BMEG3102 Bioinformatics

Lecture 4. Mutation Models and Molecular Phylogenetics (1/2)

Qi Dou

Email: qidou@cuhk.edu.hk

Office: Room 1014, 10/F, SHB

BMEG3102 Bioinformatics

The Chinese University of Hong Kong

Lecture outline

1. Evolutionary distance and mutation models

2. Substitution matrices for amino acids

3. Trees: Hierarchical structures relating different biological objects

- File formats

Part 1

Evolutionary Distance and Mutation Models

Evolutionary distance

- Suppose we have an alignment of two sequences.

  At a site, one sequence has a A and one has a C.

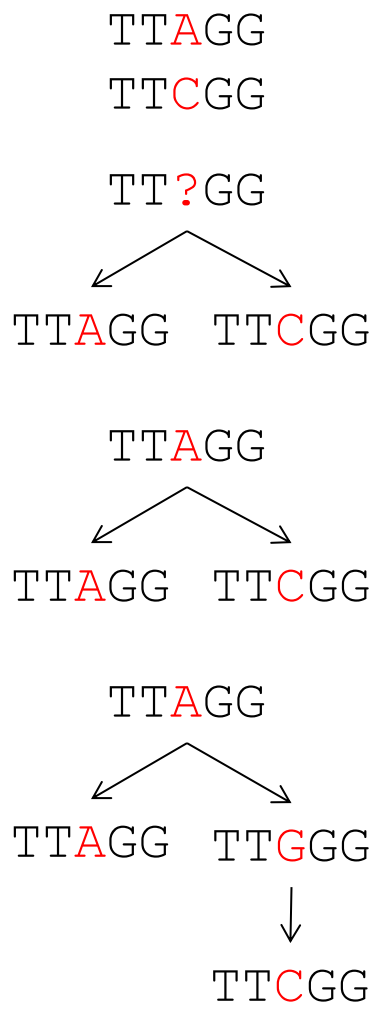
  – Assume that the sequences have a common ancestor

  – What did the common ancestor have at that site?

  – We don't know.

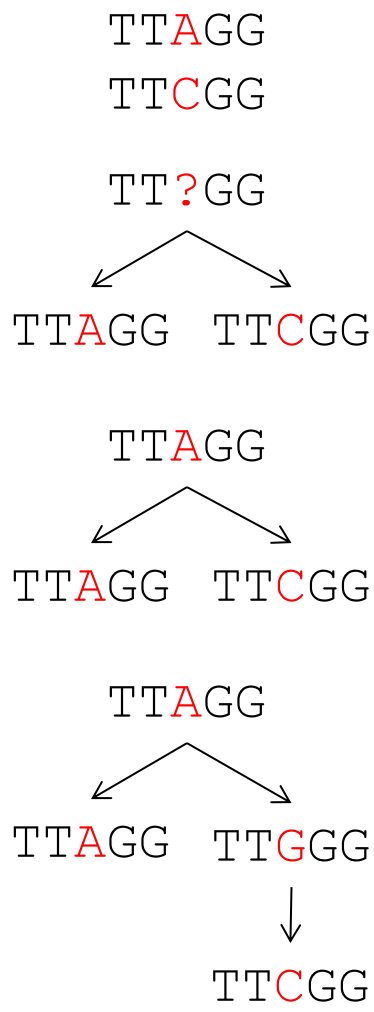
  – Let's say A. How many mutations have happened?

    - Could be one (A → C)
    - Could be more (A → G → C, A → T → C, A → C → A, etc.)



Evolutionary distance

- Suppose we have an alignment of two sequences. At a site, one sequence has a A and one has a C.
  - Assume that the sequences have a common ancestor
  - What did the common ancestor have at that site?
  - We don't know.
  - Let's say A. How many mutations have happened?
    - Could be one (A → C)
    - Could be more (A → G → C, A → T → C, A → C → A, etc.)
- We want a way to define the "evolutionary distance" between two observed sequences
  - According to the number of mutations happened or the time since their divergence
  - We need to first define a mutation model



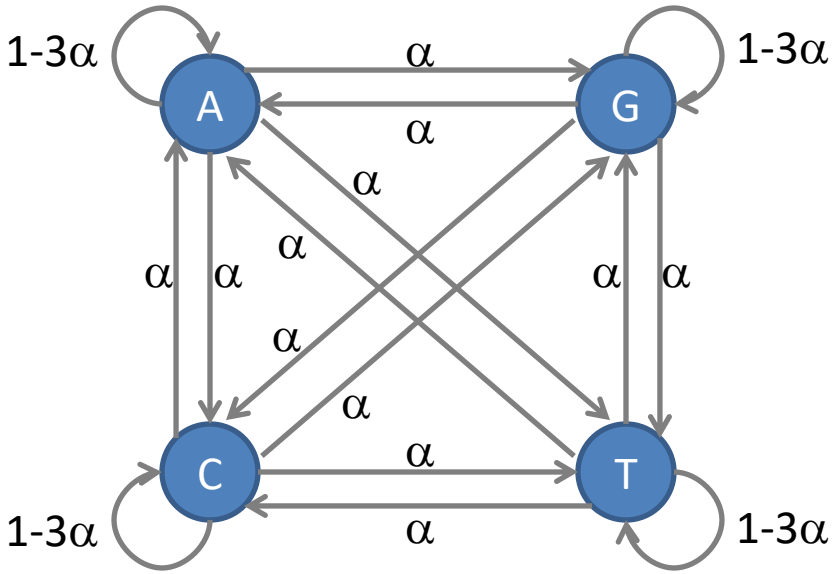
Mutation model

- A mutation model is a probabilistic model that describes how mutations happen over time
  - How often a mutation happens
  - What kinds of mutation are more frequent
- To make things simple, we will make the following assumptions:
  - Sites are independent
  - Mutation rates are the same for different sites and at different time in the history
  - Given current state, future states do not depend on past states
- We know these assumptions are usually not true, but without them the calculations can be difficult
  - More complex models that require fewer strong assumptions exist.

    We only study the simple models here

The Jukes-Cantor model

- Proposed by Jukes and Cantor in 1969
- Equal rate of substitution, α, to the other three bases **in one unit of time**
  - Assume there is at most one mutation within one unit of time – We can always make the unit smaller to ensure this



Notes: In this lecture,
1. We do not consider indels
2. A "substitution" is a point mutation actually happened, while a mismatch in an alignment could be caused by one or more substitutions

Illustration of the Jukes-Cantor model

- Suppose at time 0, site 1 of a sequence was A
- At time 1:
  - There is a probability of $1-3\alpha$ that the site was A
  - There is a probability of α that the site was C
  - There is a probability of α that the site was G
  - There is a probability of α that the site was T
- At time 2, what is the probability that the site is A, if we only know it was A at time 0 but do not know what it was at time 1?
  - Two possibilities:
    1. At time 1, the site was A, and there was no mutation from time 1 to time 2 [probability: $(1-3\alpha)^2$]
    2. At time 1, the site was C, G or T, and there was a mutation to A from time 1 to time 2 [probability: $3\alpha^2$]
  - Therefore, the total probability that the site is A at time 2 is $(1-3\alpha)^2 + 3\alpha^2$

Recursive formulas

- Denote $P_{X \to Y}(t)$ as the probability that for a base that was X at time 0, it is Y at time t for any X and Y
  - Here "site", "base", "nucleotide" all mean the same thing
  - $P_{A \to A}(1) = 1 - 3\alpha$
  - $P_{A \to A}(2) = (1 - 3\alpha)^2 + 3\alpha^2$
  - In general,
    $P_{A \to A}(t+1) = (1 - 3\alpha)P_{A \to A}(t) + \alpha[1 - P_{A \to A}(t)]$
  - Similarly:
    - $P_{X \to X}(t+1) = (1 - 3\alpha)P_{X \to X}(t) + \alpha[1 - P_{X \to X}(t)]$ for any X
    - $P_{X \to Y}(t+1) = [1 - P_{X \to X}(t+1)] / 3$ for any X and Y

Recursive formulas

- Denote $P_{X \to Y}(t)$ as the probability that for a base that was X at time 0, it is Y at time t for any X and Y
  - Here "site", "base", "nucleotide" all mean the same thing
  - $P_{A \to A}(1) = 1 - 3\alpha$
  - $P_{A \to A}(2) = (1 - 3\alpha)^2 + 3\alpha^2$
  - In general,
    $P_{A \to A}(t+1) = (1 - 3\alpha)P_{A \to A}(t) + \alpha[1 - P_{A \to A}(t)]$
  - Similarly:
    - $P_{X \to X}(t+1) = (1 - 3\alpha)P_{X \to X}(t) + \alpha[1 - P_{X \to X}(t)]$ for any X
    - $P_{X \to Y}(t+1) = [1 - P_{X \to X}(t+1)] / 3$ for any X and Y

  - We first study how to compute $P_{X \to Y}(t)$ for given mutation rate α and divergence time t, and then study how we can use $P_{X \to Y}(t)$ to estimate the number of mutations that have happened since the divergence of the two sequences

Solving $P_{A \to A}(t)$

- $P_{A \to A}(t+1) = (1 - 3\alpha)P_{A \to A}(t) + \alpha[1 - P_{A \to A}(t)]$
- $\Delta P_{A \to A}(t)$
  $\equiv P_{A \to A}(t+1) - P_{A \to A}(t)$ (here $\equiv$ means "is defined as")
  $= (1 - 3\alpha)P_{A \to A}(t) + \alpha[1 - P_{A \to A}(t)] - P_{A \to A}(t)$
  $= \alpha[1 - 4\, P_{A \to A}(t)]$

Solving $P_{A \to A}(t)$

- $P_{A \to A}(t+1) = (1 - 3\alpha)P_{A \to A}(t) + \alpha[1 - P_{A \to A}(t)]$
- $\Delta P_{A \to A}(t)$
  $\equiv P_{A \to A}(t+1) - P_{A \to A}(t)$ (here $\equiv$ means "is defined as")
  $= (1 - 3\alpha)P_{A \to A}(t) + \alpha[1 - P_{A \to A}(t)] - P_{A \to A}(t)$
  $= \alpha[1 - 4\,P_{A \to A}(t)]$

