively, express the exact value for $P_{C \to G}(2)$ in terms of the parameter(s).
- Hint: there are four possible paths:
  - {C→A→G, C→T→G, C→C→G, C→G→G}

Answer of Exercise 2

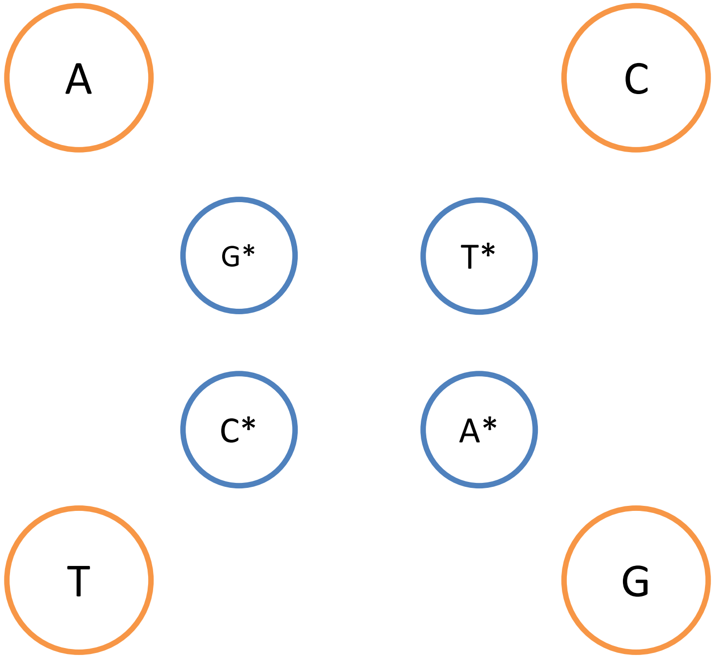
- Possible paths:
  - {C→A→G, C→T→G, C→C→G, C→G→G}
- For Jukes-Cantor model:
  - $P_{C\rightarrow G}(2) = \alpha^2 + \alpha^2 + (1-3\alpha)\alpha + \alpha(1-3\alpha)$

    $= 2\alpha - 4\alpha^2$
- For Kimura model:
  - $P_{C\rightarrow G}(2) = \beta\alpha + \alpha\beta + (1-\alpha-2\beta)\beta + \beta(1-\alpha-2\beta)$

