Topics

• Evolution mechanisms through mutations
  – Population genetics
  – Nucleic acid sequence evolution
    • Evolutionary distance vs. sequence distance
    • Jukes-Cantor model
Population Genetics

• Evolution:
  – Changes in the frequency as well as the sequence of genes in a population observed across time
  – Heritable changes in a population over many generations
  – …

• Two essential components:
  – Error-prone self replication produces genetic variants
  – Different variants incur varying levels of success at self replication through selection
    • Molecular evolution involves natural selection; selection carried out by nature
    • Unnatural selection; or artificial selection by humans forms the basis of agriculture
      – juicier and sweeter fruits
      – bigger and disease resistant crops
      – dogs and other animals bred selectively to fulfill different tasks
Case in Point: Dogs and Birds

- Dogs differ widely in their size and appearance, but belong to the same species
  - many years of selective breeding is responsible for all dog varieties
- Birds of prey look very similar but belong to different species

Source: http://www.dogbreedslist.info/all-dog-breeds/
Source: https://www.thespruce.com/types-of-birds-of-prey-387307
Nucleic Acid Sequences and Evolutionary History

- Organisms with common evolutionary ancestors share similar genetic sequences
  - At the time of genetic bifurcation, the two daughter species embark on different evolutionary paths
  - These different paths are characterized by the accumulation of different mutations
- The differences between their genetic sequences observed at the present time are related to the time of the bifurcation from the common ancestor
  - The earlier the separation, the higher the number of accumulated differences
  - The fraction of differences between sequences related to the evolutionary distances through mutation models
› Estimation of the evolutionary relationship among a given set of genetic sequences from different organisms
**Spread of Mutations**

- **An organism’s fitness**: The ability to leave descendants in future generations
  - The greater the number of descendants, the higher the fitness
    - Has little to do with the health or the general well being of an organism
    - Has more to do with how beneficial its traits are in the organism’s specific environment to leave descendants
- **Mutations can have three types of effects on the fitness**:  
  - Advantageous: Increase the chance of leaving descendants  
  - Neutral: No perceivable change in fitness  
  - Deleterious: Decrease the chance of leaving descendants
Genetic Variation Between Species

• Evolution traces out **ancestors** and **descendants**
  – Common ancestors of different species from which they have diverged some time in the distant past
  – Some evolutionary tracts lead to survival
  – Other tracts disappear into extinction
    • Evolution is competition between alternative genetic configurations
    • Species that get outcompeted by others die out
Tree of Life

Source: http://biologicalphysics.iop.org/cws/article/lectures/47042
Genetic Divergence Mechanism

- Changing environmental conditions work on the genetic variations within an ancestral species to create and shape the descendants
  - The descendants start off in the same species with slightly different genetic makeup
  - Time enhances the differences that allow exploiting different environmental niches
  - Eventually different species become “discernable”
Mutation Models on Nucleic Acid Sequences

• Genetic variations are characterized by differences in the gene sequences
  – Identical genes imply nearly identical organisms (up to chance effects from the environment)
  – Differences between organisms and species imply differences in their genes
• Quantification of these differences require stochastic models of nucleic acid sequence evolution
• These models also link sequence differences to evolutionary distances in units of evolutionary time
  – in terms of the nearest common ancestor in the evolutionary past
Modeling Nucleic Acid Substitutions

- **Objective**: to derive the relationship between the observed substitutions on different sequences and their evolutionary correspondence
  - Evolutionary correspondence refers to the amount of time in which the sequences went down independent evolutionary paths

- **Premises**:
  - Substitutions occur randomly
  - Fixation is assumed to have been…
    - achieved when comparing sequences of different species
    - not achieved when comparing sequences across individuals
  - Rates of substitution are constant for the sequences involved during the corresponding time period

- **Approach**:
  - Establish a functional relationship between a sequence distance and the corresponding evolutionary distance

\[
\text{evolutionary distance (in time units)} = \mathcal{F}(\text{sequence distance})
\]
Modeling Nucleic Acid Substitutions

• **Sequence difference** $D$:
  – Measured by the *fraction of nucleotides that are different* between two nucleic acid sequence fragments
  – Correlates linearly with the evolutionary time span for small time periods, but varies nonlinearly for large time periods
  – Can be measured quantitatively for any given two sequences simply by counting the number of sites where the sequences do not match
    • Hamming distance in coding theory