- For an infinitesimally small time unit, we get a first-order differential equation, which can be solved by using an integrating factor

$$\frac{dP_{A \to A}(t)}{dt} = \alpha[1 - 4P_{A \to A}(t)]$$

$$\Rightarrow \quad P_{A \to A}(t) = \frac{1}{4} + \left(P_{A \to A}(0) - \frac{1}{4}\right) e^{-4\alpha t}$$

$$= \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}$$

  – Observation: When t is large, the initial state (i.e., nucleotide) does not matter any more and all four bases are equally likely

Final formulas for $P_{X \to Y}(t)$

- By symmetry,

$$P_{A \to A}(t) = P_{C \to C}(t) = P_{G \to G}(t) = P_{T \to T}(t) = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}$$

- Similarly, it is easy to show that

$$P_{A \to C}(t) = P_{A \to G}(t) = P_{A \to T}(t) = \cdots = P_{T \to G}(t) = \left[1 - \left(\frac{1}{4} + \frac{3}{4} e^{-4\alpha t}\right)\right] / 3 = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t}$$

- You don't need to memorize these formulas
  - But you do need to know what the probabilities mean and how to apply them in calculations

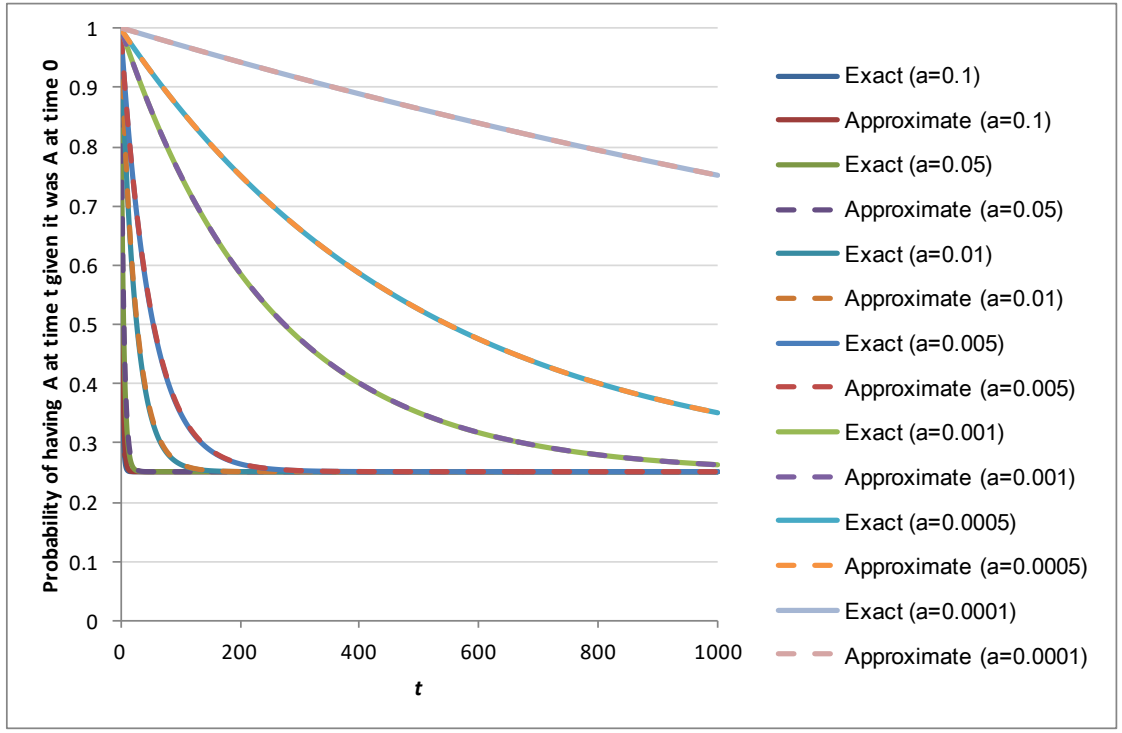
An alternative formula [optional]

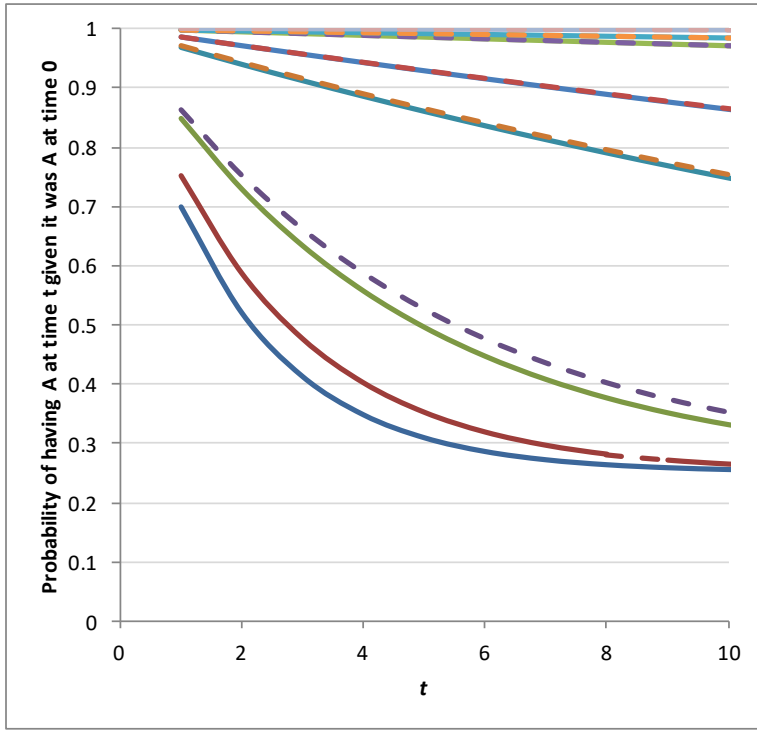
- Another way to solve the equation without using differential equations, suggested by a former student taking this class, Cao Jianquan:

$$
\begin{aligned}
& P_{A\to A}(t) && = (1 - 3\alpha)P_{A\to A}(t\text{-}1) + \alpha[1 - P_{A\to A}(t\text{-}1)] \\
& && = (1 - 4\alpha)P_{A\to A}(t\text{-}1) + \alpha \\
\Rightarrow\ & \underline{P_{A\to A}(t) - 1/4} && = (1 - 4\alpha)P_{A\to A}(t\text{-}1) + \alpha - 1/4 \\
& && = (1 - 4\alpha)P_{A\to A}(t\text{-}1) - 1/4\,(1 - 4\alpha) \\
& && = (1 - 4\alpha)(\underline{P_{A\to A}(t\text{-}1) - 1/4}) \\
& && = (1 - 4\alpha)^2(P_{A\to A}(t\text{-}2) - 1/4) \\
& && = \ldots \\
& && = (1 - 4\alpha)^{t-1}(P_{A\to A}(1) - 1/4) \\
& && = (1 - 4\alpha)^{t-1}(1 - 3\alpha - 1/4) \\
& && = (1 - 4\alpha)^{t-1}(3/4 - 3\alpha) \\
& && = 3/4\,(1 - 4\alpha)^t \\
\Rightarrow\ & P_{A\to A}(t) && = 3/4\,(1 - 4\alpha)^t + 1/4
\end{aligned}
$$

Key: The underlined parts have exactly the same form

Comparing the formulas [optional]





They differ significantly only when α ("a" in the figure) is large, and only in the first few time points

Comparing the formulas [optional]

- The exact solution on the last page:
  - $P_{A \to A}(t) = 3/4\ (1 - 4\alpha)^t + 1/4$
- The approximate solution based on differential equation:
  - $P_{A \to A}(t) = 3/4\ e^{-4\alpha t} + 1/4$
- How similar are $(1 - 4\alpha)^t$ and $e^{-4\alpha t}$?
  - By Taylor expansion, $e^x = 1 + x/1! + x^2/2! + x^3/3! + \ldots$
  - Therefore $e^{-4\alpha t} = 1 - 4\alpha t + (4\alpha t)^2/2 - (4\alpha t)^3/6 + \ldots$
  - On the other hand,

    $(1 - 4\alpha)^t = 1 - 4\alpha t + t(t\text{-}1)/2(4\alpha)^2 - t(t\text{-}1)(t\text{-}2)/6(4\alpha)^3 + \ldots$

    $= 1 - 4\alpha t + [(t^2\text{-}t)(4\alpha)^2/2] - [(t^3\text{-}3t^2\text{+}2t)(4\alpha)^3/6] + \ldots$
  - Therefore, their difference is

    $e^{-4\alpha t} - (1 - 4\alpha)^t = t(4\alpha)^2/2 - (3t^2\text{-}2t)(4\alpha)^3/6 + \ldots$

Problem revisited

- What have we done so far?
  - Given
    1. The ancestral state of a site
    2. The substitution rate α (probability of each type of mutation in unit time)
  - Determine the probability of the current state, which is *t* units of time after the separation event, where *t* is also given

- What do we really want?
  - Given the current states of two sequences
    - The ancestral state is unknown
  - Determine the number of substitutions happened in the two sequences since their divergence, both observed and unobserved
    - The substitution rate (α) and for how long the two sequences have diverged (*t*) are also unknown

Difficulties and ideas

- Difficulty #1: We do not know the ancestral state, mutation rate α or divergence time t
  - Solutions:
    - Due to symmetry, the ancestral state does not matter if we care only about whether two current sequences have the same nucleotide or not at each site
      - $P_{same}(t)$

$$= [P_{A\to A}(t)]^2 + [P_{A\to C}(t)]^2 + [P_{A\to G}(t)]^2 + [P_{A\to T}(t)]^2$$
$$= [P_{C\to A}(t)]^2 + [P_{C\to C}(t)]^2 + [P_{C\to G}(t)]^2 + [P_{C\to T}(t)]^2$$
$$= [P_{G\to A}(t)]^2 + [P_{G\to C}(t)]^2 + [P_{G\to G}(t)]^2 + [P_{G\to T}(t)]^2$$
$$= [P_{T\to A}(t)]^2 + [P_{T\to C}(t)]^2 + [P_{T\to G}(t)]^2 + [P_{T\to T}(t)]^2$$

    - We do not know α or t, but we can easily estimate their product αt and it turns out this is all that we need

Difficulties and ideas

- Difficulty #2: Even if we knew the ancestral state, mutation rate α and divergence time *t*, there would still be an infinite number of possibilities each with an associated probability

  – Solution:

  - We will talk about the *expected number* of mutations happened

    i.e., the average of all cases by considering their number of mutations and probability of happening

Expectation and variance

- A little bit on two basic statistical concepts:
  - Consider flipping an unfair coin with 60% chance of head and 40% chance of tail. We flip it 100 times. How many heads do we get?
  - We do not know before the experiment, because the number could vary every time
  - But if we repeat the experiment many times, on average we get 60 heads per experiment
    - This is called the **expectation**
  - There are variations between the numbers obtained from different experiments. We can quantify the variation by the average squared difference between the observed numbers and the expectation
    - This is called the **variance**
  - If we change to 30% chance of head but flip 200 times per experiment, the expected number of heads per experiment remains 60

How do these concepts help in solving our problem?

