Reconstructing Phylogenetic Trees: Distances, Parsimony, and Likelihood

September 25th 1, 2008

Parsimony

I. Counting evolutionary changes.

II. Searching for the most parsimonious tree.

  - Exploring tree space.
  - Sequence of taxon addition.

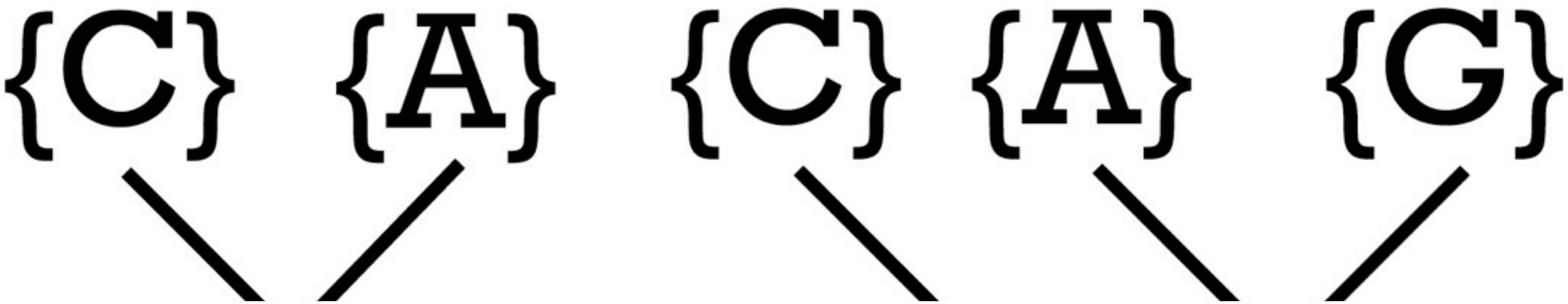
III. Performance of parsimony.

IV. Problems with parismony.

I. Counting the Number of Changes

- The Fitch/Wagner method.
- The Dollo method.

Fitch/Wagner Method

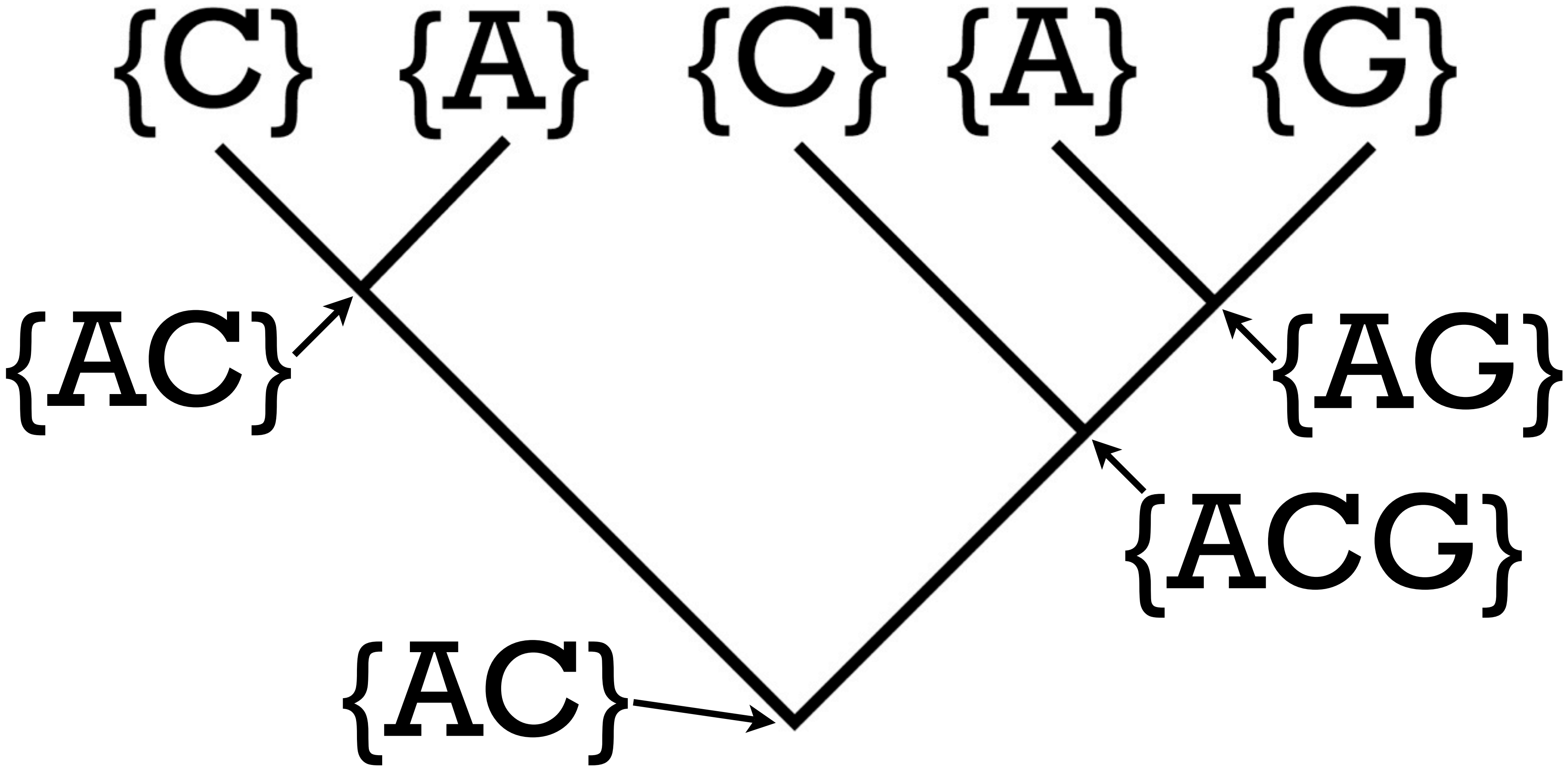


{C} {A} {C} {A} {G}

- Postorder tree transversal.
- At internal node, create intersection of descendent sets.
- If empty, create union of descendent sets.

