CSE 527
Phylogeny & RNA: Pfold

Lectures 20-21
Autumn 2006

Phylogenies
(aka Evolutionary Trees)

“Nothing in biology makes sense, except in the light of evolution”
-- Dobzhansky

Modeling Sequence Evolution

Simple but useful models; assume:
- Independence of separate positions
- Independence of separate lineages
- Stationarity - e.g., nuc freqs aren’t changing
- Markov property - nuc at a given position is independent of nuc there \( t_2 \) years ago given nuc there \( t_1 < t_2 \) years ago.

Simple Example: Jukes-Cantor

Rate matrix \( R \):

\[
\begin{array}{cccc}
A & C & G & T \\
\hline
A & -3a & a & a & a \\
C & a & -3a & a & a \\
G & a & a & -3a & a \\
T & a & a & a & -3a \\
\end{array}
\]

Consequences:
- Equilibrium nuc freqs \( \pi_t \), all = 1/4
- All changes equally likely

Rate of \( C \rightarrow T \) changes per unit time
Diagonal s.t. row sums = 0
Multiplicativity

Matrix $P_{[i,j]}$: prob of change $i \rightarrow j$ in time $t$

$P_{s+t[i,j]} = \sum_k P_{s[i,k]} P_{t[k,j]}$

I.e.,

$P_{s+t} = P_s P_t$

Finding Change Probabilities

For small time $\varepsilon$, transition probabilities

$P_{t} \approx I + \varepsilon R$

By multiplicativity

$P_{t+\varepsilon} = P_{t} P_{\varepsilon} \approx P_{t} (I + \varepsilon R)$

$(P_{t+\varepsilon} - P_{t}) / \varepsilon \approx P_{t} R$

I.e., solve system of diff eqns:

$\frac{d}{dt} P' = P'R$

Jukes-Cantor, cont.

Solving $\frac{d}{dt} P' = P'R$

Gives $P' = \begin{bmatrix} r & s & s & s \\ s & r & s & s \\ s & s & r & s \\ s & s & s & r \end{bmatrix}$

where

$r = (1+3 \exp(-4at))/4$

$s = (1 - \exp(-4at))/4$

Other Models

Jukes-Cantor is simple, but inaccurate for some uses. E.g.,

Many genomes deviate sharply from $\pi_i = 1/4$

In fact, “transversions”

(purine {A,G} ↔ pyrimidine {C,T})

less frequent than “transitions”

(pur ↔ pur or pyr ↔ pyr).

Various other models often used
General Reversible Model

Model is reversible if for all $i, j$

$\pi_i P[i,j] = \pi_j P[j,i]$

I.e., $i \rightarrow j$ and $j \rightarrow i$ changes are equally frequent; statistically, the past looks like the future

No closed form solution for but numerically solvable using eigenvalues of rate matrix $R$

Evolutionary Models: Key points

Given small number of parameters (e.g., $4 \times 4$ symmetric rate matrix, ...), an evolutionary tree, and branch lengths, you can calculate probabilities of changes on the tree.

Uses: Example 1

Probability of changes shown on this (given) tree:

$P(t_1, G \rightarrow G) \times P(t_2, G \rightarrow T)$

Uses: Example 2

What if ancestral state unknown?

$\sum \pi_a P(t_1, a \rightarrow G) \times P(t_2, a \rightarrow T)$

draw $a$ at root from equilibrium distribution
**Uses: Example 3**

What if sequences at leaves and ancestral sequence unknown?

\[ \prod_{u=1}^{\sigma} \sum_{a_u} \pi_{a_u} P(t_1, a_u \rightarrow x_u^1) P(t_2, a_u \rightarrow x_u^2) \]

**Uses: Example 4**

What if branch lengths also unknown?

Can find MLE by numerical optimization of

\[ \arg \max_{t_1, t_2} \prod_{u=1}^{\sigma} \sum_{a_u} \pi_{a_u} P(t_1, a_u \rightarrow x_u^1) P(t_2, a_u \rightarrow x_u^2) \]

**Uses: Example 5**

What if Tree also unknown?

Can try MLE of tree topology, too (>> parsimony)

---

*Figure 8.3* The log likelihood \( P(x^1, x^2|T, t_1, t_2) \) given by (8.9), with \( n_1 = 100, n_2 = 250 \), and with \( n_1 = 1000, n_2 = 2500 \). The latter curve is sharper, as there are more data to define the maximum likelihood peak. The curves have been shifted so their peaks superimpose.