- Although we do not know whether a mutation appears after one time unit, but if we consider a large number of time units, the **expected** number of mutations in t time units is $3\alpha t$

- If we can estimate the number of mutations that have happened, we can compute $3\alpha t$ even we never know the separate values of α and t alone
  - That's why we use the approximation formula for $P_{A\to A}(t) = 3/4\ e^{-4\alpha t} + 1/4$

    It only involves the product of α and t but not their separate values

Estimating the number of substitutions

- For the Jukes-Cantor model:
  - For a single site, probability for two sequences separated t units of time ago having the same state is (assuming the ancestral state was A, but the same formula holds for other ancestral states): – Idea #1

$$
\begin{aligned}
P_{same}(t) &= [P_{A\to A}(t)]^2 + [P_{A\to C}(t)]^2 + [P_{A\to G}(t)]^2 + [P_{A\to T}(t)]^2 \\
&= \left(\frac{1}{4} + \frac{3}{4}e^{-4\alpha t}\right)^2 + 3\left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right)^2 \\
&= \left(\frac{1}{16} + \frac{6}{16}e^{-4\alpha t} + \frac{9}{16}e^{-8\alpha t}\right) + \left(\frac{3}{16} - \frac{6}{16}e^{-4\alpha t} + \frac{3}{16}e^{-8\alpha t}\right) \\
&= \frac{1}{4} + \frac{3}{4}e^{-8\alpha t}
\end{aligned}
$$

  - Correspondingly, the probability that the two sequences have different states at a single site is – Idea #2

$$
\begin{aligned}
p_{diff} &\equiv 1 - \left(\frac{1}{4} + \frac{3}{4}e^{-8\alpha t}\right) = \frac{3}{4}(1 - e^{-8\alpha t}) \\
\Rightarrow \quad \alpha t &= -\frac{1}{8}\ln\left(1 - \frac{4}{3}p_{diff}\right)
\end{aligned}
$$

We don't know the value of α (substitution rate) or t (number of time units since the divergence of the two sequences), but we can estimate p_{diff}, which will give us an estimate of αt.

Estimating the number of substitutions

- How to estimate p_{diff} probability for two random sequences generated according to the above procedure to have different states at a site?
  - We estimate p_{diff} by x/n, where x is the number of sites different between the observed sequences – our best guess based on observed data
- Putting everything together:
  - Suppose we have two length-n sequences diverged t units of time ago currently with x mismatches (assuming no indels)
  - Let K_{sup} be the no. of substitutions per site happened to the two sequences since their divergence
  - According to the Jukes-Cantor model, the expected value of K_{sup} is (from previous page)

$$E[K_{sup}] = 2(3\alpha t) = 6\left[-\frac{1}{8}\ln\left(1-\frac{4}{3}p_{diff}\right)\right] = -\frac{3}{4}\ln\left(1-\frac{4}{3}p_{diff}\right) = -\frac{3}{4}\ln\left(1-\frac{4x}{3n}\right)$$

  - For large n, variance of this estimation is approximately $\frac{p_{diff}-(p_{diff})^2}{n\left(1-\frac{4}{3}p_{diff}\right)^2} = \frac{x/n-(x/n)^2}{n\left(1-\frac{4x}{3n}\right)^2}$

Estimate the variance, details refer to: https://link.springer.com/content/pdf/10.1007/BF01653945.pdf

Estimating the number of substitutions

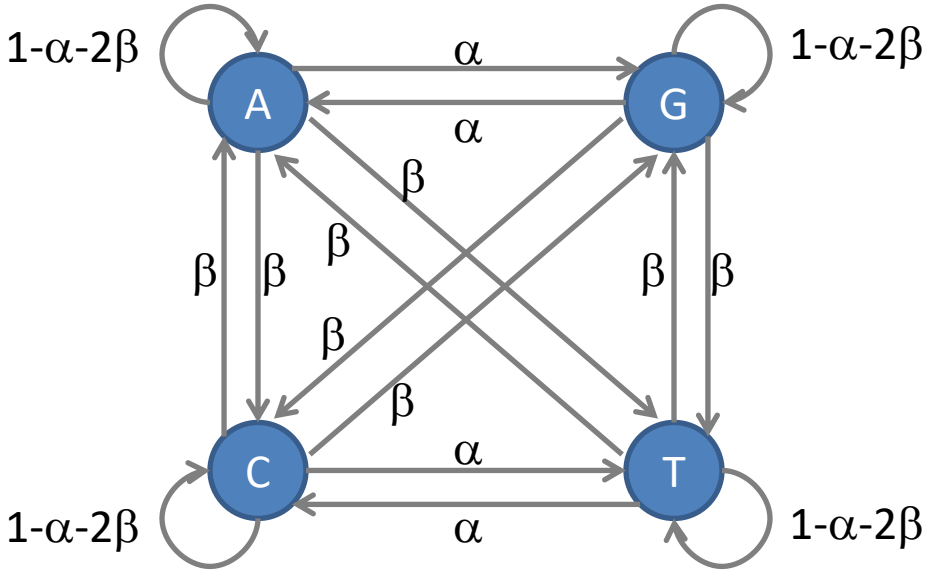
- Let's see how we can apply the results
- Example:
  - Suppose two sequences each with n=200 nucleotides have x=66 observed mismatches, then
    - $p_{diff} = x/n = 66/200 = 0.33$
    - $E[K_{sup}] = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p_{diff}\right) \approx 0.43$
    - Variance of this estimation is $\frac{p_{diff} - (p_{diff})^2}{n\left(1 - \frac{4}{3}p_{diff}\right)^2} \approx 0.0035$
  - Observations:
    1. Observed number of substitutions per site is smaller than the estimated number of (observed + unobserved) substitutions per site, as expected
    2. Variance is fairly large -- The actual number may be a bit different from this estimate (would be smaller for large n)

Evolutionary distance

- Now, we can use $E[K_{sup}]$ as a measure of the evolutionary distance between two aligned sequences without indels.
  - Using a mutation model to define the distance is more theoretical grounded than defining the substitution score matrix arbitrarily (like match=1, mismatch=-1).
  - Of course, we need to align the sequences first, at which time we still need a substitution matrix to start with
- There are other models that allow
  - More parameters (e.g., different sub-types of substitutions)
  - Variable rates at different sties
  - Dependency between different sites
  - Changing substitution rates over time
  - Indels
- Let's study one more model that considers differences between transitions and transversions

The Kimura two-parameter model

- Kimura proposed the following model in 1980:
  - Assume transitions are more frequent than transversions
  - In one unit of time, probability of α for a transition to happen and probability of β ($<\alpha$) for a transversion to happen



- Again, it is possible to estimate $E[K_{sup}]$ without knowing α, β and t

Recursive formulas [optional]

- $P_{A\to A}(t+1) = (1 - \alpha - 2\beta)P_{A\to A}(t) + \beta P_{A\to C}(t) + \alpha P_{A\to G}(t) + \beta P_{A\to T}(t)$
- $\Delta P_{A\to A}(t)$
  $= P_{A\to A}(t+1) - P_{A\to A}(t)$
  $= (1 - \alpha - 2\beta)P_{A\to A}(t) + \beta P_{A\to C}(t) + \alpha P_{A\to G}(t) + \beta P_{A\to T}(t) - P_{A\to A}(t)$
  $= -(\alpha + 2\beta)P_{A\to A}(t) + \beta P_{A\to C}(t) + \alpha P_{A\to G}(t) + \beta P_{A\to T}(t)$

- For an infinitesimally small time unit,

$$\frac{dP_{A\to A}(t)}{dt} = -[\alpha + 2\beta]P_{A\to A}(t) + \beta P_{A\to C}(t) + \alpha P_{A\to G}(t) + \beta P_{A\to T}(t)$$

- Solving the four simultaneous differential equations for

$\frac{dP_{A\to A}(t)}{dt}$, $\frac{dP_{A\to C}(t)}{dt}$, $\frac{dP_{A\to G}(t)}{dt}$ and $\frac{dP_{A\to T}(t)}{dt}$, we get the formulas on the next slide

Final formulas

- Same base after t units of time:

$$\mathrm{P}_{\mathrm{A}\to\mathrm{A}}(t) = \mathrm{P}_{\mathrm{C}\to\mathrm{C}}(t) = \mathrm{P}_{\mathrm{G}\to\mathrm{G}}(t) = \mathrm{P}_{\mathrm{T}\to\mathrm{T}}(t) = \frac{1}{4} + \frac{1}{4}\mathrm{e}^{-4\beta t} + \frac{1}{2}\mathrm{e}^{-2(\alpha+\beta)t}$$

- An observed transition after t units of time:

$$\mathrm{P}_{\mathrm{A}\to\mathrm{G}}(t) = \mathrm{P}_{\mathrm{C}\to\mathrm{T}}(t) = \mathrm{P}_{\mathrm{G}\to\mathrm{A}}(t) = \mathrm{P}_{\mathrm{T}\to\mathrm{C}}(t) = \frac{1}{4} + \frac{1}{4}\mathrm{e}^{-4\beta t} - \frac{1}{2}\mathrm{e}^{-2(\alpha+\beta)t}$$

- An observed transversion after t units of time:

$$\mathrm{P}_{\mathrm{A}\to\mathrm{C}}(t) = \mathrm{P}_{\mathrm{A}\to\mathrm{T}}(t) = \cdots = \mathrm{P}_{\mathrm{T}\to\mathrm{G}}(t) = \frac{1}{4} - \frac{1}{4}\mathrm{e}^{-4\beta t}$$