Fitch/Wagner Method

Total Number of Changes = 3



I. Counting the Number of Changes

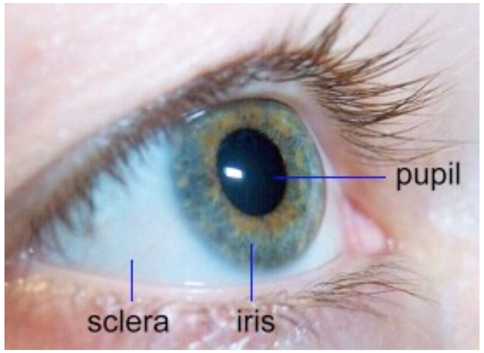
- The Fitch/Wagner method.
  - General model for reversible change.
- The Dollo method.
  - Model for characters that are considered more likely to evolve in one direction.

Dollo Method


Blind cave fish
Photo by Rick Olsen

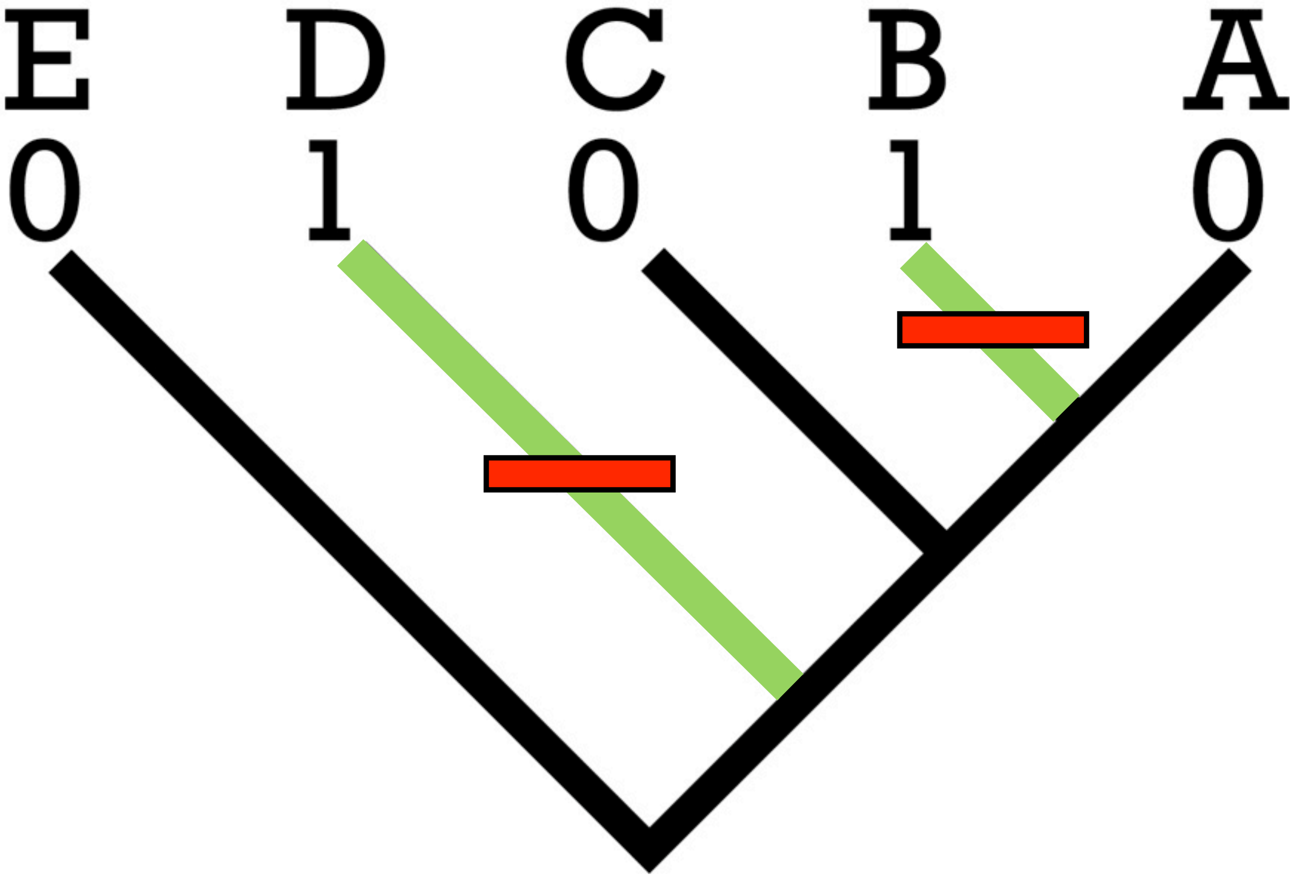
Eyeless Eye

$$0 \underset{\text{Easy}}{\overset{\text{Difficult}}{\rightleftarrows}} 1$$



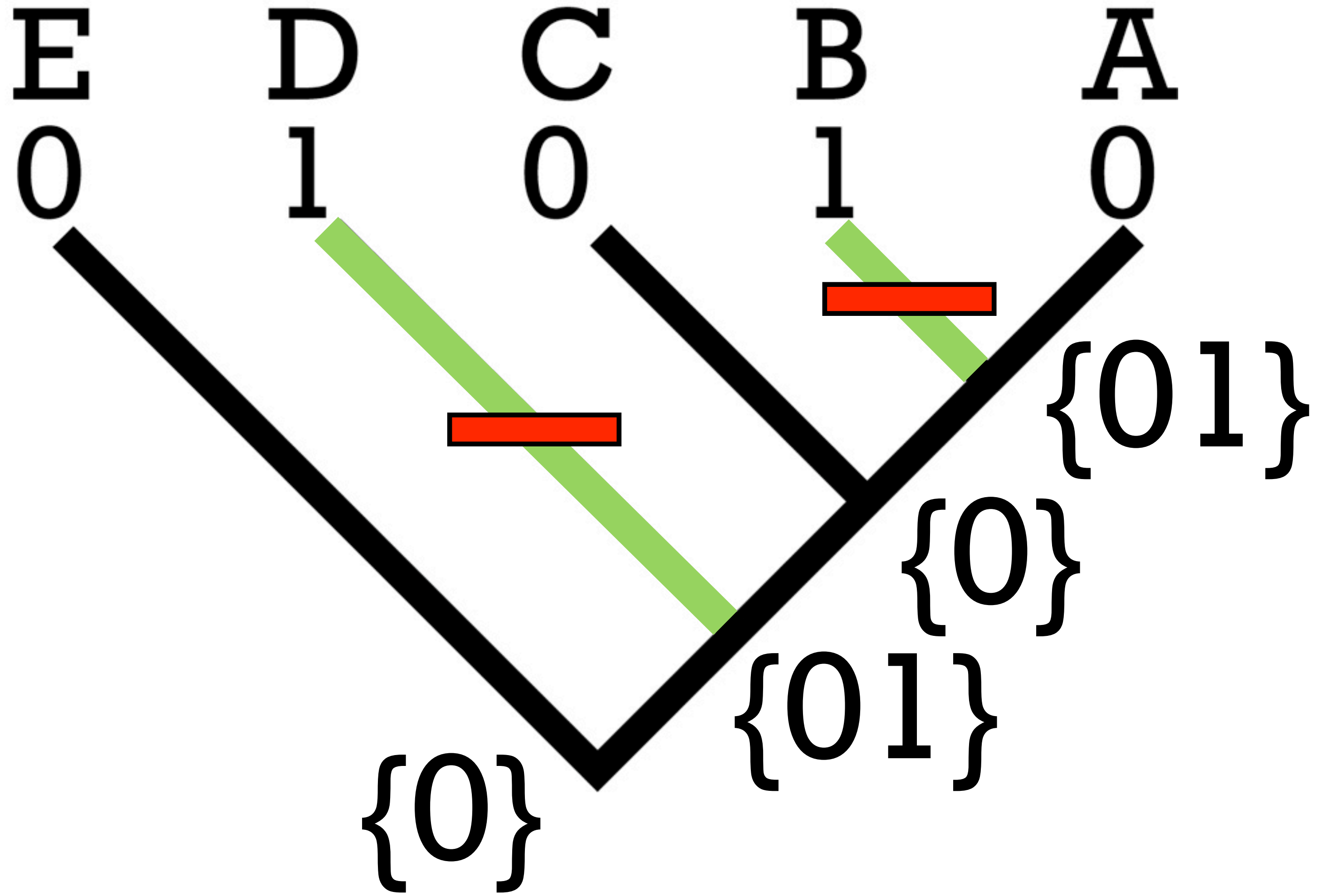
Dollo Method

Total Number of Changes = 2



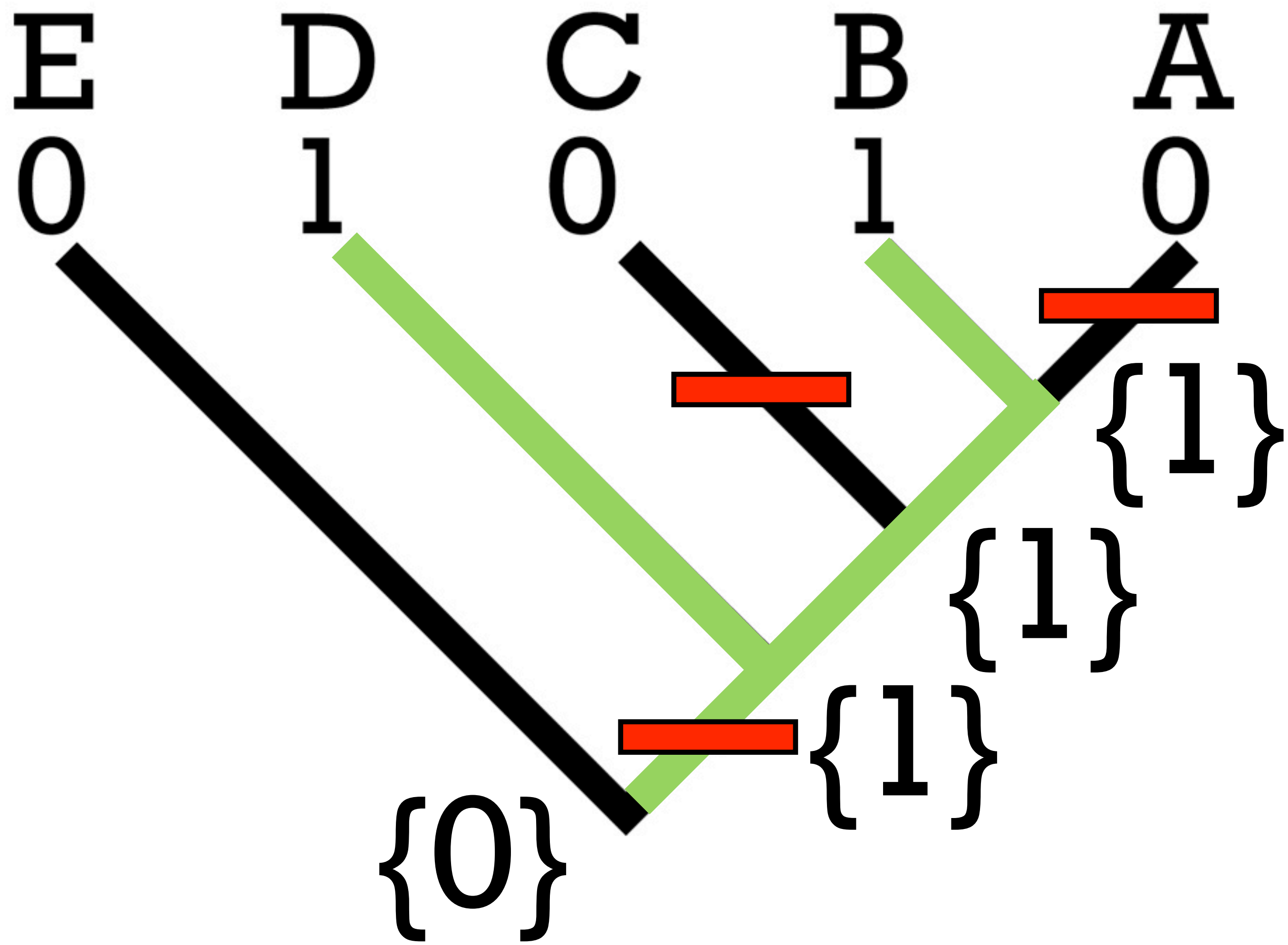
