DIRECT OPTIMIZATION

Filip Högnabba
Botanical Museum
Finnish Museum of Natural History
DNA-sequences

- **OBSERVATIONS** of the nucleotide order in a studied sequence

```
1 ggattgcctc agtaacgcg agtgaagcgg caacagctca aatttgaaat ctggttcttt
cgggggcccg agttgtaatt tgtagaggat gtttcgggtg cggcgccggt ctaagttcct
tggaacagga cgtcatagag ggtgagaatc ccgtatgcga ctggcgatca accccatgtg
aaaccccttc gacgagtcga gttgtttggg aatgcagctc aaaatgggtg gtaaatttca
tctaaagcta aataccggcc agagaccgat agcgcacaag tagagtgatc gaaagatgaa
aagcactttg gaaagagagt caaacagtac gtgaaattgt tgaaagggaa gcgcttgcaa
ccagacctgc ccgcgggtga tcaaccgtcc ttcgttgggc ggtgcactcg cccgcgaaca
ggccagcatc ggttgggacg gccggataaa ggccttggga atgtagcttc ctccggggag
tcttatagcc cggggcgcca tgcggccagc cccgaccgag gaccgcgctt cggctaggat
gctggcgtaa tggttgtaag cgacccgtct tgaaacacgg accaaggagt ctaacatcta
tgcgagtgtt tgggtgtcaa acccatgcgc ggaatgaaag tgaacggagg tgggaacccg
cgagggcgca ccatcgaccg atccagatgt cttcggatgg atttgagtac gagcatagct
gttgggaccc gaaagatggg gaactatgcc tgaatagggt gaagccagag gaaactctgg
tggaggctcg cagcggttct gacgtgcaaa tcgatcgtca aatttgggta taggggcgaa
agactaatcg aaccatctag tagctggttc ctgccgaagt ttccctcagg atagcagtaa
cgttttcagt tttatgaggt aaagcgaatg attagaggcc ttggggttga aacaacctta
acctattctc aaactttaaa tatgtaagaa gtccttgtta cttagttgaa cgtggacatt
tgaatgcacc gttactagtg ggccattttt ggtaagcaga actggcgatg cgggatgaac
cgaacgcgag gttaaggtgc cggaatgcac gctcatcaga caccacaaaa agtgttaatt
catccagaca gccggacggt ggccatggaa gtcggaaacc gctaaggagt gtgtaacaac
tcaccggccg aatgaattag ccttgaaaat ggatggcgct caagcgtgct acccatacct
cgccgccagg gtagaaacga tgccct
```
From observations to phylogenetic hypotheses

DNA sequences

Phylogenetic hypotheses
From observations to phylogenetic hypotheses

DNA sequences

1) alignment
2) phylogenetic analyses

Phylogenetic hypotheses
Alignment

- homology assumptions on nucleotide level
- insertion of gaps to indicate nucleotide insertions or deletions (indels) - gaps are NOT OBSERVATIONS
- not always self-evident, particularly for non-coding sequences
Protein-coding DNA

- no length variation
- unambiguous homology assumptions possible
Non-coding DNA

- length variation present
- unambiguous homology assumptions not possible
Alignment

- critically important in phylogenetic studies
- a dataset aligned according to different criteria or indel treatments may support contradictory phylogenies
- no structural or developmental complexity to test nucleotide homology
- nucleotide homology can be evaluated only in reference to a topology
- static homology assumptions - may not be optimal for the final phylogenetic hypothesis
**Number of alignments**

\[ f(n,m) \quad 1 \leq n \leq 10; \quad 1 \leq m \leq 5 \]

<table>
<thead>
<tr>
<th>n\m</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>13</td>
<td>75</td>
<td>541</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>409</td>
<td>23917</td>
<td>2244361</td>
</tr>
<tr>
<td>3</td>
<td>63</td>
<td>16081</td>
<td>10681263</td>
<td>14638756721</td>
</tr>
<tr>
<td>4</td>
<td>321</td>
<td>699121</td>
<td>5552351121</td>
<td>117629959485121</td>
</tr>
<tr>
<td>5</td>
<td>1683</td>
<td>32193253</td>
<td>3147728203035</td>
<td>1.05 \times 10^{18}</td>
</tr>
<tr>
<td>10</td>
<td>8097453</td>
<td>9850349744182729</td>
<td>3.32 \times 10^{26}</td>
<td>1.35 \times 10^{38}</td>
</tr>
</tbody>
</table>

*IMPOSSIBLE* to find the *OPTIMAL* alignment even for a few short sequences

---

*Slowinski (1998)*
From observations to phylogenetic hypotheses

DNA sequences

1) alignment
2) phylogenetic analyses

Problematic

Phylogenetic hypotheses
From observations to phylogenetic hypotheses

DNA sequences

1) alignment
2) phylogenetic analyses

Direct optimization

Phylogenetic hypotheses

Wheeler (1996)
Direct optimization

- algorithm for optimization of sequence data on a tree by Sankoff (1975)
- homology assumptions and phylogenetic analyses made simultaneously
- unique homology assumptions for different topologies - dynamic homology assumptions
- indels treated equivalent to any other transformation
Direct optimization