Estimating the number of substitutions

- Skipping the remaining derivations:
  - Suppose there are x_1 observed transitions and x_2 observed transversions
  - Correspondingly, we estimate the probability of having a transition and a transversion per site in the final sequences as $p_{diff1} \equiv x_1/n$ and $p_{diff2} \equiv x_2/n$, respectively
  - Then using a derivation similar as before, we get
    - $$E[K_{sup}] = \frac{1}{2}\ln\left(\frac{1}{1-2p_{diff\,1}-p_{diff\,2}}\right) + \frac{1}{4}\ln\left(\frac{1}{1-2p_{diff\,2}}\right)$$
    - For large n, , variance of this estimation

$$= \frac{1}{n}\left[p_{diff\,1}\left(\frac{1}{1-2p_{diff\,1}-p_{diff\,2}}\right)^2 + p_{diff\,2}\left(\frac{1}{2-4p_{diff\,1}-2p_{diff\,2}}+\frac{1}{2-4p_{diff\,2}}\right)^2 - \left(\frac{p_{diff\,1}}{1-2p_{diff\,1}-p_{diff\,2}}+\frac{p_{diff\,2}}{2-4p_{diff\,1}-2p_{diff\,2}}+\frac{p_{diff\,2}}{2-4p_{diff\,2}}\right)^2\right]$$

Estimating the number of substitutions

- Example:
  - Suppose two sequences each with n=200 nucleotides have x_1=50 observed transitions and x_2=16 transversions, then
    - $p_{diff1} = x_1/n = 50/200 = 0.25$
    - $p_{diff2} = x_2/n = 16/200 = 0.08$
    - $$E[K_{sup}] = \frac{1}{2}\ln\left(\frac{1}{1-2p_{diff\,1}-p_{diff\,2}}\right) + \frac{1}{4}\ln\left(\frac{1}{1-2p_{diff\,2}}\right) \approx 0.48$$
    - Variance of this estimation ≈ 0.0056
    - This estimated number of substitutions per site is even larger than the estimate from the Jukes-Cantor model
      - This one may be more accurate for more diverged sequences

Part 2

Substitution Matrices for Amino Acids

Amino acid substitutions

- If we only see amino acid sequences, it is more difficult to estimate the number of DNA substitutions
  - Because one amino acid substitution can be caused by several possible DNA substitutions
  - However, similar ideas still apply. For example, the observed number of substitutions is likely smaller than the actual number of substitutions happened

Observed Percent Difference	Evolutionary Distance in PAM
1	1
5	5
10	11
15	17
20	23
25	30
30	38
35	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246
85	328

Table source: Dayhoff et al., *Atlas of Protein Sequence and Structure* 5(3):345-352, (1978)

Amino acid substitutions

- For non-coding DNA, we can build simple yet highly reasonable mutation models using just one or two parameters

- In contrast, amino acid substitutions depend heavily on biochemical properties, and it is difficult to form a simple model
  - For example, a leucine is more likely substituted by an isoleucine than a valine
  - Instead, people have estimated substitution rates from data in large databases

PAM

- One commonly used series of substitution matrices for amino acids is PAM (point accepted mutation)
  - "Accepted" means survived, so that we can observe today
- Created by Dayhoff et al. in 1978 based on 1,572 observed substitutions in 71 families of closely related proteins

PAM1

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	0.9867	0.0001	0.0004	0.0006	0.0001	0.0003	0.0010	0.0021	0.0001	0.0002	0.0003	0.0002	0.0001	0.0001	0.0013	0.0028	0.0022	0.0000	0.0001	0.0013
Arg R	0.0002	0.9913	0.0001	0.0000	0.0001	0.0009	0.0000	0.0001	0.0008	0.0002	0.0001	0.0037	0.0001	0.0001	0.0005	0.0011	0.0002	0.0002	0.0000	0.0002
Asn N	0.0009	0.0001	0.9822	0.0042	0.0000	0.0004	0.0007	0.0012	0.0018	0.0003	0.0003	0.0025	0.0000	0.0001	0.0002	0.0034	0.0013	0.0000	0.0003	0.0001
Asp D	0.0010	0.0000	0.0036	0.9859	0.0000	0.0005	0.0056	0.0011	0.0003	0.0001	0.0000	0.0006	0.0000	0.0000	0.0001	0.0007	0.0004	0.0000	0.0000	0.0001
Cys C	0.0003	0.0001	0.0000	0.0000	0.9973	0.0000	0.0000	0.0001	0.0001	0.0002	0.0000	0.0000	0.0000	0.0000	0.0001	0.0011	0.0001	0.0000	0.0003	0.0003
Gln Q	0.0008	0.0010	0.0004	0.0006	0.0000	0.9876	0.0035	0.0003	0.0020	0.0001	0.0006	0.0012	0.0002	0.0000	0.0008	0.0004	0.0003	0.0000	0.0000	0.0002
Glu E	0.0017	0.0000	0.0006	0.0053	0.0000	0.0027	0.9865	0.0007	0.0001	0.0002	0.0001	0.0007	0.0000	0.0000	0.0003	0.0006	0.0002	0.0000	0.0001	0.0002
Gly G	0.0021	0.0000	0.0006	0.0006	0.0000	0.0001	0.0004	0.9935	0.0000	0.0000	0.0001	0.0002	0.0000	0.0001	0.0002	0.0016	0.0002	0.0000	0.0000	0.0003
His H	0.0002	0.0010	0.0021	0.0004	0.0001	0.0023	0.0002	0.0001	0.9912	0.0000	0.0004	0.0002	0.0000	0.0002	0.0005	0.0002	0.0001	0.0000	0.0004	0.0003
Ile I	0.0006	0.0003	0.0003	0.0001	0.0001	0.0001	0.0003	0.0000	0.0000	0.9872	0.0022	0.0004	0.0005	0.0008	0.0001	0.0002	0.0011	0.0000	0.0001	0.0057
Leu L	0.0004	0.0001	0.0001	0.0000	0.0000	0.0003	0.0001	0.0001	0.0001	0.0009	0.9947	0.0001	0.0008	0.0006	0.0002	0.0001	0.0002	0.0000	0.0001	0.0011
Lys K	0.0002	0.0019	0.0013	0.0003	0.0000	0.0006	0.0004	0.0002	0.0001	0.0002	0.0002	0.9926	0.0004	0.0000	0.0002	0.0007	0.0008	0.0000	0.0000	0.0001
Met M	0.0006	0.0004	0.0000	0.0000	0.0000	0.0004	0.0001	0.0001	0.0000	0.0012	0.0045	0.0020	0.9874	0.0004	0.0001	0.0004	0.0006	0.0000	0.0000	0.0017
Phe F	0.0002	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0007	0.0013	0.0000	0.0001	0.9946	0.0001	0.0003	0.0001	0.0001	0.0021	0.0001
Pro P	0.0022	0.0004	0.0002	0.0001	0.0001	0.0006	0.0003	0.0003	0.0003	0.0000	0.0003	0.0003	0.0000	0.0000	0.9926	0.0017	0.0005	0.0000	0.0000	0.0003
Ser S	0.0035	0.0006	0.0020	0.0005	0.0005	0.0002	0.0004	0.0021	0.0001	0.0001	0.0001	0.0008	0.0001	0.0002	0.0012	0.9840	0.0032	0.0001	0.0001	0.0002
Thr T	0.0032	0.0001	0.0009	0.0003	0.0001	0.0002	0.0002	0.0003	0.0001	0.0007	0.0003	0.0011	0.0002	0.0001	0.0004	0.0038	0.9871	0.0000	0.0001	0.0010
Trp W	0.0000	0.0008	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0004	0.0000	0.0000	0.0003	0.0000	0.0005	0.0000	0.9976	0.0002	0.0000
Tyr Y	0.0002	0.0000	0.0004	0.0000	0.0003	0.0000	0.0001	0.0000	0.0004	0.0001	0.0002	0.0001	0.0000	0.0028	0.0000	0.0002	0.0002	0.0001	0.9945	0.0002
Val V	0.0018	0.0001	0.0001	0.0001	0.0002	0.0001	0.0002	0.0005	0.0001	0.0033	0.0015	0.0001	0.0004	0.0000	0.0002	0.0002	0.0009	0.0000	0.0001	0.9901

Source of the original table (transpose of this one): http://www.icp.ucl.ac.be/~opperd/private/pam1.html

PAM

- One commonly used series of substitution matrices for amino acids is PAM (point accepted mutation)
  - "Accepted" means survived, so that we can observe today
- Created by Dayhoff et al. in 1978 based on 1,572 observed substitutions in 71 families of closely related proteins
- In the PAM1 matrix, each element records **the probability of that substitution given a mutation rate of one substitution per 100 amino acids**
  - Reason to define in this way is to setup a time context without knowing exactly how long it was
  - For any $x \geq 1$, the PAMx matrix records substitution probabilities given a mutation rate of x substitutions per 100 amino acids
    - Larger $x \Rightarrow$ more substitutions
    - It is equal to PAM1 to the power x (matrix multiplication)
  - The matrix is asymmetric