    $= 2\beta - 4\beta^2$

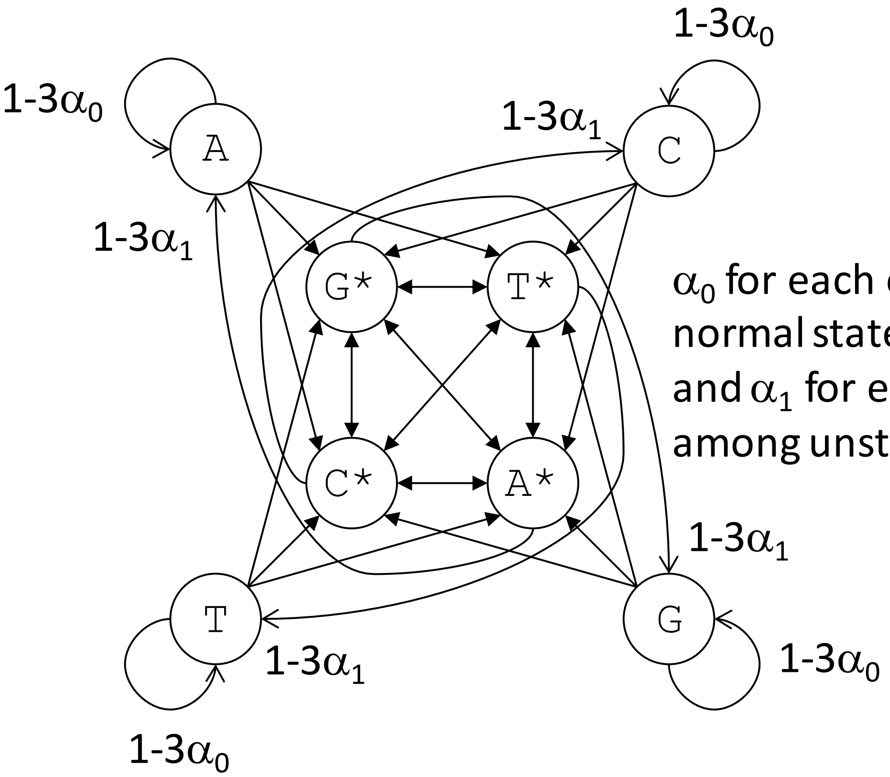
Exercise 3: DNA mutation models

- Suppose we want to develop a new mutation model for the following situation. A nucleotide can be in either a **normal state**, or an **unstable state**.
- In the normal state, in every time unit there is a probability of α_0 that it would mutate to each of the three other nucleotides, just like the Jukes-Cantor model, and enters the unstable state. Otherwise, it remains unmutated and stays in the normal state.
- When the nucleotide is in the unstable state, it has a probability of α_1 that it would mutate to each of the three other nucleotides and remains in the unstable state in every time unit, where $\alpha_1 > \alpha_0$. If it does not mutate to another nucleotide, it returns to the normal state.
- Draw the transition diagram for this mutation model. Explain any additional symbol you introduce in the diagram.



- A, C, G and T correspond to the normal states, while A*, C*, G* and T* correspond to the unstable states.

Answer to Exercise 3



α_0 for each of the 12 edges from a normal state to an unstable state, and α_1 for each of the 12 edges among unstable states

- A, C, G and T correspond to the normal states, while A*, C*, G* and T* correspond to the unstable states.

Substitution for amino acids

- Amino acid substitutions depend heavily on biochemical properties, and thus more difficult to model than DNA substitution
- Two common substitution matrices:
  - PAM (Point accepted mutation)
  - BLOSUM (Blocks of amino acid substitution matrix)

Comparison between PAM and BLOSUM

	PAM	BLOSUM
Definition	PAMx (x ≥ 1): mutation rate is x substitutions per 100 amino acid. $S_{ij} = P_{i\to j}$ (probability)	BLOSUMy (0 ≤ y ≤ 100): local alignment involving sequences more than y% identical. $S_{ij} = \frac{1}{\lambda} log \frac{p_{ij}}{p_i p_j}$ (log-odd score)
Similarities	• Substitution matrix of amino acid (with dimension of 20 x 20) • Obtained from taking sets of high-confidence alignments of many homologous proteins and assessing the frequencies of all substitutions.	
Differences	1. Each entry corresponds to a probability of substitution 2. Asymmetric 3. Based on global alignment of closely related proteins 4. Larger x implies larger evolutionary distance	1. Each entry correspond to a log-odd of the observed substitutions and expectation 2. Symmetric 3. Based on local alignment of highly conserved regions of proteins 4. Larger y implies smaller evolutionary distance

More on PAM and BLOSUM

- If two protein sequences are very similar, for each pair of matrices below, which one will you choose?
  - [✓ PAM100], PAM150
  - BLOSUM50, [✓ BLOSUM62]

- From biological aspect, what factors will contribute to the different values in PAM, or BLOSUM matrices?
  - Size, charge, hydrophobicity, etc

More on PAM and BLOSUM

- If we know PAM1, how to compute PAMx from PAM1?
  - $\mathrm{PAMx} = (\mathrm{PAM1})^{\mathrm{x}}$ (matrix multiplication)

- In BLOSUM, there are zeros, positive and negative numbers, what do they mean?
  - Zero: Two amino acids have no preference for or against substituting to the other
  - Positive: Two amino acids are more similar, and the alignment of them is found more often in the database
  - Negative: Two amino acids are less similar, and the alignment of them is found less often in the database

Exercise 4: PAM and BLOSUM

- To simplify the question, here we use DNA sequence instead of protein sequence.
- Suppose we have a database containing some similar DNA sequences (diverged from a common ancestor) and performed optimal local sequence alignment. The counts for all kinds of matches/mismatches are summarized as follows (alignments of reversed complementary strands are also considered):

"X-Y" means you see "X" aligned with "Y" in the alignment

Match/mismatch	Number of sites
C-C, G-G	250
A-A, T-T	150
A-G, C-T	50
A-C, A-T, C-G, G-T	25

This means we have 250 C-C and 250 G-G

Exercise 4: PAM and BLOSUM

a) Suppose on average there are x substitutions in 100 nucleotides at this moment. Construct the substitution matrix PAM*x* that summarizes the observed substitutions between the DNA sequence pairs.

b) Suppose there is a given period of time equivalent to the time having 2x mutations per 100 nucleotides. Construct the substitution matrix PAM2*x.*

c) [optional] Suppose there is a given period of time equivalent to the time having 0.5x mutations per 100 nucleotides. Construct the substitution matrix PAM0.5*x.*

d) Construct BLOSUM, with scale factor λ = 1.