TGCA

CGCA

TAAA

GAAT
Direct optimization

Rule I
if a character state is shared by the terminals it is marked for their common ancestor (interception)

Rule II
if terminals have different character states the union is marked for their common ancestor
Direct optimization

TGCA → CGCA

TGA = R
GT = K
GC = S
GAT = D
GAC = V
GTC = B
ATC = H
AT = W
AC = M
CT = Y
ACGT = N
Direct optimization

TGCA

YGCA

CGCA

TAAA

GAAT

three length: 5 changes

T G C A

C G C A

Y G C A

T A A A

G A A T

GA  = R
GT  = K
GC  = S
GAT = D
GAC = V
GTC = B
ATC = H
AT  = W
AC  = M
CT  = Y
ACGT = N
Direct optimization

c > t

A > G

A > C

g > t
t > a

three length: 5 changes

TGCA

CGCA

TAAA

GAAT

YGCA

TRMA

KAAW

GA = R
GT = K
GC = S
GAT = D
GAC = V
GTC = B
ATC = H
AT = W
AC = M
CT = Y
ACGT = N
Direct optimization

TGCA
- T > C

CGCA
- G > A
- C > A

TAAA
- T > G

GAAT

YGCA

TRMA
- T > G
- A > T

three length: 5 changes

GA = R
GT = K
GC = S
GAT = D
GAC = V
GTC = B
ATC = H
AT = W
AC = M
CT = Y
ACGT = N
Direct optimization

three length: 7 changes

| GA   | = R |
| GT   | = K |
| GC   | = S |
| GAT  | = D |
| GAC  | = V |
| GTC  | = B |
| ATC  | = H |
| AT   | = W |
| AC   | = M |
| CT   | = Y |
| ACGT | = N |
Direct optimization

AGC

AGCT

ATCA

ATC
Direct optimization

AGC
AGC

AGCT

AT

ATCA

ATC

AGC = R
GT = K
GC = S
GAT = D
GAC = V
GTC = B
ATC = H
AT = W
AC = M
CT = Y
ACGT = N
Direct optimization

AGC

AGCT

AGC(T/-)
Direct optimization

AGC

AGCT

ATCA

ATC

AGC(T/-)

e.g. gap cost = 2
substitution cost = 1

AGC (T/-)

ATC A

GA = R
GT = K
GC = S
GAT = D
GAC = V
GTC = B
ATC = H
AT = W
AC = M
CT = Y
ACGT = N
Direct optimization

AGC

AGC(T/-)

AGCT

AKC

ATCA

gap cost = substitution cost

ATC

GA = R
GT = K
GC = S
GAT = D
GAC = V
GTC = B
ATC = H
AT = W
AC = M
CT = Y
ACGT = N
Direct optimization

AGC

AGCT

ATCA

ATC

AGC(T/-)

AGC

AGC

ATC

ATC

GA = R
GT = K
GC = S
GAT = D
GAC = V
GTC = B
ATC = H
AT = W
AC = M
CT = Y
ACGT = N
Direct optimization

Rule III
if an internal node has one character state it will not be changed

Rule VI
if an internal node has two character states and the node preceding it had only one the intersection is given to the one with both character states
Direct optimization

**AGC**

**AGCT**

**ATCA**

**ATC**

**AGC(T/-)**

**AKC**

**ATC**

- **GA** = R
- **GT** = K
- **GC** = S
- **GAT** = D
- **GAC** = V
- **GTC** = B
- **ATC** = H
- **AT** = W
- **AC** = M
- **CT** = Y
- **ACGT** = N
Direct optimization

AGC

AGC(T/-)

AGCT

ATCA

ATC

ATC

ATC

GA = R
GT = K
GC = S
GAT = D
GAC = V
GTC = B
ATC = H
AT = W
AC = M
CT = Y
ACGT = N
Direct optimization

AGC
AGCT
ATCA
ATC

AGC(T/-)

ATC
ATC

GA = R
GT = K
GC = S
GAT = D
GAC = V
GTC = B
ATC = H
AT = W
AC = M
CT = Y
ACGT = N
Direct optimization

![Diagram showing the relationships between different nucleotide sequences with a legend for the mapping of nucleotide pairs to amino acids.]

Legend:
- GA = R
- GT = K
- GC = S
- GAT = D
- GAC = V
- GTC = B
- ATC = H
- AT = W
- AC = M
- CT = Y
- ACGT = N
Direct optimization

three length: 3 changes
Direct optimization in practice

- direct optimization is implemented in the program POY
POY

- implements classic heuristic tree search strategies, e.g. branch swapping, treedrifting, treefusing, and ratcheting
- heuristic procedures to optimize sequences on trees, Direct optimization and Fixed states is implemented and tightly integrated with the tree search heuristics
• computationally demanding
• various techniques needed to reduce calculation time
How to reduce calculation time?

- parallel computing - essential even for moderate sized data-sets
- cut sequences into shorter fragments in CONSERVED regions
- perform analyses in pieces - use intermediate results
Lets try!