Week 3: Parsimony variants, compatibility, statistics and parsimony

Genome 570

April, 2008
Camin-Sokal parsimony

0 → 1

4 changes

Week 3: Parsimony variants, compatibility, statistics and parsimony – p.2/43
Dollo parsimony

4 changes
(3 losses)
Polymorphism parsimony

5 retentions of polymorphism
nonadditive binary coding

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>a</td>
<td>b</td>
<td>c</td>
<td>d</td>
<td>e</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Week 3: Parsimony variants, compatibility, statistics and parsimony – p.5/43
Dollo parsimony – a paradox

Week 3: Parsimony variants, compatibility, statistics and parsimony – p.6/43
Successive weighting

There are 15 possible unrooted trees, which fall into 5 types according to how many changes they have in each character. The table shows the total weighted number of changes when each tree type is evaluated using the weights implied by the 5 different tree types.

<table>
<thead>
<tr>
<th>number of trees</th>
<th>have pattern of changes:</th>
<th>type of tree</th>
<th>tree type used for weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1,1,1,2,2,1)</td>
<td>I</td>
<td>2.333 2.250 2.167 2.417 2.083</td>
</tr>
<tr>
<td>2</td>
<td>(1,2,1,2,2,1)</td>
<td>II</td>
<td>2.667 2.500 2.500 2.667 2.333</td>
</tr>
<tr>
<td>2</td>
<td>(2,1,2,2,2,1)</td>
<td>III</td>
<td>3.000 2.917 2.667 2.917 2.583</td>
</tr>
<tr>
<td>3</td>
<td>(2,2,2,1,1,1)</td>
<td>IV</td>
<td>2.833 2.667 2.500 2.500 2.333</td>
</tr>
<tr>
<td>7</td>
<td>(2,2,2,2,2,1)</td>
<td>V</td>
<td>3.333 3.167 3.000 3.167 2.833</td>
</tr>
</tbody>
</table>
Ties in successive weighting

An example of successive weighting that would show the difficulty it has in detecting ties. The table shows the total weighted number of changes when each tree type is evaluated using the weights implied by the different tree types.

<table>
<thead>
<tr>
<th>number of trees</th>
<th>have pattern of changes:</th>
<th>type of tree</th>
<th>tree type used for weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1,1,2,2)</td>
<td>I</td>
<td>1.667 1.833 1.5</td>
</tr>
<tr>
<td>1</td>
<td>(2,2,1,1)</td>
<td>II</td>
<td>1.833 1.667 1.5</td>
</tr>
<tr>
<td>1</td>
<td>(2,2,2,2)</td>
<td>III</td>
<td>2.333 2.333 2</td>
</tr>
</tbody>
</table>
Data example for compatibility

The data set Table 1.1 with an added species all of whose characters are 0.

<table>
<thead>
<tr>
<th>Species</th>
<th>Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Alpha</td>
<td>1</td>
</tr>
<tr>
<td>Beta</td>
<td>0</td>
</tr>
<tr>
<td>Gamma</td>
<td>1</td>
</tr>
<tr>
<td>Delta</td>
<td>1</td>
</tr>
<tr>
<td>Epsilon</td>
<td>0</td>
</tr>
<tr>
<td>Omega</td>
<td>0</td>
</tr>
</tbody>
</table>
The compatibility matrix for this data set
The Pairwise Compatibility Theorem

A set $S$ of characters has all pairs of characters compatible with each other if and only if all of the characters in the set are jointly compatible (in that there exists a tree with which all of them are compatible).
The compatibility graph

Maximal cliques? Largest?
One of the cliques it contains
A maximal clique
The largest maximal clique
Making the tree by popping

Week 3: Parsimony variants, compatibility, statistics and parsimony – p.16/43
Making the tree by popping

Character 1
Character 2
Character 3

Week 3: Parsimony variants, compatibility, statistics and parsimony – p.17/43
Making the tree by popping

Week 3: Parsimony variants, compatibility, statistics and parsimony – p.18/43
Making the tree by popping

Week 3: Parsimony variants, compatibility, statistics and parsimony – p.19/43
Unknown states and compatibility

A data set that has all pairs of characters compatible, but which cannot have all characters compatible with the same tree. This violates the Pairwise Compatibility Theorem, owing to the unknown ("?") states.

<table>
<thead>
<tr>
<th></th>
<th>Alpha</th>
<th>Beta</th>
<th>Gamma</th>
<th>Delta</th>
<th>Epsilon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Beta</td>
<td>?</td>
<td>0</td>
<td>1</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Gamma</td>
<td>1</td>
<td>?</td>
<td>0</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>Delta</td>
<td>0</td>
<td>1</td>
<td>?</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>Epsilon</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Multiple character states and compatibility

Fitch’s set of nucleotide sequences that have each pair of sites compatible, but which are not all compatible with the same tree.