Dollo Method

Total Number of Changes = 2



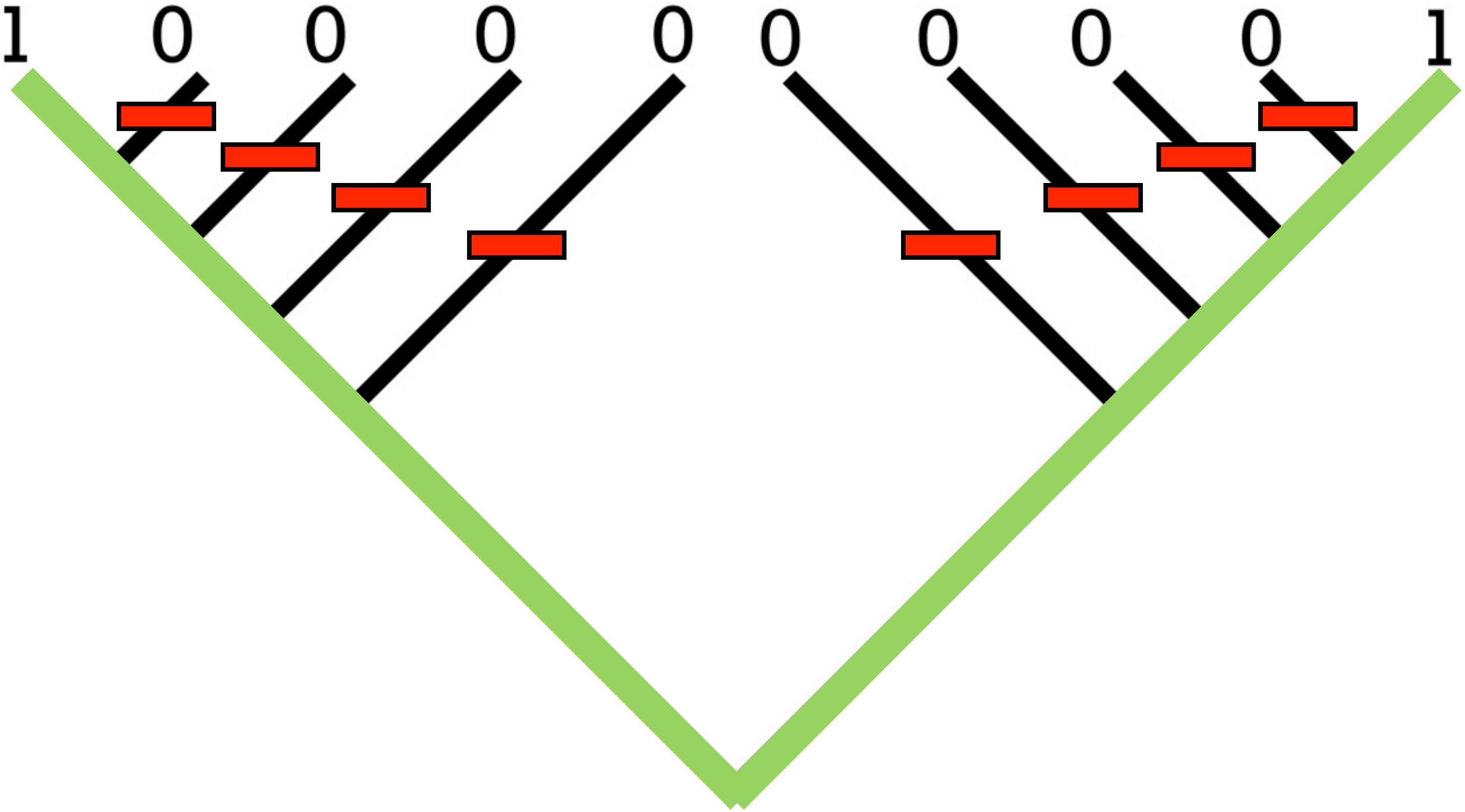
Dollo Method

Total Number of Changes = 3



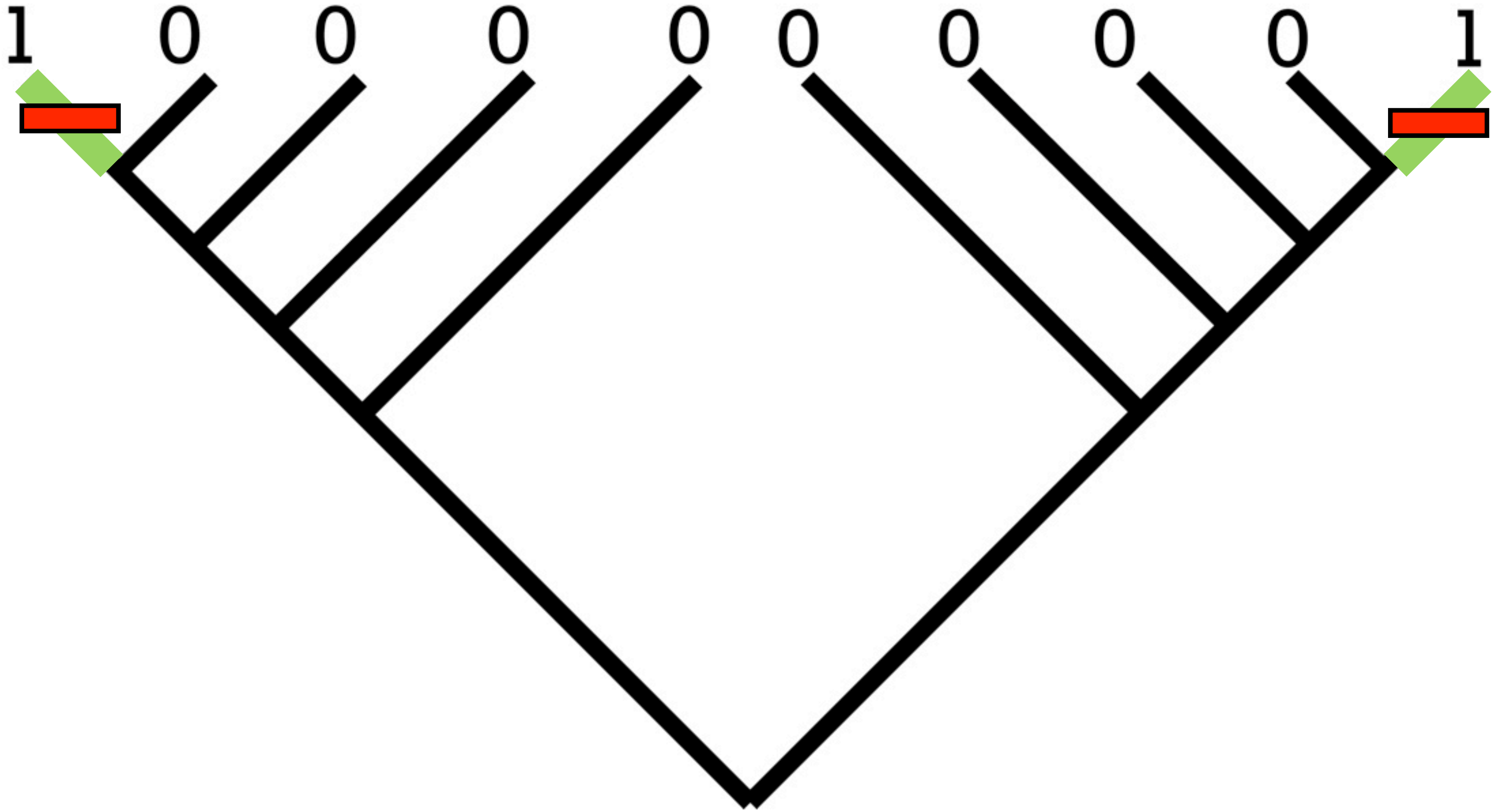
Dollo Method

Total Number of Changes = 8



Dollo Method

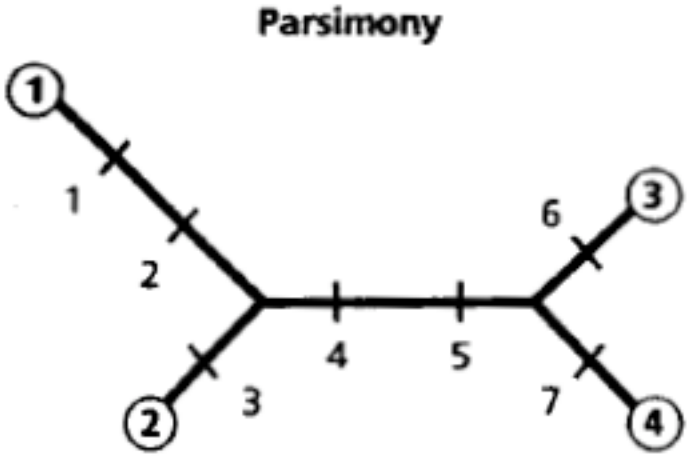
Total Number of Changes = 2