PAM1

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	0.9867	0.0001	0.0004	0.0006	0.0001	0.0003	0.0010	0.0021	0.0001	0.0002	0.0003	0.0002	0.0001	0.0001	0.0013	0.0028	0.0022	0.0000	0.0001	0.0013
Arg R	0.0002	0.9913	0.0001	0.0000	0.0001	0.0009	0.0000	0.0001	0.0008	0.0002	0.0001	0.0037	0.0001	0.0001	0.0005	0.0011	0.0002	0.0002	0.0000	0.0002
Asn N	0.0009	0.0001	0.9822	0.0042	0.0000	0.0004	0.0007	0.0012	0.0018	0.0003	0.0003	0.0025	0.0000	0.0001	0.0002	0.0034	0.0013	0.0000	0.0003	0.0001
Asp D	0.0010	0.0000	0.0036	0.9859	0.0000	0.0005	0.0056	0.0011	0.0003	0.0001	0.0000	0.0006	0.0000	0.0000	0.0001	0.0007	0.0004	0.0000	0.0000	0.0001
Cys C	0.0003	0.0001	0.0000	0.0000	0.9973	0.0000	0.0000	0.0001	0.0001	0.0002	0.0000	0.0000	0.0000	0.0000	0.0001	0.0011	0.0001	0.0000	0.0003	0.0003
Gln Q	0.0008	0.0010	0.0004	0.0006	0.0000	0.9876	0.0035	0.0003	0.0020	0.0001	0.0006	0.0012	0.0002	0.0000	0.0008	0.0004	0.0003	0.0000	0.0000	0.0002
Glu E	0.0017	0.0000	0.0006	0.0053	0.0000	0.0027	0.9865	0.0007	0.0001	0.0002	0.0001	0.0007	0.0000	0.0000	0.0003	0.0006	0.0002	0.0000	0.0001	0.0002
Gly G	0.0021	0.0000	0.0006	0.0006	0.0000	0.0001	0.0004	0.9935	0.0000	0.0000	0.0001	0.0002	0.0000	0.0001	0.0002	0.0016	0.0002	0.0000	0.0000	0.0003
His H	0.0002	0.0010	0.0021	0.0004	0.0001	0.0023	0.0002	0.0001	0.9912	0.0000	0.0004	0.0002	0.0000	0.0002	0.0005	0.0002	0.0001	0.0000	0.0004	0.0003
Ile I	0.0006	0.0003	0.0003	0.0001	0.0001	0.0001	0.0003	0.0000	0.0000	0.9872	0.0022	0.0004	0.0005	0.0008	0.0001	0.0002	0.0011	0.0000	0.0001	0.0057
Leu L	0.0004	0.0001	0.0001	0.0000	0.0000	0.0003	0.0001	0.0001	0.0001	0.0009	0.9947	0.0001	0.0008	0.0006	0.0002	0.0001	0.0002	0.0000	0.0001	0.0011
Lys K	0.0002	0.0019	0.0013	0.0003	0.0000	0.0006	0.0004	0.0002	0.0001	0.0002	0.0002	0.9926	0.0004	0.0000	0.0002	0.0007	0.0008	0.0000	0.0000	0.0001
Met M	0.0006	0.0004	0.0000	0.0000	0.0000	0.0004	0.0001	0.0001	0.0000	0.0012	0.0045	0.0020	0.9874	0.0004	0.0001	0.0004	0.0006	0.0000	0.0000	0.0017
Phe F	0.0002	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0007	0.0013	0.0000	0.0001	0.9946	0.0001	0.0003	0.0001	0.0001	0.0021	0.0001
Pro P	0.0022	0.0004	0.0002	0.0001	0.0001	0.0006	0.0003	0.0003	0.0003	0.0000	0.0003	0.0003	0.0000	0.0000	0.9926	0.0017	0.0005	0.0000	0.0000	0.0003
Ser S	0.0035	0.0006	0.0020	0.0005	0.0005	0.0002	0.0004	0.0021	0.0001	0.0001	0.0001	0.0008	0.0001	0.0002	0.0012	0.9840	0.0032	0.0001	0.0001	0.0002
Thr T	0.0032	0.0001	0.0009	0.0003	0.0001	0.0002	0.0002	0.0003	0.0001	0.0007	0.0003	0.0011	0.0002	0.0001	0.0004	0.0038	0.9871	0.0000	0.0001	0.0010
Trp W	0.0000	0.0008	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0004	0.0000	0.0000	0.0003	0.0000	0.0005	0.0000	0.9976	0.0002	0.0000
Tyr Y	0.0002	0.0000	0.0004	0.0000	0.0003	0.0000	0.0001	0.0000	0.0004	0.0001	0.0002	0.0001	0.0000	0.0028	0.0000	0.0002	0.0002	0.0001	0.9945	0.0002
Val V	0.0018	0.0001	0.0001	0.0001	0.0002	0.0001	0.0002	0.0005	0.0001	0.0033	0.0015	0.0001	0.0004	0.0000	0.0002	0.0002	0.0009	0.0000	0.0001	0.9901

If there is a mutation rate of 1 substitution per 100 amino acids, there is a probability of about 0.0033 that a valine would be mutated to an isoleucine (why much larger than others?)

Please also take note on this number, which is smaller

Source of the original table (transpose of this one): http://www.icp.ucl.ac.be/~opperd/private/pam1.html

Valine and isoleucine

Amino Acid	Side-chain polarity	Side-chain charge (pH 7.4)	Hydropathy index
Alanine	nonpolar	neutral	1.8
Arginine	polar	positive	−4.5
Asparagine	polar	neutral	−3.5
Aspartic acid	polar	negative	−3.5
Cysteine	polar	neutral	2.5
Glutamic acid	polar	negative	−3.5
Glutamine	polar	neutral	−3.5
Glycine	nonpolar	neutral	−0.4
Histidine	polar	positive(10%) neutral(90%)	−3.2
Isoleucine	nonpolar	neutral	4.5

Amino Acid	Side-chain polarity	Side-chain charge (pH 7.4)	Hydropathy index
Leucine	nonpolar	neutral	3.8
Lysine	polar	positive	−3.9
Methionine	nonpolar	neutral	1.9
Phenylalanine	nonpolar	neutral	2.8
Proline	nonpolar	neutral	−1.6
Serine	polar	neutral	−0.8
Threonine	polar	neutral	−0.7
Tryptophan	nonpolar	neutral	−0.9
Tyrosine	polar	neutral	−1.3
Valine	nonpolar	neutral	4.2

- Similar chemical properties

→ Less impact upon substitution

→ Substitution more likely to occur

Information source: Wikipedia

PAM250

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	0.13	0.03	0.04	0.05	0.02	0.03	0.05	0.12	0.02	0.03	0.06	0.06	0.01	0.02	0.07	0.09	0.08	0.00	0.01	0.07
Arg R	0.06	0.17	0.04	0.04	0.01	0.05	0.04	0.05	0.05	0.02	0.04	0.18	0.01	0.01	0.05	0.06	0.05	0.02	0.01	0.04
Asn N	0.09	0.04	0.06	0.08	0.01	0.05	0.07	0.10	0.05	0.02	0.04	0.10	0.01	0.02	0.05	0.08	0.06	0.00	0.02	0.04
Asp D	0.09	0.03	0.07	0.11	0.01	0.06	0.11	0.10	0.04	0.02	0.03	0.08	0.01	0.01	0.04	0.07	0.06	0.00	0.01	0.04
Cys C	0.05	0.02	0.02	0.01	0.52	0.01	0.01	0.04	0.02	0.02	0.02	0.02	0.00	0.01	0.03	0.07	0.04	0.00	0.03	0.04
Gln Q	0.08	0.05	0.05	0.07	0.01	0.10	0.09	0.07	0.07	0.02	0.06	0.10	0.01	0.01	0.05	0.06	0.05	0.00	0.01	0.04
Glu E	0.09	0.03	0.06	0.10	0.01	0.07	0.12	0.09	0.04	0.02	0.04	0.08	0.01	0.01	0.04	0.07	0.05	0.00	0.01	0.04
Gly G	0.12	0.02	0.04	0.05	0.02	0.03	0.05	0.27	0.02	0.02	0.03	0.05	0.01	0.01	0.05	0.09	0.06	0.00	0.01	0.04
His H	0.06	0.06	0.06	0.06	0.02	0.07	0.06	0.05	0.15	0.02	0.05	0.08	0.01	0.03	0.05	0.06	0.04	0.01	0.03	0.05
Ile I	0.08	0.03	0.03	0.03	0.02	0.02	0.03	0.05	0.02	0.10	0.15	0.05	0.02	0.05	0.03	0.05	0.06	0.00	0.02	0.04
Leu L	0.06	0.02	0.02	0.02	0.01	0.03	0.02	0.04	0.02	0.06	0.34	0.04	0.03	0.06	0.03	0.04	0.04	0.01	0.02	0.15
Lys K	0.07	0.09	0.05	0.05	0.01	0.05	0.05	0.06	0.03	0.02	0.04	0.24	0.02	0.01	0.04	0.07	0.06	0.00	0.01	0.10
Met M	0.07	0.04	0.03	0.03	0.01	0.03	0.03	0.05	0.02	0.06	0.20	0.09	0.06	0.04	0.03	0.05	0.05	0.00	0.02	0.04
Phe F	0.04	0.01	0.02	0.01	0.01	0.01	0.01	0.03	0.02	0.05	0.13	0.02	0.02	0.32	0.02	0.03	0.03	0.01	0.15	0.10
Pro P	0.11	0.04	0.04	0.04	0.02	0.04	0.04	0.08	0.03	0.02	0.05	0.06	0.01	0.01	0.20	0.09	0.06	0.00	0.01	0.05
Ser S	0.11	0.04	0.05	0.05	0.03	0.03	0.05	0.11	0.03	0.03	0.04	0.08	0.01	0.02	0.06	0.10	0.08	0.01	0.02	0.05
Thr T	0.11	0.03	0.04	0.05	0.02	0.03	0.05	0.09	0.02	0.04	0.06	0.08	0.01	0.02	0.05	0.09	0.11	0.00	0.02	0.05
Trp W	0.02	0.07	0.02	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.06	0.04	0.01	0.04	0.01	0.04	0.02	0.55	0.03	0.72
Tyr Y	0.04	0.02	0.03	0.02	0.04	0.02	0.02	0.03	0.03	0.03	0.07	0.03	0.01	0.20	0.02	0.04	0.03	0.01	0.31	0.04
Val V	0.09	0.02	0.03	0.03	0.02	0.03	0.03	0.07	0.02	0.09	0.13	0.05	0.02	0.03	0.04	0.06	0.06	0.00	0.02	0.17

If there is a mutation rate of 250 substitutions per 100 amino acids, there is a probability of about 0.09 that a valine would be mutated to an isoleucine

Why V→L is more likely than V→I?
Because (1) there are indirect paths for V to get mutated to L, (2) I is a source and (3) L is a sink

Source of the original table (transpose of this one): http://www.icp.ucl.ac.be/~opperd/private/pam250.html

Another matrix: BLOSUM

- Blocks of amino acid substitution matrix is another set of commonly used substitution matrix for amino acids.
- Based on local alignments of very conserved protein regions

BLOSUM62

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
Arg R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
Asn N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
Asp D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
Cys C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Gln Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
Glu E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
Gly G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
His H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
Ile I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
Leu L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
Lys K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
Met M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
Phe F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
Pro P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
Ser S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
Thr T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
Trp W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Tyr Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
Val V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Table source: http://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt

Another matrix: BLOSUM

- Blocks of amino acid substitution matrix is another set of commonly used substitution matrix for amino acids.
- Based on local alignments of very conserved protein regions
- The BLOSUM*y* matrix (*y* between 0 and 100), the local alignments involve sequences that are more than *y*% identical
  - Larger $y \Rightarrow$ less substitutions
- The entry at the *i*-th row and *j*-th column is the log-odd score:
  - $S_{ij} = 1/\lambda \log_2(p_{ij} / (p_i p_j))$, where
    - p_{ij} is the fraction of observed substitutions between amino acids *i* and *j*
    - p_i and p_j are the fraction of sites with amino acids *i* and *j*, respectively
    - λ is a scaling factor to make the numbers close to integers
  - For example, if amino acids *i* and *j* are independent, i.e., $p_i p_j = p_{ij}$, then $\log_2(p_{ij} / (p_i p_j)) = 0$