Reversible model; you can't place root.
A Complex Question:
Given data (sequences, anatomy, ...) infer the phylogeny

A Simpler Question:
Given data and a phylogeny, evaluate “how much change” is needed to fit data to tree

Parsimony
General idea ~ Occam’s Razor: If change is rare, prefer explanations requiring few changes

Human  A T G A T ...
Chimp  A T G A T ...
Gorilla A T G A G ...
Rat    A T G C G ...
Mouse  A T G C T ...

0 changes

Parsimony
General idea ~ Occam’s Razor: If change is rare, prefer explanations requiring few changes

Human  A T G A T ...
Chimp  A T G A T ...
Gorilla A T G A G ...
Rat    A T G C G ...
Mouse  A T G C T ...

1 change
**Parsimony**

General idea ~ Occam’s Razor: If change is rare, prefer explanations requiring few changes

- Human: A T G A T ...
- Chimp: A T G A T ...
- Gorilla: A T G A G ...
- Rat: A T G C G ...
- Mouse: A T G C T ...

2 changes

**Likelihood**

Given a statistical model of evolutionary change, prefer the explanation of *maximum likelihood*

- Human: A T G A T ...
- Chimp: A T G A T ...
- Gorilla: A T G A G ...
- Rat: A T G C G ...
- Mouse: A T G C T ...

**Sankoff & Rousseau, ‘75**

\[ P_u(s) = \text{best parsimony score of subtree rooted at node } u, \text{ assuming } u \text{ is labeled by character } s \]

**Sankoff-Rousseau Recurrence**

\[ P_u(s) = \text{best parsimony score of subtree rooted at node } u, \text{ assuming } u \text{ is labeled by character } s \]

For leaf \( u \):

\[ P_u(s) = \begin{cases} 0 & \text{if } u \text{ is a leaf labeled } s \\ \infty & \text{if } u \text{ is a leaf not labeled } s \end{cases} \]

For internal node \( u \):

\[ P_u(s) = \min_{t \in \text{child}(u)} \text{cost}(s, t) + P_v(t) \]

Time: \( O(\text{alphabet}^2 \times \text{tree size}) \)
So, Parsimony easy; What about Likelihood?

Straightforward generalization of “simple” formula for 2-leaf tree

\[
\prod_{v \in T} \sum_{a_v} \pi_v P(t_v, a_v \rightarrow x_v^1) P(t_v, a_v \rightarrow x_v^2)
\]

is infeasible, since you need to consider all (exponentially many) labelings of non-leaf nodes. Fortunately, there’s a better way…

\[ L_u(s | \theta) = \text{Likelihood of subtree rooted at node } u, \text{ assuming } u \text{ is labeled by character } s, \text{ given } \theta \]

For Leaf \( u \):

\[ L_u(s | \theta) = \begin{cases} 1 & \text{if } u \text{ is a leaf labeled } s \\ 0 & \text{if } u \text{ is a leaf not labeled } s \end{cases} \]

For Internal node \( u \):

\[ L_u(s | \theta) = \prod_{v \in \text{child}(u)} \sum_{t \in \{A,C,G,T\}} P(s \rightarrow t | \text{length}(u,v), \theta) \cdot L_v(t | \theta) \]

Felsenstein Recurrence

Another Application: RNA folding

Using Evolution for RNA Folding

Assume you have

1. Training set of trusted RNA alignments
   - build evo model for unpaired columns
   - build evo model for paired columns
2. Alignment (\& tree) for some RNAs presumed to have an (unknown) common structure
   - look at every col pair - better fit to paired model or 2 indp unpaired models?

(Alternative to mutual information, using evo)
Training Data

Trusted alignments of 1968 tRNAs + 305 LSU rRNAs

Table 1. Base frequencies, showing nearly equal overall distribution of bases, with a slight underrepresentation of Cs. Stems have high GC/CG base pair frequencies, while loops have low content of Cs and Gs. The lowest row shows the distribution of bases between loops and stems.