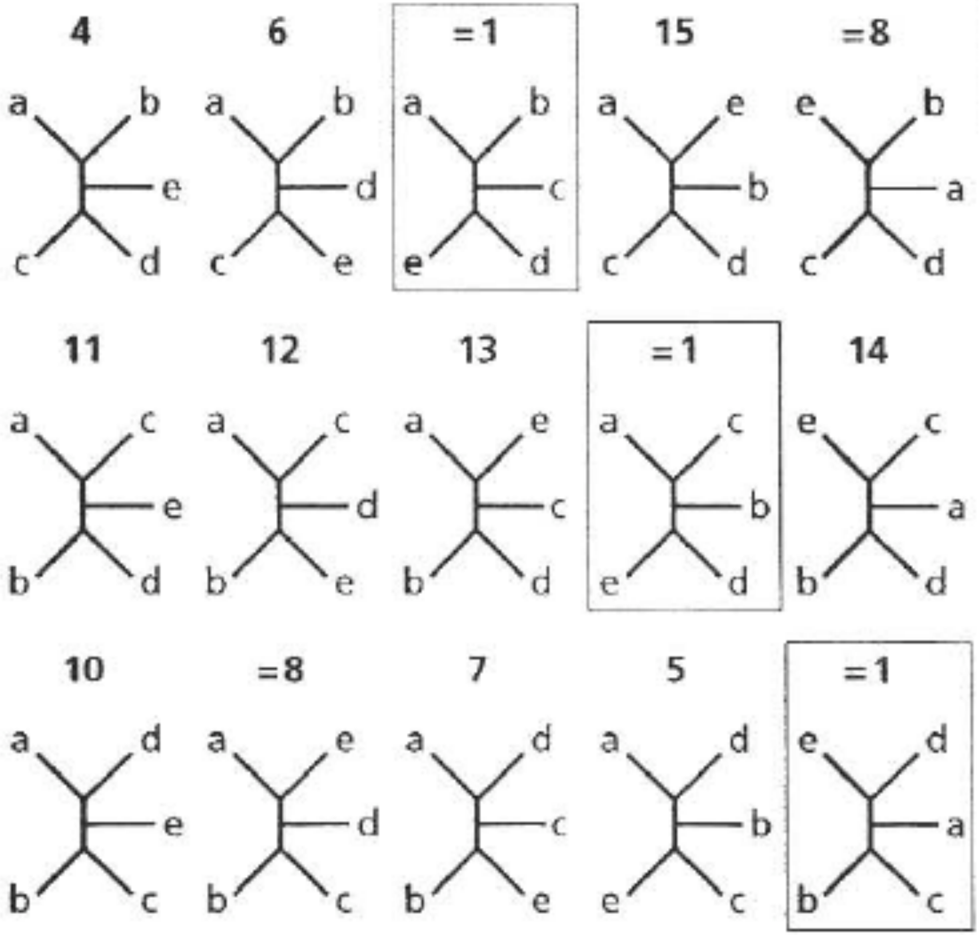


I. Counting the Number of Changes

- The Fitch/Wagner method.
  - General model for reversible change.
- The Dollo method.
  - Model for characters that are considered more likely to evolve in one direction.

Optimality Criterion





How Many Trees?