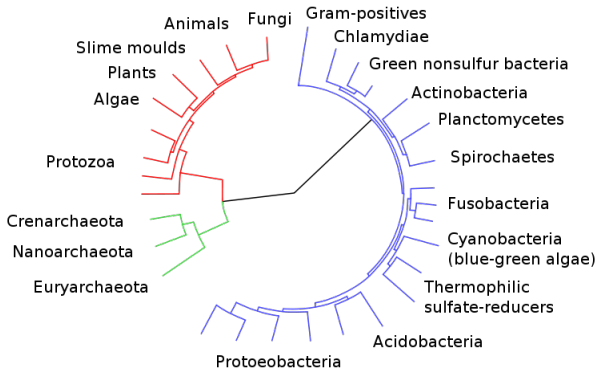
PAM vs. BLOSUM

- Meaning of numbers in the matrices:
  - PAM: probabilities of substitution $P_{i \to j}(x)$
    - Asymmetric
  - BLOSUM: log odds of observed substitutions and expectation
    - Symmetric
  - Can also compute log odds for the PAM probabilities
    (some "PAM matrices" that you can find on the Web are actually log odds)
- Construction methods:
  - PAM: groups of related proteins
  - BLOSUM: local alignments of very conserved regions of proteins
- Meaning of *x* and *y* in PAM*x* and BLOSUM*y*
  - PAM: mutation rate of *x* substitutions per 100 amino acids.
    Larger *x* means more substitutions
  - BLOSUM: identity threshold above which sequences are grouped to perform local alignment.
    Larger *y* means less substitutions

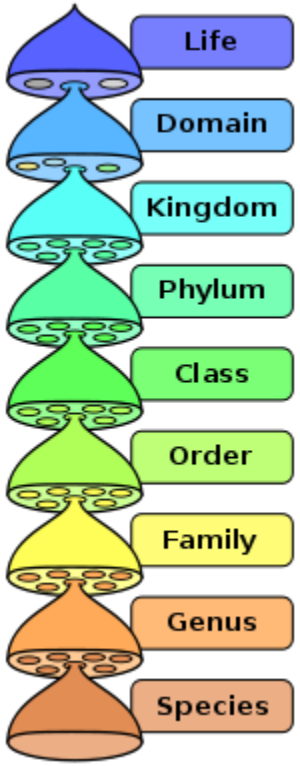
Part 3

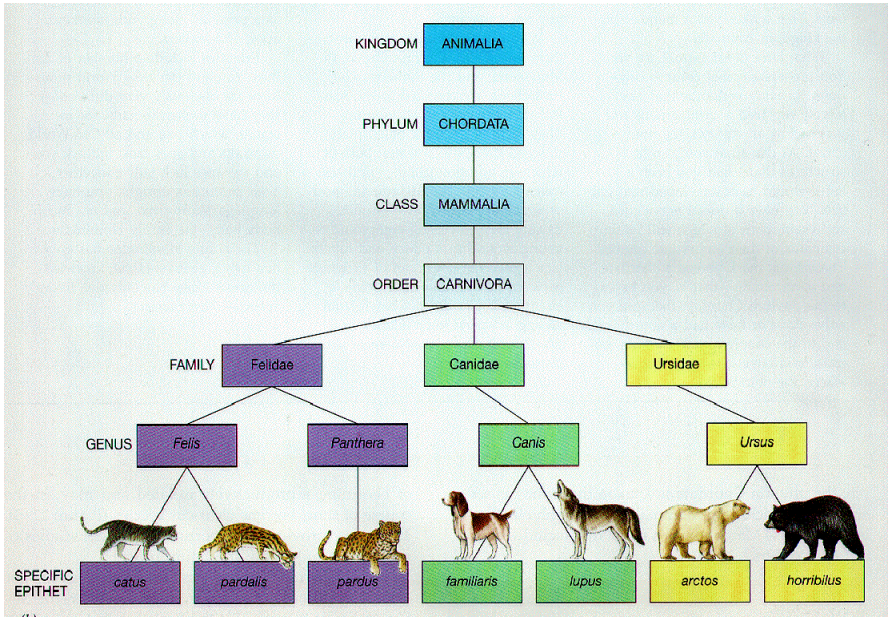
Trees

Classification of species



Domains and kingdoms

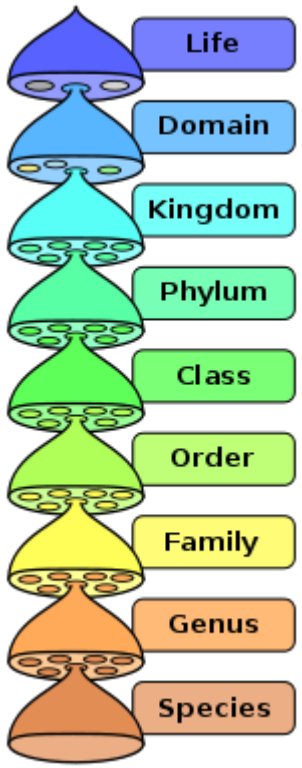


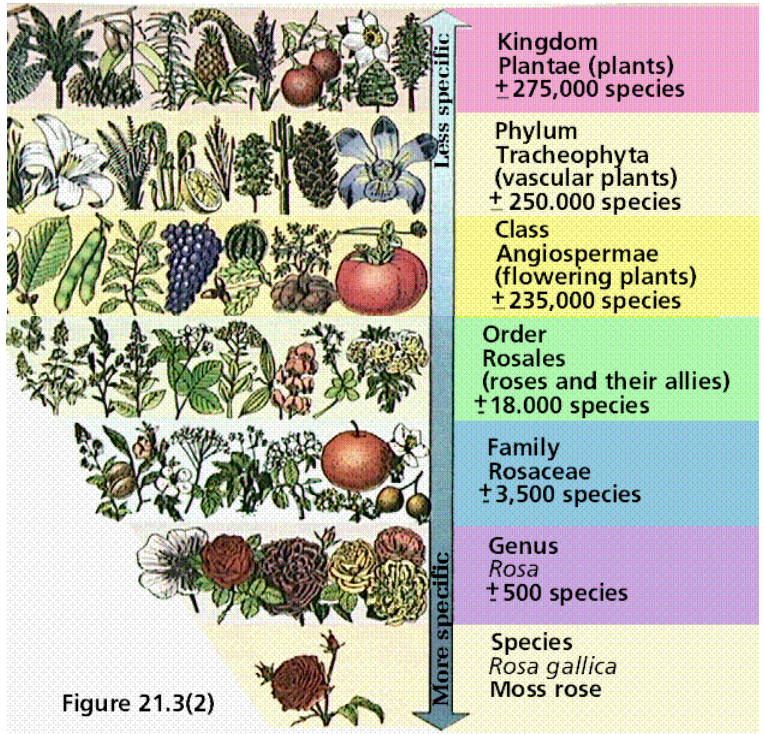


The animal kingdom

Image credit: Wikipedia, http://ridge.icu.ac.jp/gen-ed/classif-gifs/animal-class-example.gif

Taxonomy





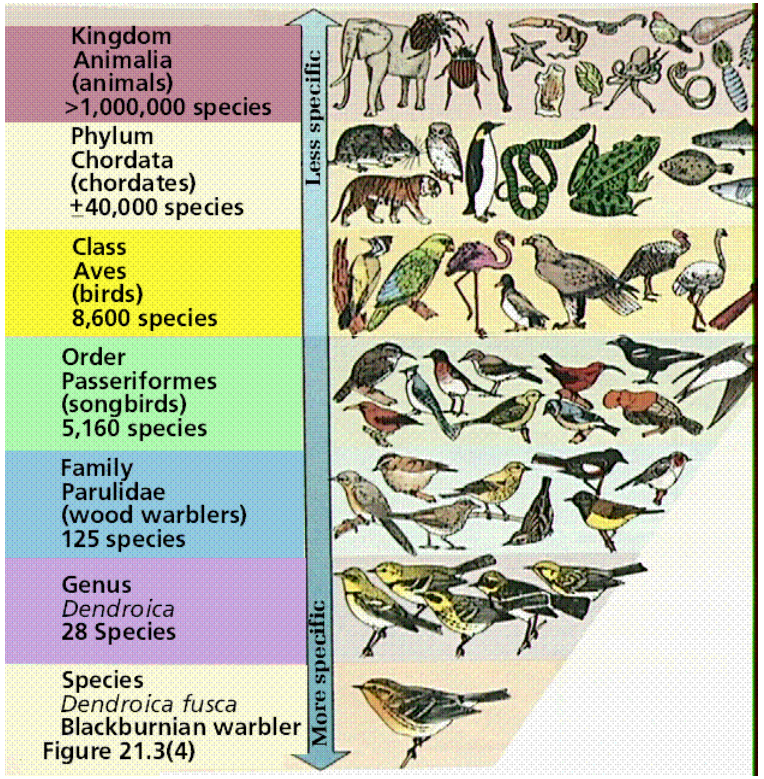
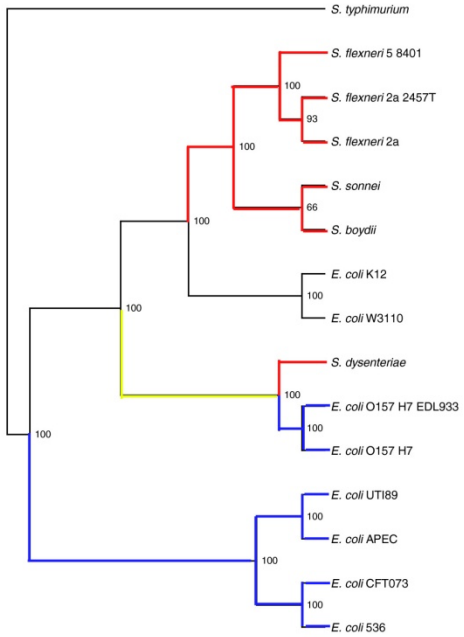
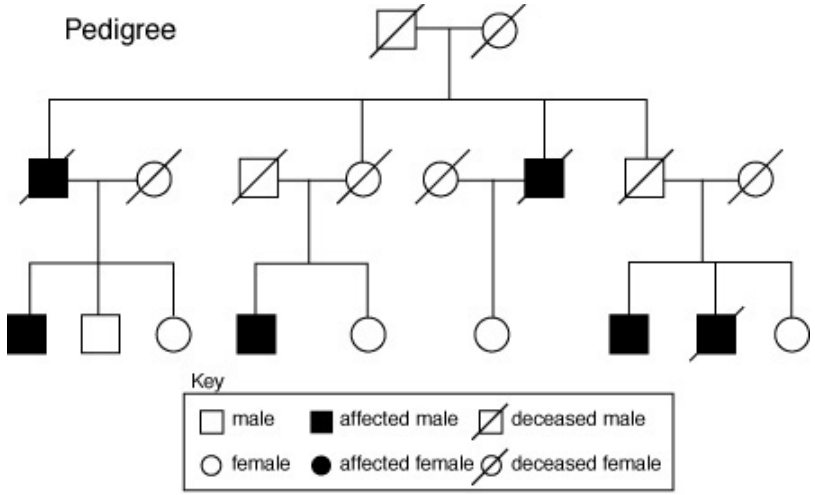


Image credit: Wikipedia, http://www2.estrellamountain.edu/faculty/farabee/biobk/BioBookDivers_class.html

Finer scales

- Same idea can be applied to classify different strains of a type of bacteria
- Or even family relationships





What genes tell us about inheriting diabetes.