Answer of Exercise 4(a)

Match/mismatch	Number of sites
C-C, G-G	250
A-A, T-T	150
A-G, C-T	50
A-C, A-T, C-G, G-T	25

$$\boxed{S_{ij} = P_{i \to j}}$$

	A	C	G	T
A	$\frac{2(150)}{2(150)+50+25+25} = \frac{3}{4}$	$\frac{25}{2(150)+50+25+25} = \frac{1}{16}$	$\frac{50}{2(150)+50+25+25} = \frac{1}{8}$	$\frac{25}{2(150)+50+25+25} = \frac{1}{16}$
C	$\frac{25}{2(250)+50+25+25} = \frac{1}{24}$	$\frac{2(250)}{2(250)+50+25+25} = \frac{5}{6}$	$\frac{25}{2(250)+50+25+25} = \frac{1}{24}$	$\frac{50}{2(250)+50+25+25} = \frac{1}{12}$
G	$\frac{50}{2(250)+50+25+25} = \frac{1}{12}$	$\frac{25}{2(250)+50+25+25} = \frac{1}{24}$	$\frac{2(250)}{2(250)+50+25+25} = \frac{5}{6}$	$\frac{25}{2(250)+50+25+25} = \frac{1}{24}$
T	$\frac{25}{2(150)+50+25+25} = \frac{1}{16}$	$\frac{50}{2(150)+50+25+25} = \frac{1}{8}$	$\frac{25}{2(150)+50+25+25} = \frac{1}{16}$	$\frac{2(150)}{2(150)+50+25+25} = \frac{3}{4}$

Answer of Exercise 4(b)

$$\mathrm{PAM}x = \begin{bmatrix} \frac{3}{4} & \frac{1}{16} & \frac{1}{8} & \frac{1}{16} \\ \frac{1}{24} & \frac{5}{6} & \frac{1}{24} & \frac{1}{12} \\ \frac{1}{12} & \frac{1}{24} & \frac{5}{6} & \frac{1}{24} \\ \frac{1}{16} & \frac{1}{8} & \frac{1}{16} & \frac{3}{4} \end{bmatrix}$$

$$\mathrm{PAM2x} = \mathrm{PAM}x \times \mathrm{PAM}x = \begin{bmatrix} \frac{3}{4} & \frac{1}{16} & \frac{1}{8} & \frac{1}{16} \\ \frac{1}{24} & \frac{5}{6} & \frac{1}{24} & \frac{1}{12} \\ \frac{1}{12} & \frac{1}{24} & \frac{5}{6} & \frac{1}{24} \\ \frac{1}{16} & \frac{1}{8} & \frac{1}{16} & \frac{3}{4} \end{bmatrix}^2 = \begin{bmatrix} \frac{445}{768} & \frac{43}{384} & \frac{157}{768} & \frac{5}{48} \\ \frac{43}{576} & \frac{817}{1152} & \frac{23}{288} & \frac{157}{1152} \\ \frac{157}{1152} & \frac{23}{288} & \frac{817}{1152} & \frac{43}{576} \\ \frac{5}{48} & \frac{157}{768} & \frac{43}{384} & \frac{445}{768} \end{bmatrix}$$

Ideas of Exercise 4(c) [optional]

- Method 1: Solving system of equations
  - Let PAM0.5*x* be a 4×4 matrix of 16 unknowns
  - We know that PAM0.5*x* × PAM0.5*x* = PAM*x*, then, we will get a system of 16 quadratic equations of 16 unknowns
  - After solving the equations, we will find the PAM0.5*x*
- Method 2: Diagonalize the PAM*x* matrix
  - We diagonalize PAM*x* into QDQ^{-1}, where D is a diagonal matrix, Q is a collection of eigenvectors of PAM*x*
  - PAM0.5*x* = $QD^{1/2}Q^{-1}$

Answer of Exercise 4(d)