<table>
<thead>
<tr>
<th></th>
<th>Alpha</th>
<th>Beta</th>
<th>Gamma</th>
<th>Delta</th>
<th>Epsilon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
</tbody>
</table>
Transition probabilities in a two-state model

\[
\text{Prob} \left( 1 \mid 0, t_j, r_i \right) = \frac{1}{2} \left( 1 - e^{-2r_it_j} \right)
\]
Likelihood and parsimony

\[ L = \text{Prob} (\text{Data}|\text{Tree}) \]

\[ = \prod_{i=1}^{\text{chars}} \sum_{\text{recon-structions}} \left( \frac{1}{2} \prod_{j=1}^{B} \begin{cases} r_{ij} t_{ij} & \text{if this character changes} \\ 1 - r_{ij} t_{ij} & \text{if it does not change} \end{cases} \right) \]
approximating ...

For each character the likelihood is approximately

\[ \frac{1}{2} \prod_{i=1}^{B} (r_i t_j)^{n_{ij}}. \]

For them all it is approximately

\[ L \approx \prod_{i=1}^{\text{chars}} \left( \frac{1}{2} \prod_{j=1}^{\text{branches}} (r_i t_j)^{n_{ij}} \right). \]

Tossing out the 1/2 terms and taking logarithms:

\[ -\ln L \approx \sum_{i=1}^{\text{chars}} \sum_{j=1}^{\text{branches}} n_{ij} (-\ln (r_i t_j)). \]
Weights and likelihood

Probabilities of change and resulting weights for an imaginary case.

<table>
<thead>
<tr>
<th>Character</th>
<th>$r_i t_j$</th>
<th>changes</th>
<th>total weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.01</td>
<td>1</td>
<td>4.605</td>
</tr>
<tr>
<td>2</td>
<td>0.01</td>
<td>2</td>
<td>9.210</td>
</tr>
<tr>
<td>3</td>
<td>0.00001</td>
<td>1</td>
<td>11.519</td>
</tr>
</tbody>
</table>
with one-tenth as much change ...

<table>
<thead>
<tr>
<th>Character</th>
<th>$r_i t_j$</th>
<th>changes</th>
<th>total weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.001</td>
<td>1</td>
<td>6.908</td>
</tr>
<tr>
<td>2</td>
<td>0.001</td>
<td>2</td>
<td>13.816</td>
</tr>
<tr>
<td>3</td>
<td>0.000001</td>
<td>1</td>
<td>13.816</td>
</tr>
</tbody>
</table>
... and with one-tenth less again

<table>
<thead>
<tr>
<th>Character</th>
<th>$r_i t_j$</th>
<th>changes</th>
<th>total weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0001</td>
<td>1</td>
<td>9.210</td>
</tr>
<tr>
<td>2</td>
<td>0.0001</td>
<td>2</td>
<td>18.421</td>
</tr>
<tr>
<td>3</td>
<td>0.0000001</td>
<td>1</td>
<td>16.118</td>
</tr>
</tbody>
</table>
Trees, steps, and patterns

A
B
C
D

A
B
C
D

A
B
C
D

0000 0 0 0
0001 1 1 1
0010 1 1 1
0011 1 2 2
0100 1 1 1
0101 2 1 2
0110 2 2 1
0111 1 1 1
1000 1 1 1
1001 2 2 1
1010 2 1 2
1011 1 1 1
1100 1 2 2
1101 1 1 1
1110 1 1 1
1111 0 0 0

Week 3: Parsimony variants, compatibility, statistics and parsimony – p.28/43
Trees, steps, and patterns with DNA

Week 3: Parsimony variants, compatibility, statistics and parsimony – p.29/43
Parsimony and patterns

\[ n_{xxyy} + 2n_{xyxy} + 2n_{xyyx} = 2(n_{xxyy} + n_{xyxy} + n_{xyyx}) - n_{xxyy} \]

\[ 2n_{xxyy} + n_{xyxy} + 2n_{xyyx} = 2(n_{xxyy} + n_{xyxy} + n_{xyyx}) - n_{xyxy} \]

\[ 2n_{xxyy} + 2n_{xyxy} + n_{xyyx} = 2(n_{xxyy} + n_{xyxy} + n_{xyyx}) - n_{xyyx} \]
The counterexample

Week 3: Parsimony variants, compatibility, statistics and parsimony – p.31/43
Pattern probabilities