Image credit: Hershberg et al., Genome Biology 8:R164 (2007), http://www.accessexcellence.org/RC/VL/GG/images/pedigree.gif, http://www.jdrf.ca/

Relating biological objects

- How were the hierarchies determined?
  - Species: traditionally by morphological and behavioral similarities, or paleontological evidences
  - Bacterial strains: by physical, chemical and biological properties
- Question: Which features should be used first?

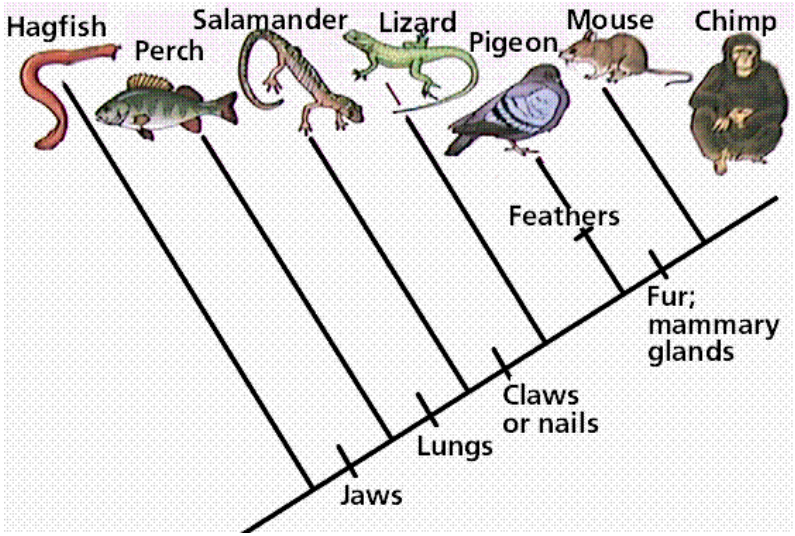


Image credit: http://www2.estrellamountain.edu/faculty/farabee/biobk/BioBookDivers_class.html

Phylogeny

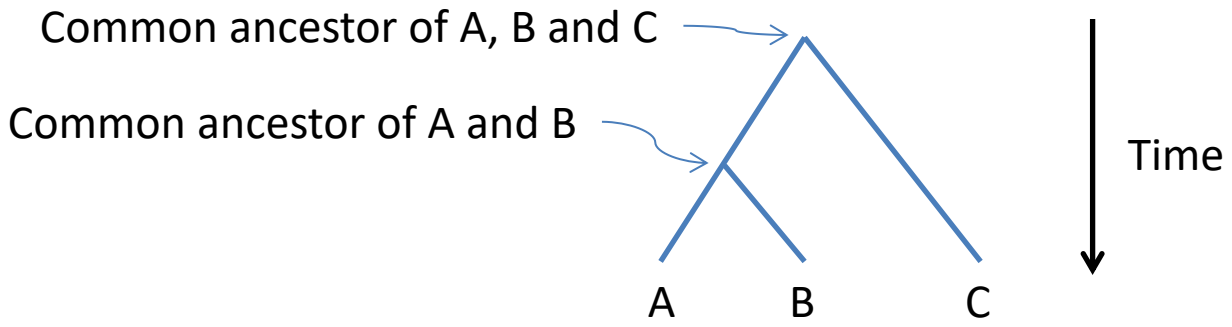
- A systematic and objective way to construct these trees is by comparing DNA/protein sequences

- In these two lectures, we study trees that relate objects sufficiently different
  - Different species
  - Different strains/populations of a species

- Our goal is to reconstruct the actual evolutionary relationships based on observable sequences

Assumptions

- Basic assumptions behind phylogenetic trees:
  1. The current sequences share a common ancestor
  2. All were mutated from the common ancestor
  3. Mutations are rare. Therefore, if the DNA of A and B are more similar than both A and C as well as B and C, likely C was separated from A and B before their separation



Terminology

- A tree is an acyclic graph with nodes connected by edges

- A phylogenetic tree is a binary tree with sequences (nodes) connected by branches (edges)
  - Leaf nodes are the observed sequences
  - Internal nodes are the unobserved ancestral sequences
  - The root node is the common ancestor of all the observed sequences
  - Branch lengths may represent evolutionary distances

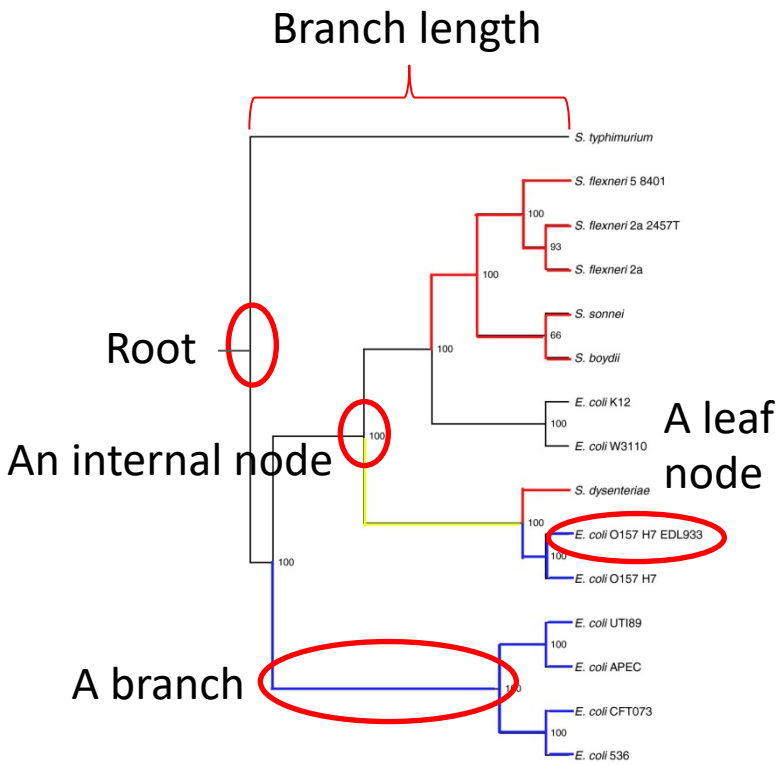


Image credit: Hershberg et al., Genome Biology 8:R164 (2007) , http://www.jdrf.ca/

Rooted and unrooted trees

- Sometimes it is not very clear where the common ancestor should be put
  - We can have a tree without root – an unrooted tree

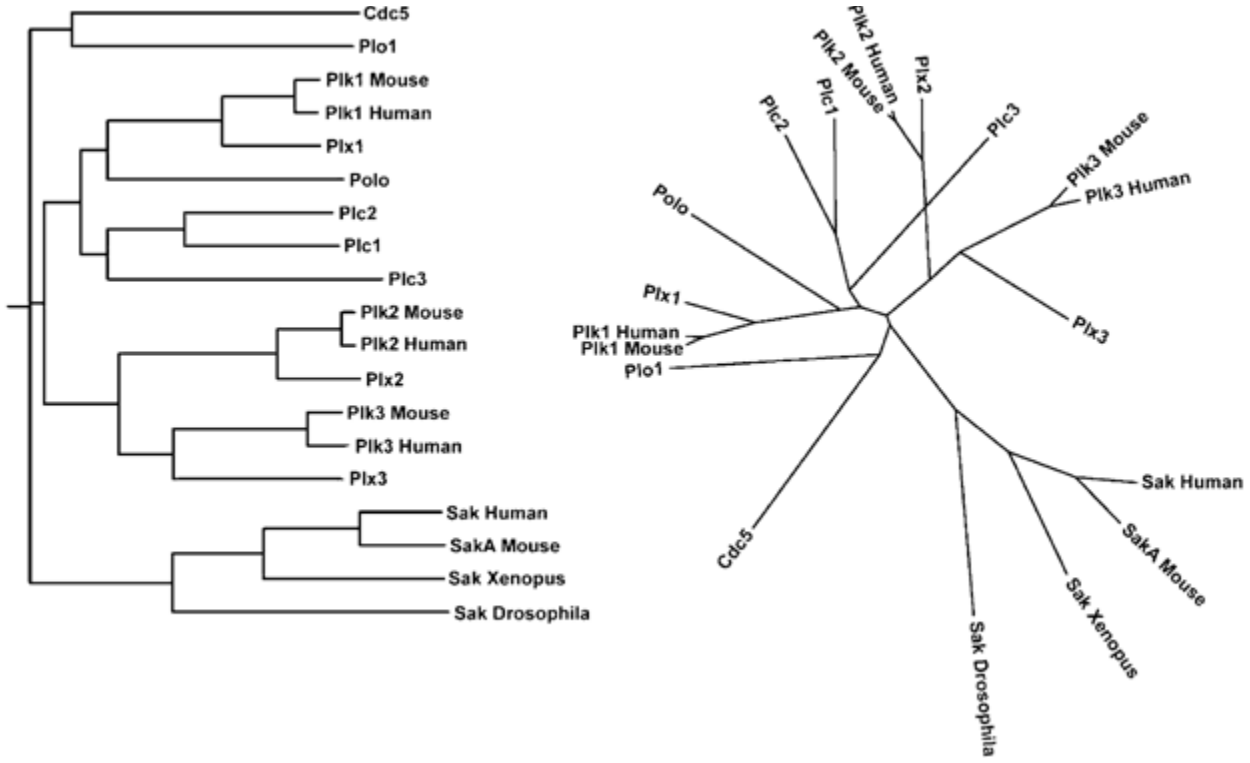


Image credit: Lowery et al., *Oncogene* 24(2):248-259, (2005)