<table>
<thead>
<tr>
<th>Stem/Loop</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>AU/AU</td>
<td>35.6%</td>
</tr>
<tr>
<td>GC/GC</td>
<td>53.4%</td>
</tr>
<tr>
<td>UG/GU</td>
<td>9.8%</td>
</tr>
<tr>
<td>Other</td>
<td>1.2%</td>
</tr>
<tr>
<td>Total</td>
<td>52.6%</td>
</tr>
</tbody>
</table>

Rate Matrix (Unpaired)

Table 2. The entries, r_{XY}, for the loop rate matrix. Transitions are more frequent than transversions.

<table>
<thead>
<tr>
<th>X\Y</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-0.75</td>
<td>0.16</td>
<td>0.32</td>
<td>0.26</td>
</tr>
<tr>
<td>C</td>
<td>0.40</td>
<td>-1.57</td>
<td>0.24</td>
<td>0.93</td>
</tr>
<tr>
<td>G</td>
<td>0.55</td>
<td>0.17</td>
<td>-0.96</td>
<td>0.24</td>
</tr>
<tr>
<td>U</td>
<td>0.35</td>
<td>0.51</td>
<td>0.19</td>
<td>-1.05</td>
</tr>
</tbody>
</table>

Rate Matrix (Paired)

Table 3. Some of the entries for the stem rate matrix. Only rates between the six most frequent base pairs are shown.

<table>
<thead>
<tr>
<th>X\Y</th>
<th>AU</th>
<th>UA</th>
<th>GC</th>
<th>CG</th>
<th>UG</th>
<th>GU</th>
</tr>
</thead>
<tbody>
<tr>
<td>AU</td>
<td>0.18</td>
<td>0.18</td>
<td>0.50</td>
<td>0.12</td>
<td>0.02</td>
<td>0.27</td>
</tr>
<tr>
<td>UA</td>
<td>0.18</td>
<td>0.16</td>
<td>0.12</td>
<td>0.50</td>
<td>0.27</td>
<td>0.02</td>
</tr>
<tr>
<td>GC</td>
<td>0.33</td>
<td>0.08</td>
<td>-0.82</td>
<td>0.13</td>
<td>0.02</td>
<td>0.23</td>
</tr>
<tr>
<td>CG</td>
<td>0.08</td>
<td>0.33</td>
<td>0.13</td>
<td>-0.82</td>
<td>0.23</td>
<td>0.02</td>
</tr>
<tr>
<td>UG</td>
<td>0.08</td>
<td>1.00</td>
<td>0.10</td>
<td>1.26</td>
<td>-2.56</td>
<td>0.04</td>
</tr>
<tr>
<td>GU</td>
<td>1.00</td>
<td>0.08</td>
<td>1.26</td>
<td>0.10</td>
<td>0.04</td>
<td>-2.56</td>
</tr>
</tbody>
</table>

What about Gaps?

option 1: evo model for them
- hard & slow

option 2: treat “-” as a 5th character
- they don’t “evolve” quite like others

option 3: treat “-” as unknown
- ditto
- end up pairing?
  (drop if < 75%)
+ easy
Which Tree?

KH-99 : try to find MLE tree (using SCFG et al.)
good but slow
KH-03 : est tree without structure
  average unpaired & (marginalized) paired rates
calc pairwise distances between seqs
tree topology from “neighbor joining”
  MLE tree branch lengths

Synopsis of last lecture

Based on simplifying assumptions (stationarity, independence, Markov, reversible), there are simple sequence-evolution models with a modest number of parameters giving, e.g., $Pr(G\rightarrow T \mid 1.5 \text{ m yr})$. …
It can model base-pairing in RNA, too
Felsenstein allows ML estimation of probabilities, branch lengths, even trees,... in this model.
(Somewhat like “parsimony” algorithm, but better.)
Goal: Use all this for inference of RNA 2ary struct

Phylogeny vs Mutual Information

CCGUAGAUUA
CCGUAGAUUA
CCGUAGAUUA
CAGUAUAUA
CAGUAUAUA

MI = 1 bit in both cases, but green pair is more compelling evidence of interaction: 3 events, not 1

The Glue: An SCFG

a) $S \rightarrow LS \mid L$
$b) \rightarrow ssLssss \rightarrow ssddFdssss$
$c) \rightarrow ssddLdssss$
$d) \rightarrow ssddLdssss$

b) $S \rightarrow LS \mid L$
$b) \rightarrow ssLssss \rightarrow ssddFdssss$
$c) \rightarrow ssddLdssss$
$d) \rightarrow ssddLdssss$