If tip pattern is 1100, and internal nodes are 1, 1 the probability is

\[ \frac{1}{2} (1 - p)(1 - q)(1 - q)pq \]

and in general summing over all four possibilities:

\[ P_{1100} = \frac{1}{2} ( (1 - p)(1 - q)^2 pq + (1 - p)^2(1 - q)^2 q \]

\[ + p^2 q^3 + pq(1 - p)(1 - q)^2) \]
Pattern probabilities

\[ P_{xyyy} = (1 - p)(1 - q)[q(1 - q)(1 - p) + q(1 - q)p] \]
\[ + pq[(1 - q)^2(1 - p) + q^2p] \]

\[ P_{xyxy} = (1 - p)q[q(1 - q)p + q(1 - q)(1 - p)] \]
\[ + p(1 - q)[p(1 - q)^2 + (1 - p)q^2] \]

\[ P_{xyyx} = (1 - p)q[(1 - p)q^2 + p(1 - q)^2] \]
\[ + p(1 - q)[q(1 - q)p + q(1 - q)(1 - p)] \]

Week 3: Parsimony variants, compatibility, statistics and parsimony – p.33/43
Taking differences

\[ P_{xyxy} - P_{xxyy} = (1 - 2q) \left[ q^2(1 - p)^2 + (1 - q)^2p^2 \right] \]

Which is always positive as long as \( q < 1/2 \) and either \( p \) or \( q \) is positive. Thus \( P_{xyxy} > P_{xxyy} \)

So that if

\[ P_{xxyy} - P_{xyxy} > 0 \]

To have \( P_{xxyy} \) be the largest of the three we need to have

\[ (1 - 2q) \left[ q(1 - q) - p^2 \right] > 0 \]

which is true if and only if

\[ q(1 - q) > p^2 \]
Conditions for inconsistency

\[
p^2 < q \cdot (1-q)
\]

consistent

inconsistent
Example for patterns with DNA

Week 3: Parsimony variants, compatibility, statistics and parsimony – p.36/43
Calculating a pattern frequency

\[
\text{Prob } [\text{CCAA}] = \frac{1}{18} (1 - p)(1 - q)^2 pq + \frac{1}{27}pq^2(1 - p)(1 - q) \\
+ \frac{1}{162} p^2q^2(1 - q) + \frac{7}{972} p^2q^3 \\
+ \frac{1}{12} (1 - p)^2(1 - q)^2q.
\]
Pattern frequencies

\[
\text{Prob [xxyy]} = (1 - p)^2 q (1 - q)^2 + \frac{2}{3} p (1 - p) q (1 - q)^2 \\
+ \frac{4}{9} p (1 - p) q^2 (1 - q) + \frac{2}{27} p^2 q^2 (1 - q) \\
+ \frac{7}{81} p^2 q^3
\]

\[
\text{Prob [xyxy]} = \frac{1}{3} (1 - p)^2 q^2 (1 - q) + \frac{2}{9} p (1 - p) q^2 (1 - q) \\
+ \frac{4}{27} p (1 - p) q^3 + \frac{1}{3} p^2 (1 - q)^3 \\
+ \frac{2}{9} p^2 q^2 (1 - q) + \frac{2}{81} p^2 q^3
\]

\[
\text{Prob [xyyx]} = \frac{1}{81} (1 - p)^2 q^3 + \frac{2}{3} p (1 - p) q (1 - q)^2 \\
+ \frac{4}{9} p (1 - p) q^3 + \frac{1}{9} p^2 q (1 - q)^2 \\
+ \frac{6}{27} p^2 q^2 (1 - q) + \frac{2}{81} p^2 q^3.
\]
Conditions for inconsistency with DNA

\[ p < \frac{-18q + 24q^2 + \sqrt{243q - 567q^2 + 648q^3 - 288q^4}}{9 - 24q + 32q^2} \]

For small \( p \) and \( q \) this is approximately

\[ \frac{1}{3} p^2 < q \]
Conditions for inconsistency with DNA

\[ q \]

\[ p \]

consistent

inconsistent

Week 3: Parsimony variants, compatibility, statistics and parsimony – p.40/43
Inconsistency with a clock

Week 3: Parsimony variants, compatibility, statistics and parsimony – p.41/43
Example showing inconsistency with a clock
This freeware-friendly presentation prepared with

- Linux (operating system)
- PDFLaTeX (mathematical typesetting and PDF preparation)
- ps2pdf to turn Postscript file into a PDF
- Idraw (drawing program to modify plots and draw figures)
- Adobe Acrobat Reader (to display the PDF in full-screen mode)