• **Evolutionary distance** $d$:
  – Measured by the *average number of substitutions* that have occurred per site between the two sequences during the time span of independent evolution
  – Correlates linearly with the time span of independent evolution for all time ranges, small AND large
  – Cannot be measured directly but can be inferred from $D$ using a stochastic model
Modeling Nucleic Acid Substitutions

- **Visible substitutions:**
  - One sequence remains the same and the other incurs a substitution, or
  - Both sequences incur substitutions into different nucleotides

- **Invisible substitutions:**
  - Neither sequence incurs a substitution (i.e., the original nucleotide remains preserved/conserved in both sequences), or
  - Both sequences incur substitutions into the same nucleotide

- **Annulled substitutions:**
  - Successive substitutions in both sequences result in the same nucleotide

\[ D = \frac{4}{20} \]

**Example:***

- **Original Sequence:** CCACGAGTCCACCGCAGCAC
- **Modified Sequence:** CCACGAGTCCACCGCAGC

Both sequences incur substitutions into the same nucleotide.
The Jukes-Cantor Model

- The substitution phenomenon is modeled by a Markov chain.
- In the Jukes-Cantor model, the rate of substitution from one base to any other is denoted by $\alpha$, in number of substitutions per unit time.
  - Thus, the net rate of change of a base is $3\alpha$.
  - $\alpha \ll 1$
- The corresponding state transition rate matrix is given by:

$$Q = \begin{bmatrix}
-3\alpha & \alpha & \alpha & \alpha \\
\alpha & -3\alpha & \alpha & \alpha \\
\alpha & \alpha & -3\alpha & \alpha \\
\alpha & \alpha & \alpha & -3\alpha
\end{bmatrix}$$
The Jukes-Cantor Model

- The resulting transition probability matrix is

\[
P(t) = e^{Qt} = \begin{bmatrix}
\frac{1}{4} + \frac{3}{4} e^{-4\alpha t} & \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} & \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} & \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} \\
\frac{1}{4} & \frac{1}{4} + \frac{3}{4} e^{-4\alpha t} & \frac{1}{4} & \frac{1}{4} \\
\frac{1}{4} - \frac{1}{4} e^{-4\alpha t} & \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} & \frac{1}{4} & \frac{1}{4} \\
\frac{1}{4} - \frac{1}{4} e^{-4\alpha t} & \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} & \frac{1}{4} & \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}
\end{bmatrix}
\]

or, more simply,

\[
P_{i,j}(t) = \begin{cases}
\frac{1}{4} + \frac{3}{4} e^{-4\alpha t} \quad & \text{if } i = j \\
\frac{1}{4} - \frac{1}{4} e^{-4\alpha t} \quad & \text{otherwise}
\end{cases}
\]
The Jukes-Cantor Model

• Note that $P_{i,j}(t)$ represents the probability with which the $i$’th nucleotide occupying a specific site on the original DNA sequence will be replaced by the $j$’th nucleotide in $t$ units of time.

• This allows calculating the average sequence difference between the original sequence and the evolving sequence as the expected value

$$D(t) = \sum_{i,j} 1(i \neq j)P_{i,j}(t)\pi_i$$

• Assuming an equal rate of nucleotides across the DNA, i.e. $\pi_i = 1/4$ for all $i = 1,2,3,4$, we get

$$D(t) = 12 \left( \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} \right) \frac{1}{4} = \frac{3}{4} - \frac{3}{4} e^{-4\alpha t}$$

• In addition, the incurred evolutionary distance by the evolving sequence to the original sequence is given by

$$d(t) = \sum_{i,j} 1(i \neq j)Q_{i,j}t\pi_i = 3\alpha t$$
The Jukes-Cantor Model

• To relate the observed sequence distance $D$ between two evolved sequences to the evolutionary distance $d$ between them:
  – the first sequence incurs $3\alpha t$ from the original
  – the second sequence incurs another $3\alpha t$ from the original, independent of the substitutions of the first one
  – this implies a total evolutionary distance of
    $$d(t) = 6\alpha t$$
  between the independently evolving sequences
  – furthermore, a combined evolution time of $2t$ produces a sequence distance of
    $$D(t) = \frac{3}{4} - \frac{3}{4} e^{-8\alpha t}$$
  – solving for the two in terms of $\alpha t$, we get
    $$D = \frac{3}{4} - \frac{3}{4} e^{-\frac{4}{3}d} \quad \text{or} \quad d = -\frac{3}{4} \log \left( 1 - \frac{4}{3}D \right)$$
The Jukes-Cantor Model