Match/mismatch	Number of sites
C-C, G-G	250
A-A, T-T	150
A-G, C-T	50
A-C, A-T, C-G, G-T	25

$$\boxed{S_{ij} = \frac{1}{\lambda} log \frac{p_{ij}}{p_i p_j}}$$

	A	C	G	T
A	$\log \frac{150/1000}{200/1000 \times 200/1000} = 1.9069$	$\log \frac{25/1000}{200/1000 \times 300/1000} = -1.2630$	$\log \frac{50/1000}{200/1000 \times 300/1000} = -0.2630$	$\log \frac{25/1000}{200/1000 \times 200/1000} = -0.6781$
C	$\log \frac{25/1000}{300/1000 \times 200/1000} = -1.2630$	$\log \frac{250/1000}{300/1000 \times 300/1000} = 1.4739$	$\log \frac{25/1000}{300/1000 \times 300/1000} = -1.8480$	$\log \frac{50/1000}{300/1000 \times 200/1000} = -0.2630$
G	$\log \frac{50/1000}{300/1000 \times 200/1000} = -0.2630$	$\log \frac{25/1000}{300/1000 \times 300/1000} = -1.8480$	$\log \frac{250/1000}{300/1000 \times 300/1000} = 1.4739$	$\log \frac{25/1000}{300/1000 \times 200/1000} = -1.2630$
T	$\log \frac{25/1000}{200/1000 \times 200/1000} = -0.6781$	$\log \frac{50/1000}{200/1000 \times 300/1000} = -0.2630$	$\log \frac{25/1000}{200/1000 \times 300/1000} = -1.2630$	$\log \frac{150/1000}{200/1000 \times 200/1000} = 1.9069$

Introduction to phylogenetic tree

- Given a set of DNA/protein sequences
- Construct a phylogenetic tree such that it presents the historical evolutionary events based on observable sequences

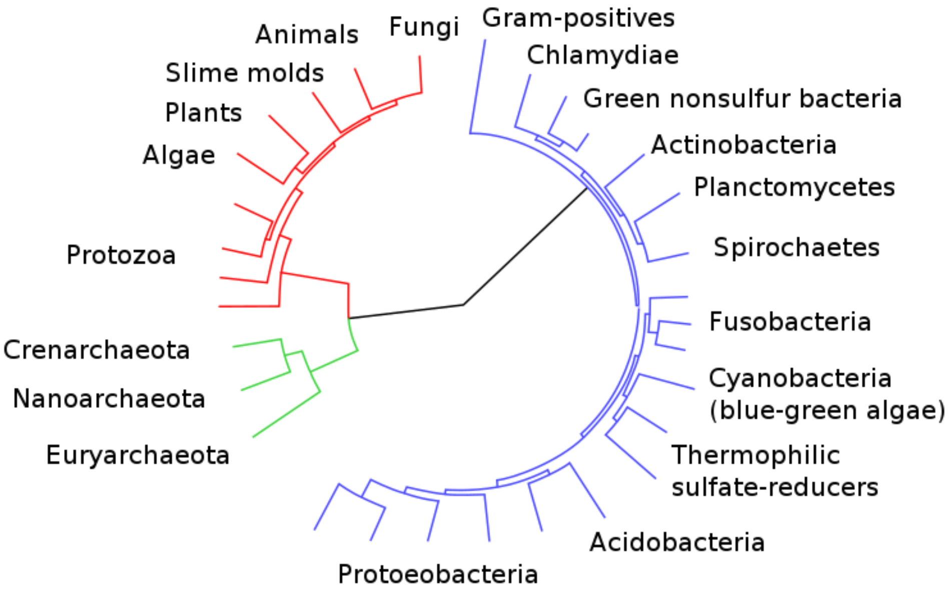
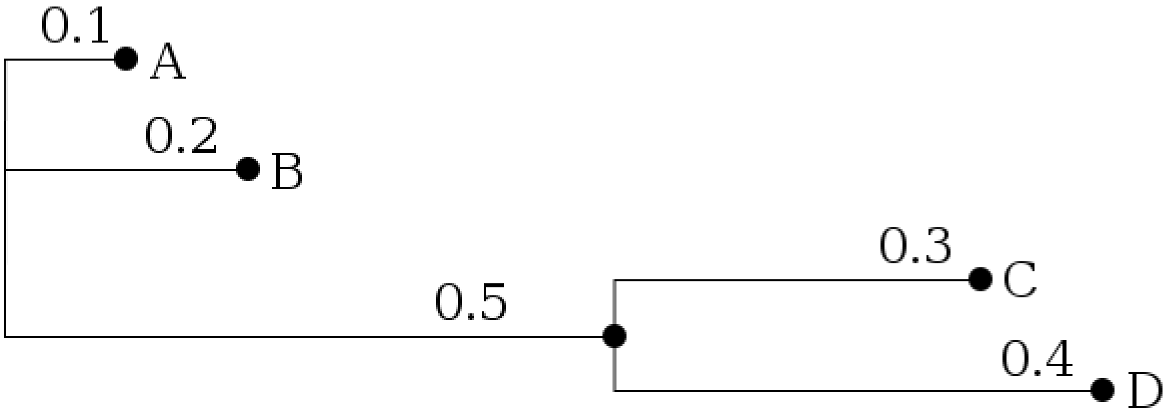