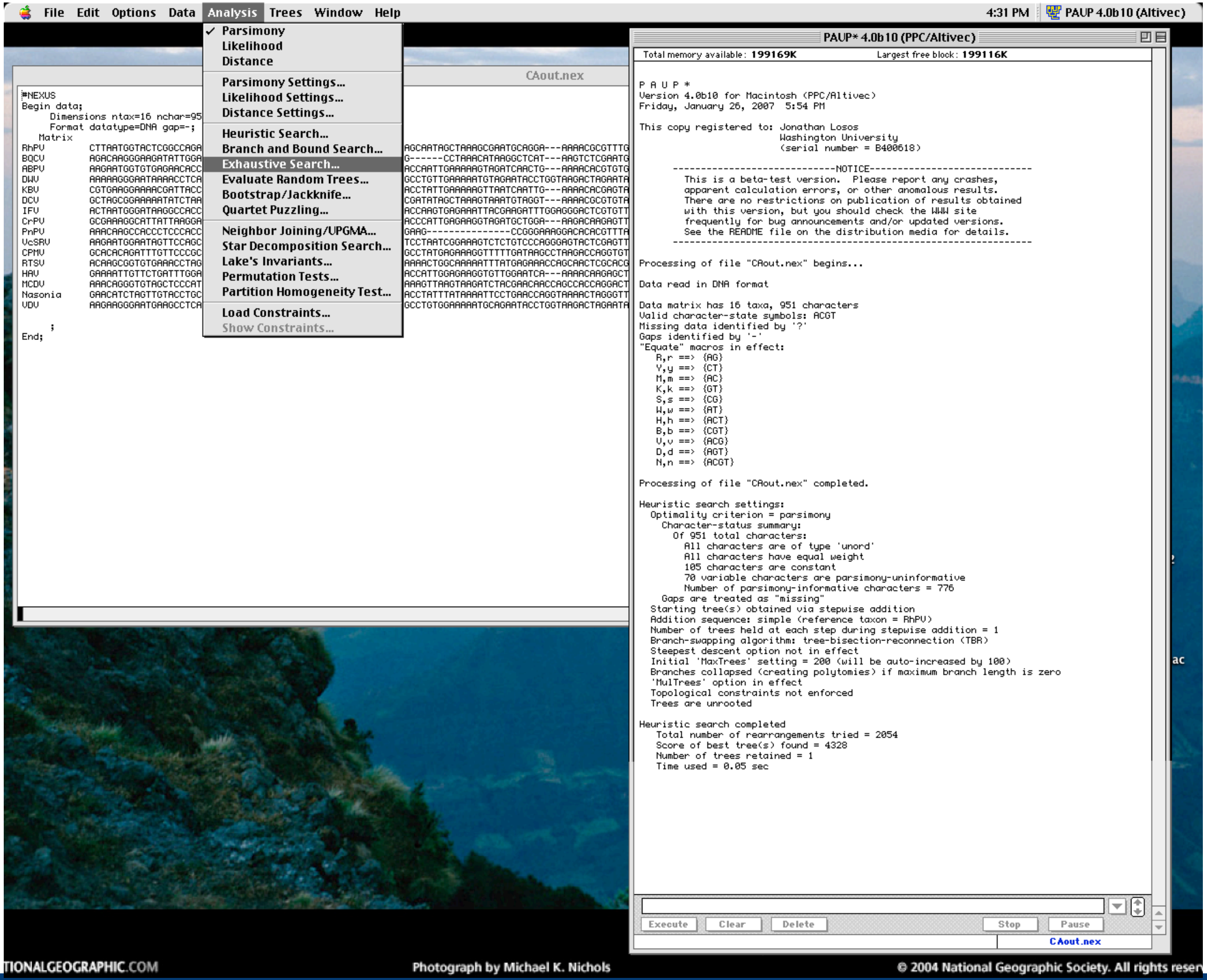
Taxa	Unrooted Trees	Rooted Trees
4	3	15
8	10,395	135,135
10	2,027,025	34,459,425
22	3x10^23	
50	3x10^74*	

* More trees than there are atoms in the universe.

II. Searching For the Optimal Topology

1. Exhaustive
    - Examines all possible trees.
    - Only works for small number of taxa (<12).
2. Branch & Bound
    - Ignores trees that are sure to be worse.
    - Again, only works for small numbers (<18).
3. Heuristic
    - Uses algorithms to explore tree space.
    - Only method that works for large datasets.