Common file formats for phylogenetic trees

- Newick (nested brackets, with distances)
- NEXUS (giving short IDs to sequences, with more metadata)
- PhyloXML (using XML's structure)

The Newick format

- Use brackets and comma to group two sub-trees
- Use colon to indicate distance to parent, if available
- End with a semicolon

Graphical representation:

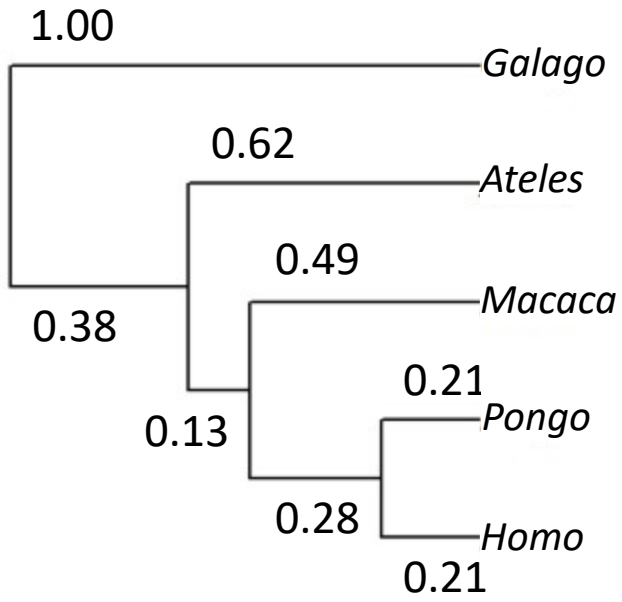


Image credit: http://www.zoology.ubc.ca/~schluter/zoo502stats/Rtips.phylogeny.html

Newick:

((((Homo:0.21,Pongo:0.21):0.28, Macaca:0.49):0.13,Ateles:0.62): 0.38,Galago:1.00);

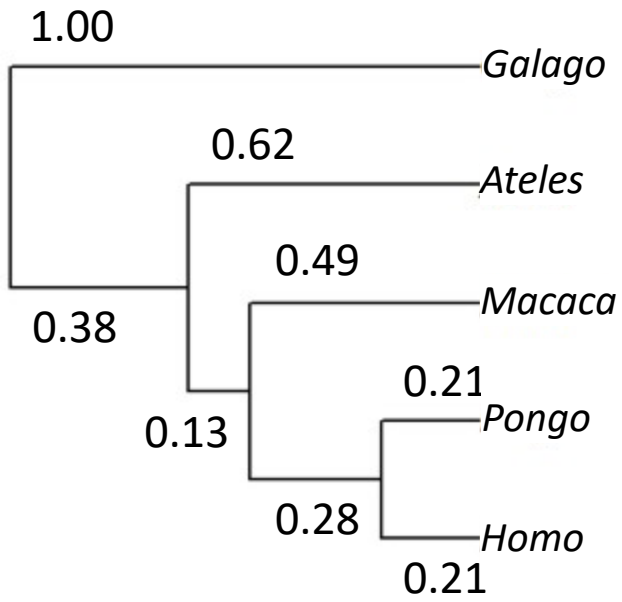
Remarks:

- For an unrooted tree, one simple way to represent it using the Newick format is to root the tree arbitrarily
- You can name internal nodes by giving the label after the close bracket (e.g., (Homo:0.21, Pongo:0.21)HP:0.28

Common file formats

- Example

Graphical representation:



Newick:

((((Homo:0.21,Pongo:0.21):0.28,Macaca:0.49):
0.13,Ateles:0.62):0.38,Galago:1.00);

NEXUS:

```
BEGIN TAXA;
    DIMENSIONS NTAX = 5;
    TAXLABELS
        Homo
        Pongo
        Macaca
        Ateles
        Galago
    ;
END;
BEGIN TREES;
    TRANSLATE
        1    Homo,
        2    Pongo,
        3    Macaca,
        4    Ateles,
        5    Galago
    ;
    TREE * UNTITLED = [&R]
((((1:0.21,2:0.21):0.28,3:0.49):0.13,4:0.62):0.38,5:1);
END;
```

Image credit: http://www.zoology.ubc.ca/~schluter/zoo502stats/Rtips.phylogeny.html

Common file formats

- PhyloXML

```
<phyloxml xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://www.phyloxml.org" xsi:schemaLocation="http://www.phyloxml.org
http://www.phyloxml.org/1.10/phyloxml.xsd">
    <phylogeny rooted="true">
        <name>Alcohol dehydrogenases</name>
        <description>contains examples of commonly used elements</description>
        <clade>
            <events>
                <speciations>1</speciations>
            </events>
            <clade>
                <taxonomy>
                    <id provider="ncbi">6645</id>
                    <scientific_name>Octopus vulgaris</scientific_name>
                </taxonomy>
                <sequence>
                    <accession source="UniProtKB">P81431</accession>
                    <name>Alcohol dehydrogenase class-3</name>
                </sequence>
            </clade>
            ...
        </clade>
    </phylogeny>
</phyloxml>
```

Information source: http://www.phyloxml.org/examples_syntax/phyloxml_syntax_example_1.html

Epilogue

Case Study, Summary and Further Readings

Case study: Unexpected classifications

- In the old days, biologists classified species based on their high-level features
  - If a species possesses features that make the organisms similar to multiple other types of species, it could be difficult to classify
  - When molecular features (e.g., DNA sequences) become available, they can be used to classify species in a systematic way
    - Some previous classifications were found to be inconsistent with molecular evidence

Case study: Unexpected classifications

- Example 1: Mammals
  - –Bats look like birds, dolphins look like fish, but both are actually mammals
- Kingdom: Animalia (animals)
  - Superphylum: Deuterostomia
    - Phylum: Chordata
      - Subphylum: Vertebrata (animals with backbones)
        - Infraphylum: Gnathostomata (jawed vertebrates)
          - Class: Chondrichthyes (cartilaginous fish)
          - Superclass: Osteichthyes (bony fish)
          - Superclass: Tetrapoda (four-limbed vertebrates)
            - Class: Aves (birds)
            - Class Mammalia (mammals)

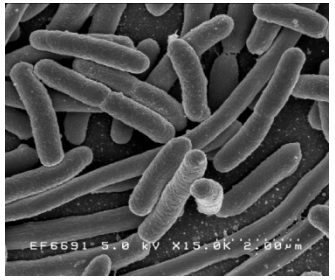
Image source: Wikipedia

Case study: Unexpected classifications

- Example 2: The three domains
- All species on earth belong to one of the three domains
  - Archaea
    - Single-celled, no nucleus
    - Usually live in places with extreme conditions (e.g., high temperature or salinity – “extremophiles”)
  - Bacteria
    - Single-celled, no nucleus
  - Eukaryote
    - Many are multi-celled, with nucleus

Image source: Wikipedia

Halobacteria sp. strain NRC-1, an archaaeon



Escherichia coli, a beacterium

Various eukaryotic species

Case study: Unexpected classifications

- It seems reasonable to assume that eukaryotes separated from the other two first
- However, based on the sequence of ribosomal RNAs, something so important that evolve slowly, archaea are closer to eukaryotes than bacteria

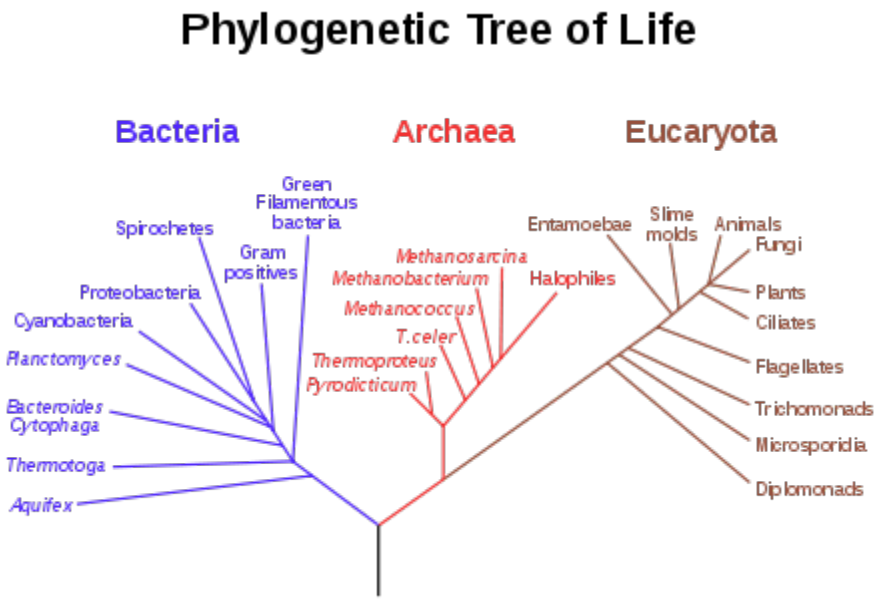


Image source: Wikipedia

Summary

- Mutation models allow us to formally estimate number of mutations happened based on observed data
  - Jukes-Cantor one parameter model
  - Kimura's two parameter model
  - PAM and BLOSUM matrices

- Phylogenetic trees capture separation events and when they happened
- Common file formats for trees

Further readings

- Chapter 3 of the book “Fundamentals of Molecular Evolution (Second Edition)” by Dan Graur and Wen-Hsiung Li. Sinauer Associates, Inc., 2000
  - Other mutation models
  - Models for coding sequences

Further readings

- The 1000 Genomes Project Consortium, A Global Reference for Human Genetic Variation. *Nature* 526(7571):68-74, (2015)
  - The 1000 Genomes Project aims at studying genetic differences among different human populations.
  - This paper is one of the latest reports from this consortium.