Image credit: wikipedia

Newick file format

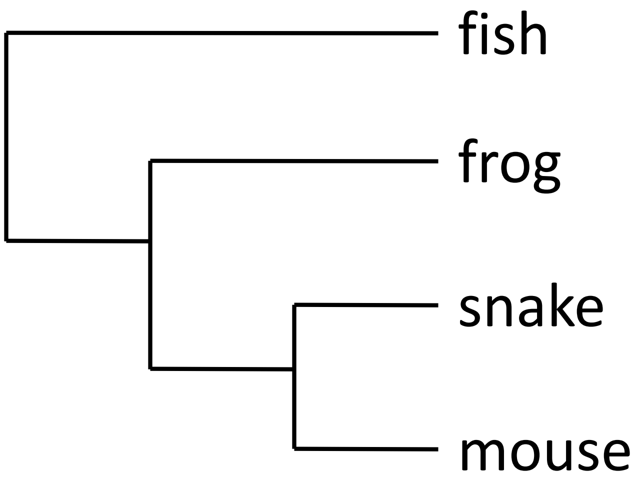
- Nested brackets with distance
- E.g.,



(A:0.1,B:0.2,(C:0.3,D:0.4):0.5);

NEXUS file format

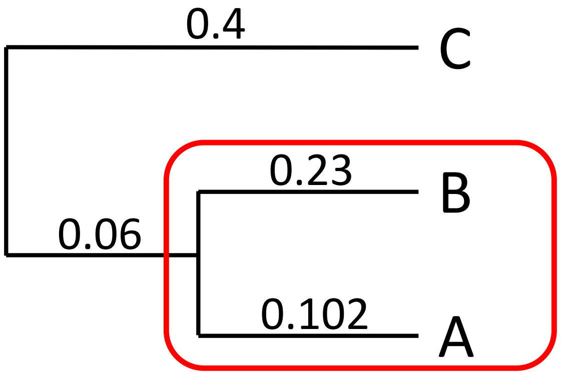
- Giving short IDs to sequences, with more metadata



```
#NEXUS
BEGIN TAXA;                                             TAXA Block
        DIMENSIONS NTAX = 4;
        TAXLABELS fish frog snake mouse;
END;
BEGIN CHARACTERS;                                       CHARACTERS Block
        DIMENSIONS NCHAR = 20;
        FORMAT DATATYPE = DNA;
        MATRIX                                          DATA Block
                fish    ACATA GACCG TACCT CTAAG
                frog    ACTTA GACCC TACCT CTACG
                snake   ACTCA CTGGG TACCT TTGCG
                mouse   ACTCA GACGG TACCT TTGCG
END;
BEGIN TREES;                                            TREES Block
        TREE BEST = ((fish, (frog, (snake, mouse)));
END;
```

For more information:
http://hydrodictyon.eeb.uconn.edu/eebedia/index.php/Phylogenetics:_NEXUS_Format

PhyloXML file format



```
<phyloxml>
        <phylogeny rooted=“false”>
                <name>example</name>
                <description>this is in phyloXML</description>
                <clade>
                        <clade branch_length="0.06">
                                <clade branch_length="0.102">
                                        <name>A</name>
                                </clade>
                                <clade branch_length="0.23">
                                        <name>B</name>
                                </clade>
                        </clade>
                        <clade branch_length="0.4">
                                <name>C</name>
                        </clade>
                </clade>
        </phylogeny>
</phyloxml>
```

Check list

- What are the assumptions for the Jukes-Cantor model and Kimura model?
- How can we estimate the expected number of substitutions based on current observations?
- What are the similarities and differences between PAM and BLOSUM?
- What does each entry of PAM and BLOSUM represent?
- What are the meanings of the numbers in PAM150 and BLOSUM62?
- How can we represent a phylogenetic tree?