F \rightarrow LS \mid dFd

S \rightarrow LS \mid L
Full SCFG

\[ S \rightarrow LS \quad (0.868534) \quad | \quad L \quad (0.131466) \]
\[ L \rightarrow s \quad (0.894603 \ast p(s)) \quad | \quad dFd \quad (0.105397 \ast p(dd)) \]
\[ F \rightarrow LS \quad (0.212360) \quad | \quad dFd \quad (0.787640 \ast p(dd)) \]

Where \( p(s) \) & \( p(dd) \) are the probabilities of the single/paired alignment columns \( s/dd \) as calculated by the Felsenstein algorithm based on the fixed evolutionary model and the given tree topology and branch lengths.

What SCFG Gives

“Prior” probabilities for fraction paired vs unpaired lengths of each frequency of bulges in stems etc., and

Inherits column probabilities from evo model

Cocke-Kasami-Younger for CFG

Suppose all rules of form \( A \rightarrow BC \) or \( A \rightarrow a \)
(by mechanical grammar transform, or use orig grammar & mechanically transform alg below...)

Given \( x = x_1...x_n \), want \( M_{ij} = \{ A \mid A \rightarrow x_{i+1}...x_j \} \)

For \( j=2 \) to \( n \)

\[ M[i-1,j] = \{ A \mid A \rightarrow x_j \text{ is a rule} \} \]

for \( i = j-1 \) down to \( 1 \)

\[ M[i,j] = \bigcup_{i < k < j} M[i,k] \otimes M[k,j] \]

Where \( X \otimes Y = \{ A \mid A \rightarrow BC, B \in X, \text{and } C \in Y \} \)

The “Inside” Algorithm for SCFG

(analogous to HMM “forward” alg)

Just like CKY, but instead of just recording possibility of \( A \) in \( M[i,j] \), record its probability:
For each \( A \), do sum instead of union, over all possible \( k \) and all possible \( A \rightarrow BC \) rules, of products of their respective probabilities.

Result: for each \( i, j, A \), have \( \Pr(A \rightarrow x_{i+1}...x_j) \)

(There’s also an “outside” alg, analogous to backward...)

10
The “Viterbi” algorithm for SCFGs

Just like inside, but use max instead of sum.

So what’s the structure?

The usual dynamic programming traceback:
Starting from $S$ in upper right corner of matrix, find which $k$ and which $S \rightarrow BC$ gave max probability, then (recursively) find where that $B$ and that $C$ came from…

(Really, you want to do it with the $F \rightarrow dFd$ grammar, and where those rules are used tells you where the base pairs are.)

Results & Validation
KH-99: 4 bacterial RNAse P, 340-380 nt

1: *Klebsiella pneumoniae*

2: *Serratia marscens*

3: *Pseudomonas fluorescens*

4: *Thiobacillus ferrooxidans*

![Fig. 2: The phylogenetic tree relating the four analysed sequences, as calculated using the ML estimation described above. The length units correspond to the rate matrices of the model.](image)

Good overall structure prediction

*Klebsiella pneumoniae* RNAse P
Good Overall Structure Prediction

1. More sequences help
2. So do phylogeny and a good alignment

Not bad, even with only one seq

Table 7. Accuracy table, showing comparisons of single sequence predictions using the method described in this paper and Mfold Version 3.0, by Zuker (1989) and Zuker et al. (1994). Predictions of secondary structures were made on single sequences, which is the only possibility using Mfold. The average results are comparable.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>SCFG method</th>
<th>Mfold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq 1</td>
<td>57.7%</td>
<td>67.1%</td>
</tr>
<tr>
<td>Seq 2</td>
<td>48.2%</td>
<td>54.0%</td>
</tr>
<tr>
<td>Seq 3</td>
<td>41.2%</td>
<td>35.6%</td>
</tr>
<tr>
<td>Seq 4</td>
<td>46.2%</td>
<td>50.3%</td>
</tr>
<tr>
<td>Average</td>
<td>48.3%</td>
<td>51.7%</td>
</tr>
</tbody>
</table>

Fig. 3. The alignment of the four HIV-1 RNA sequences. The predicted structure, using all four sequences, is shown. The structure for the database is depicted with square brackets denoting parts of the helix. The square brackets must begin with a helix terminating point, with the lines defining the structure. The square brackets drawn points where the structure differs from the structure at the top and bottom. The lines denote the positions where the structure differs from the structure at the top and bottom.
Results & Validation