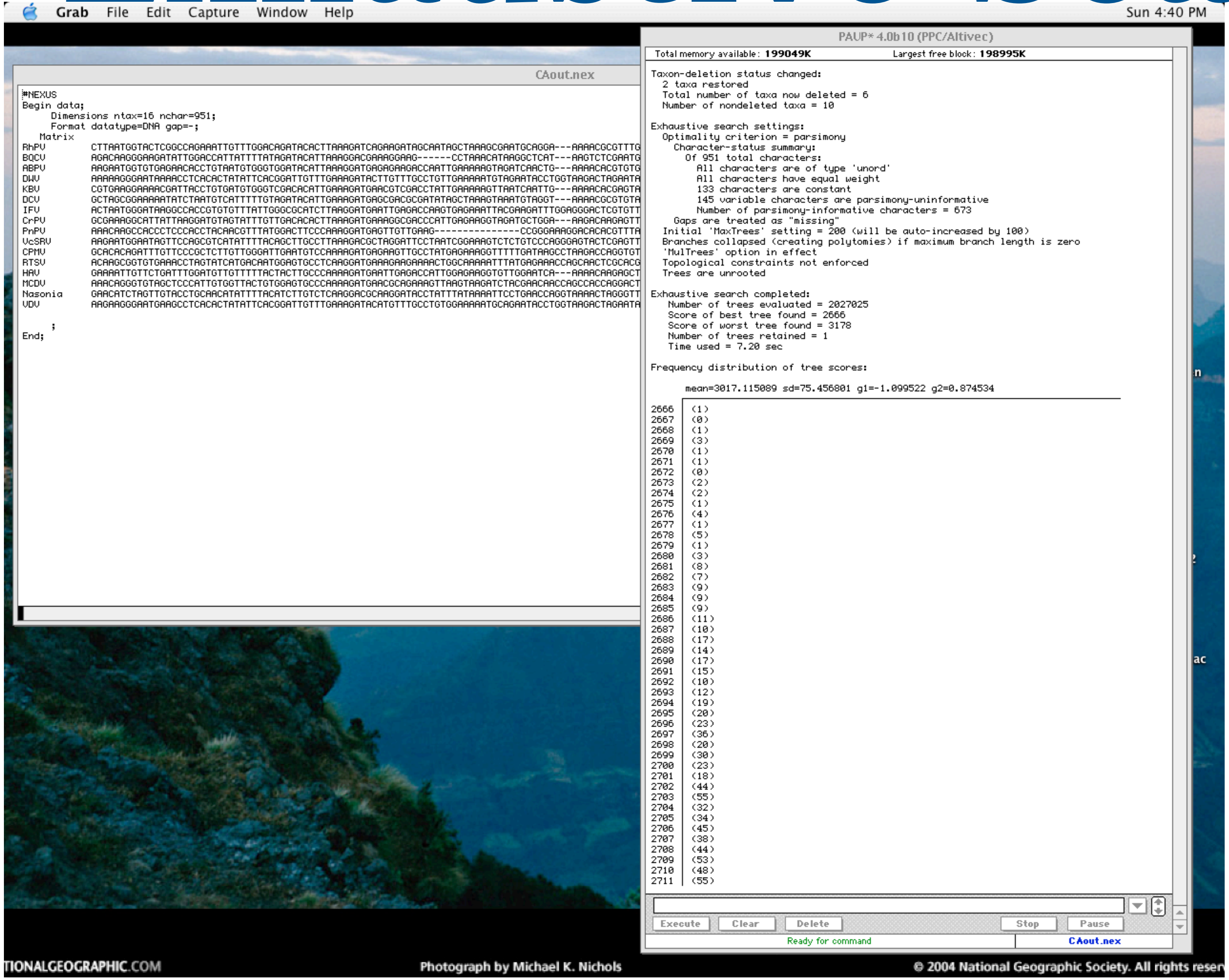
1. Exhaustive Search



```
#NEXUS
Begin data;
     Dimensions ntax=16 nchar=95
     Format datatype=DNA gap=-;
   Matrix
RhPV        CTTAATGGTACTCGGCCAGA    AGCAATAGCTAAAGCGAATGCAGGA---AAAACGCGTTTG
BQCV        AGACAAGGGAAGATATTGGA    G------CCTAAACATAAGGCTCAT---AAGTCTCGAATG
ABPV        AAGAATGGTGTGAGAACACC    ACCAATTGAAAAAGTAGATCAACTG---AAAACACGTGTG
DWV         AAAAAGGGAATAAAACCTCA    GCCTGTTGAAAAATGTAGAATACCTGGTAAGACTAGAATA
KBV         CGTGAAGGAAAACGATTACC    ACCTATTGAAAAAGTTAATCAATTG---AAAACACGAGTA
DCV         GCTAGCGGAAAAATATCTAA    CGATATAGCTAAAGTAAATGTAGGT---AAAACGCGTGTA
IFV         ACTAATGGGATAAGGCCACC    ACCAAGTGAGAAATTACGAAGATTTGGAGGGACTCGTGTT
CrPV        GCGAAAGGCATTATTAAGGA    ACCCATTGAGAAGGTAGATGCTGGA---AAGACAAGAGTT
PnPV        AAACAAGCCACCCTCCCACC    GAAG--------------CCGGGAAAGGACACACGTTTA
VcSRV       AAGAATGGAATAGTTCCAGC    TCCTAATCGGAAAGTCTCTGTCCCAGGGAGTACTCGAGTT
CPMV        GCACACAGATTTGTTCCCGC    GCCTATGAGAAAGGTTTTTGATAAGCCTAAGACCAGGTGT
RTSV        ACAAGCGGTGTGAAACCTAG    AAAACTGGCAAAAATTTATGAGAAACCAGCAACTCGCACG
HAV         GAAAATTGTTCTGATTTGGA    ACCATTGGAGAAGGTGTTGGAATCA---AAAACAAGAGCT
MCDV        AAACAGGGTGTAGCTCCCAT    AAAGTTAAGTAAGATCTACGAACAACCAGCCACCAGGACT
Nasonia     GAACATCTAGTTGTACCTGC    ACCTATTTATAAAATTCCTGAACCAGGTAAAACTAGGGTT
VDV         AAGAAGGGAATGAAGCCTCA    GCCTGTGGAAAAATGCAGAATACCTGGTAAGACTAGAATA

   ;
End;
```

```
Total memory available: 199169K          Largest free block: 199116K

P A U P *
Version 4.0b10 for Macintosh (PPC/Altivec)
Friday, January 26, 2007  5:54 PM

This copy registered to: Jonathan Losos
                         Washington University
                         (serial number = B400618)

      -----------------------------NOTICE-----------------------------
        This is a beta-test version.  Please report any crashes,
        apparent calculation errors, or other anomalous results.
        There are no restrictions on publication of results obtained
        with this version, but you should check the WWW site
        frequently for bug announcements and/or updated versions.
        See the README file on the distribution media for details.
      ----------------------------------------------------------------

Processing of file "CAout.nex" begins...

Data read in DNA format

Data matrix has 16 taxa, 951 characters
Valid character-state symbols: ACGT
Missing data identified by '?'
Gaps identified by '-'
"Equate" macros in effect:
   R,r ==> {AG}
   Y,y ==> {CT}
   M,m ==> {AC}
   K,k ==> {GT}
   S,s ==> {CG}
   W,w ==> {AT}
   H,h ==> {ACT}
   B,b ==> {CGT}
   V,v ==> {ACG}
   D,d ==> {AGT}
   N,n ==> {ACGT}

Processing of file "CAout.nex" completed.

Heuristic search settings:
  Optimality criterion = parsimony
    Character-status summary:
      Of 951 total characters:
        All characters are of type 'unord'
        All characters have equal weight
        105 characters are constant
        70 variable characters are parsimony-uninformative
        Number of parsimony-informative characters = 776
    Gaps are treated as "missing"
  Starting tree(s) obtained via stepwise addition
  Addition sequence: simple (reference taxon = RhPV)
  Number of trees held at each step during stepwise addition = 1
  Branch-swapping algorithm: tree-bisection-reconnection (TBR)
  Steepest descent option not in effect
  Initial 'MaxTrees' setting = 200 (will be auto-increased by 100)
  Branches collapsed (creating polytomies) if maximum branch length is zero
  'MulTrees' option in effect
  Topological constraints not enforced
  Trees are unrooted

Heuristic search completed
   Total number of rearrangements tried = 2054
   Score of best tree(s) found = 4328
   Number of trees retained = 1
   Time used = 0.05 sec
```