\( \alpha_1 < \alpha_2 < \alpha_3 \)
Example: Slow Evolution of a Single Sequence
Example: Fast Evolution of a Single Sequence
Example: Simultaneous Evolution of Two Sequences
# Alternative Models

### Jukes-Cantor

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>*</td>
<td>α</td>
<td>α</td>
<td>α</td>
</tr>
<tr>
<td>G</td>
<td>α</td>
<td>*</td>
<td>α</td>
<td>α</td>
</tr>
<tr>
<td>C</td>
<td>α</td>
<td>α</td>
<td>*</td>
<td>α</td>
</tr>
<tr>
<td>T</td>
<td>α</td>
<td>α</td>
<td>α</td>
<td>*</td>
</tr>
</tbody>
</table>

### HKY

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>*</td>
<td>απ_G</td>
<td>βπ_C</td>
<td>βπ_T</td>
</tr>
<tr>
<td>G</td>
<td>απ_A</td>
<td>*</td>
<td>βπ_C</td>
<td>βπ_T</td>
</tr>
<tr>
<td>C</td>
<td>βπ_A</td>
<td>βπ_G</td>
<td>*</td>
<td>απ_T</td>
</tr>
<tr>
<td>T</td>
<td>βπ_A</td>
<td>βπ_G</td>
<td>απ_C</td>
<td>*</td>
</tr>
</tbody>
</table>

### Kimura 2-parameter

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>*</td>
<td>α</td>
<td>β</td>
<td>β</td>
</tr>
<tr>
<td>G</td>
<td>α</td>
<td>*</td>
<td>β</td>
<td>β</td>
</tr>
<tr>
<td>C</td>
<td>β</td>
<td>β</td>
<td>*</td>
<td>α</td>
</tr>
<tr>
<td>T</td>
<td>β</td>
<td>β</td>
<td>α</td>
<td>*</td>
</tr>
</tbody>
</table>

### General Reversible

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>*</td>
<td>α_{A→G}</td>
<td>α_{A→C}</td>
<td>α_{A→T}</td>
</tr>
<tr>
<td>G</td>
<td>α_{G→A}</td>
<td>*</td>
<td>α_{G→C}</td>
<td>α_{G→T}</td>
</tr>
<tr>
<td>C</td>
<td>α_{C→A}</td>
<td>α_{C→G}</td>
<td>*</td>
<td>α_{C→T}</td>
</tr>
<tr>
<td>T</td>
<td>α_{T→A}</td>
<td>α_{T→G}</td>
<td>α_{T→C}</td>
<td>*</td>
</tr>
</tbody>
</table>
Variable Substitution Rates

• The Jukes-Cantor model as well as the more sophisticated ones assume that all sites along the DNA are equally prone to base substitutions
  – $P_{i,j}(t)$ is assumed to be the same regardless of the position of the nucleotide on the sequence
• This assumption simplifies the analysis, but does not exactly hold in reality
  – Some sites are structurally or functionally important, and evolve more slowly
    • Due to strong selective pressure
    • Some very important sites are practically invariant
Variable Substitution Rates

- Relaxing this assumption requires incorporating site-specific variation in observed differences
  - Jukes-Cantor model with a fixed fraction $q$ of invariable sites:
    \[
    d = -\frac{3}{4} (1 - q) \log \left( 1 - \frac{4D}{3 - 3q} \right)
    \]
  - Jukes-Cantor model where the variability of sites is governed by a gamma distribution:
    \[
    d = \frac{3}{4} a \left( \left( 1 - \frac{4}{3} D \right)^{-1/a} - 1 \right)
    \]
    where $a$ is the shape parameter of the gamma distribution governing the probability of a site being subject to a substitution rate of $r$, described by the probability density function
    \[
    f_R(r; a) = Z r a^{-1} e^{-ar}
    \]
Example: Evolutionary Siblinghood

• Data
  – A random “original” nucleic acid sequence $SQ$ of length $N = 100$ nucleotides undergoing point mutations according to a Jukes-Cantor model
  – Molecular evolution carried out in silico for 1000 epochs
    • $SQ^{(k)}$: The evolved sequence at the $k$’th epoch
    • $SQ^{(0)} = SQ$ (the original sequence)
    • $SQ_0 = SQ^{(1000)}$
  – A total of 5 sibling sequences, $SQ_1$, $SQ_2$, $SQ_3$, $SQ_4$ and $SQ_5$ identified as
    • $SQ_1^{(0)} = SQ^{(0)}$, $SQ_1^{(1000)} = SQ_1$
    • $SQ_2^{(0)} = SQ^{(200)}$, $SQ_2^{(800)} = SQ_2$
    • $SQ_3^{(0)} = SQ^{(400)}$, $SQ_3^{(600)} = SQ_3$
    • $SQ_4^{(0)} = SQ^{(600)}$, $SQ_4^{(400)} = SQ_4$
    • $SQ_5^{(0)} = SQ^{(800)}$, $SQ_5^{(200)} = SQ_5$
  evolved independently through the remaining epochs.