KH-03

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: 9 tmRNAs (363.8)</td>
<td>act, act, bac, inf, kle, pon, pas, mul, sal, par, sal, lyp, ase, put, wb, cho, yar, pes</td>
</tr>
<tr>
<td>B: 13 bacterial SRP RNAs (270.5)</td>
<td>bac, alc, bac, bre, bac, cir, bac, mac, bac, meg, bac, pol, bac, pum, bac, ph, bac, ste, bac, thu, bre, bre, clo, per</td>
</tr>
<tr>
<td>C: 10 eukaryotic SRP RNAs (300.9)</td>
<td>ory, sat, tri, ae-a, tri, ae-b, zea, ma-a, zea, ma-b, zea, ma-c, zea, ma-d, zea, ma-e, zea, ma-f, zea, ma-h</td>
</tr>
<tr>
<td>D: 51 eukaryotic SRP RNAs (297.4)</td>
<td>ara, th-a, ara, th-b, cae, el-a, cae, el-b, cae, el-c, cae, el-d, can, spe, cin, hyb, dro, mel, fig, rub, hom, sa-a, hom, sa-b, hom, sa-c, hum, ja, hum, ja-b, hum, la, hum, la-b, hum, lac, hum, la-c, lap, col, lyc, es-a, lyc, es-b, lyc, es-c, lyc, es-d, lyc, es-e, lyc, es-f, lyc, es-g, lyc, es-h, lyc, es-i, lyc, es-j, lyc, es-k, lyc, es-l, lyc, es-m, lyc, es-n, lyc, es-o, ory, sat, rat, rat, sch, pom, tct, ros, tet, the, tri, ae-a, tri, ae-b, tri, br-a, tri, br-b, xen, lae, yar, li-a, yar, li-b, zea, ma-a, zea, ma-b, zea, ma-c, zea, ma-d, zea, ma-e, zea, ma-f</td>
</tr>
</tbody>
</table>

Figure 6. Accuracy vs number of sequences used in the prediction. Crosses: ‘correct’ alignments, boxes: ClustalW alignments. Each point: average results for either all possible combinations or 50 random combinations, whichever is the lower.

Figure 7. Accuracy as a function of pairwise distance between two sequences being analysed. As in Figure 6, crosses are from results using ‘correct’ alignments, while boxes are from ClustalW alignments. The pairs were grouped according to their Jukes-Cantor distances, in the intervals [0.0, 0.2], [0.2, 0.4], [0.4, 0.6] etc. The points represent average results for 50 random sequence combinations from a specific range of distances. The x-value of a point is the average of the 50 distances.

Course Wrap Up
Course Project Presentations

Wednesday, 12/13, 1:00-2:30
CSE 674

Everyone’s invited

"High-Throughput BioTech"

Sensors
DNA sequencing
Microarrays/Gene expression
Mass Spectrometry/Proteomics
Protein/protein & DNA/protein interaction
Controls
Cloning
Gene knock out/knock in
RNAi

Floods of data
"Grand Challenge" problems

CS/Math/Stats Points of Contact

Scientific visualization
Gene expression patterns
Databases
Integration of disparate, overlapping data sources
Distributed genome annotation in face of shifting underlying coordinates
AI/NLP/Text Mining
Information extraction from journal texts with inconsistent nomenclature, indirect interactions, incomplete/inaccurate models,…
Machine learning
System level synthesis of cell behavior from low-level heterogeneous data (DNA sequence, gene expression, protein interaction, mass spec, Algorithms
…

Frontiers & Opportunities

New data:
Proteomics, SNP, arrays CGH, comparative sequence information, methylation, chromatin structure, ncRNA, interactome

New methods:
Graphical models? rigorous filtering?
Data integration
many, complex, noisy sources

Systems Biology
Frontiers & Opportunities

Open Problems:
- splicing, alternative splicing
- multiple sequence alignment (genome scale, w/ RNA etc.)
- protein & RNA structure
- interaction modeling
- network models
- RNA trafficking
- ncRNA discovery
- ...

Exciting Times

Lots to do
Various skills needed
I hope I’ve given you a taste of it

Thanks!