• Procedure:
  – Compute the sequence distances $D_{0,j}$ between $SQ_0$ and $SQ_1$, $SQ_2$, $SQ_3$, $SQ_4$ and $SQ_5$
  – Calculate the evolutionary distances $d_{0,j}$ from $D_{0,j}$ using the Jukes-Cantor model
Example: Sequence Data

$s_0$, $s_1$, $s_2$, $s_3$, $s_4$, $s_5$
Example: Sequence Data

\[ SQ_0 \]
AGTACCCGGGCGCATCGAAG...

\[ SQ_1 \]
ATTTCGGGCTCGAGATCGAAT...

\[ SQ \]
ATTACCCGGTTCGAGGGAAG...

\[ SQ_2 \]
ATTACCCGGTCGATCGATG...

\[ SQ_3 \]
AGTACACGGCAATCGAGG...

\[ SQ_4 \]
AGCAACCGTGCCCATCGAAG...

\[ SQ_5 \]
AGTACCTGCGGCCCATCGAAG...
Example: Evolutionary Distances

- **Sequence distances:**
  - \( D_{0,1} = 0.4900 \Rightarrow d_{0,1} = 0.7945 \)
    - AGTACC GGGGG CGATCGAAG...
    - 1 1 11 11 1...
    - ATTT CCCG TCGAG ATCGAAT...
  - \( D_{0,2} = 0.3400 \Rightarrow d_{0,2} = 0.4529 \)
    - AGTACCC GGGGC CATCGAAG...
    - 1 11 11 1...
    - ATTA CCCG TGGCGA AGGGAAG...
  - \( D_{0,3} = 0.2300 \Rightarrow d_{0,3} = 0.2747 \)
    - AGTACCC GGGCC ATCGAAG...
    - 1 1 1 1 1...
    - AGTACAC GTGCA ATCGAAG...
  - \( D_{0,4} = 0.1700 \Rightarrow d_{0,4} = 0.1928 \)
    - AGTACC CCGGG CGATCGAAG...
    - 1 1 1 1 1...
    - AGCA ACGT GCCGATCGAAG...
  - \( D_{0,5} = 0.0900 \Rightarrow d_{0,5} = 0.0959 \)
    - AGTACCC GGGGC CATCGAAG...
    - 1 1 1...
    - AGTACCT GCGGG CGATCGAAG...
Repeat Example: Sequence Data

$sq$
Repeat Example: Evolutionary Distances

- Distances:
  \[ D_{0,1} = 0.4400 \Rightarrow d_{0,1} = 0.6626 \]
  \[ D_{0,2} = 0.3000 \Rightarrow d_{0,2} = 0.3831 \]
  \[ D_{0,3} = 0.3000 \Rightarrow d_{0,3} = 0.3831 \]
  \[ D_{0,4} = 0.1500 \Rightarrow d_{0,4} = 0.1674 \]
  \[ D_{0,5} = 0.0800 \Rightarrow d_{0,5} = 0.0846 \]

- Remark:
  - Even though the experiment setup is exactly the same, the distances vary
    - Sequence evolution is a stochastic process
  - The variation even produces a rather strange and quite disagreeable result:
    \[ D_{0,2} = D_{0,3} \text{ providing } d_{0,2} = d_{0,3} !!! \]
Example: Variability in Computed Evolutionary Distances

![Histograms of d_{01} to d_{05}](image)

Each histogram represents the distribution of evolutionary distances for different comparisons.
Remarks

• Models of nucleic acid sequence evolution links the sequence differences to evolutionary distances
• The parameters of these models are fitted to the available data to capture reality as much as possible
  – More sophisticated models better fit the available data
  – With better fits, the risk of losing general validity increases
• The viability of these models depends on the validity of the premises on the given application data
  – Assumptions may not hold
• The estimated evolutionary distances, however, are subject to estimation errors
  – These errors may switch the order of evolutionary siblinghood
• The extent of errors are directly proportional to the expected evolutionary distances
  – For sequences that are similar, the expected error is small
  – For sequences that are significantly different, the errors are large