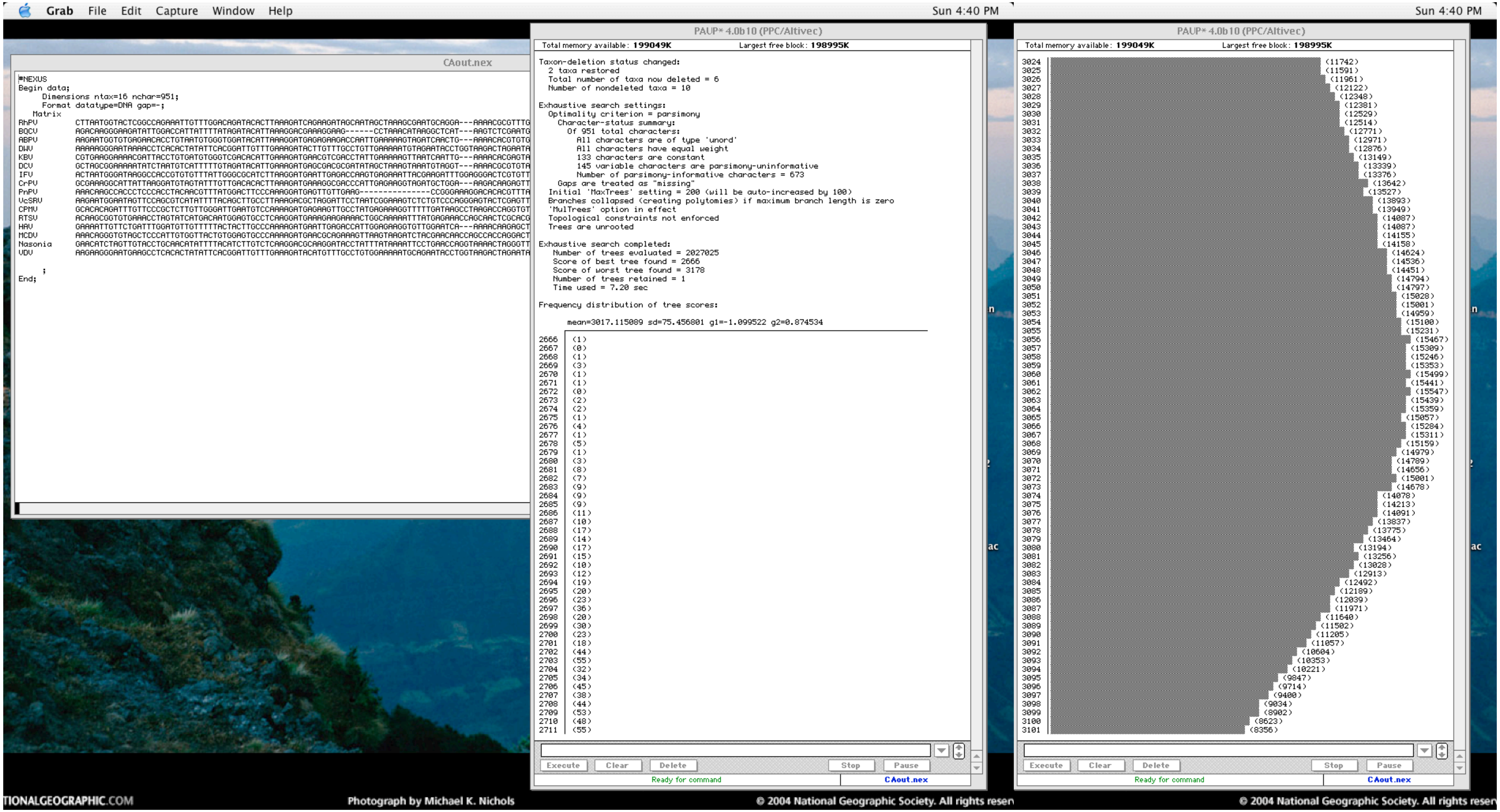
1. Exhaustive Search



```
#NEXUS
Begin data;
     Dimensions ntax=16 nchar=951;
     Format datatype=DNA gap=-;
   Matrix
RhPV        CTTAATGGTACTCGGCCAGAAATTGTTTGGACAGATACACTTAAAGATCAGAAGATAGCAATAGCTAAAGCGAATGCAGGA---AAAACGCGTTTG
BQCV        AGACAAGGGAAGATATTGGACCATTATTTTATAGATACATTAAAGGACGAAAGGAAG------CCTAAACATAAGGCTCAT---AAGTCTCGAATG
ABPV        AAGAATGGTGTGAGAACACCTGTAATGTGGGTGGATACATTAAAGGATGAGAGAAGACCAATTGAAAAAGTAGATCAACTG---AAAACACGTGTG
DWV         AAAAAGGGAATAAAACCTCACACTATATTCACGGATTGTTTGAAAGATACTTGTTTGCCTGTTGAAAAAATGTAGAATACCTGGTAAGACTAGAATA
KBV         CGTGAAGGAAAACGATTACCTGTGATGTGGGTCGACACATTGAAAGATGAACGTCGACCTATTGAAAAAGTTAATCAATTG---AAAACACGAGTA
DCV         GCTAGCGGAAAAAATATCTAATGTCATTTTTGTAGATACATTGAAAGATGAGCGACGCGATATAGCTAAAGTAAATGTAGGT---AAAACGCGTGTA
IFV         ACTAATGGGATAAGGCCACCGTGTGTTTATTGGGCGCATCTTAAGGATGAATTGAGACCAAGTGAGAAATTACGAAGATTTGGAGGGACTCGTGTT
CrPV        GCGAAAGGCATTATTAAGGATGTAGTATTTGTTGACACACTTAAAGATGAAAGGCGACCCATTGAGAAGGTAGATGCTGGA---AAGACAAGAGTT
PnPV        AAACAAGCCACCCTCCCACCTACAACGTTTATGGACTTCCCAAAGGATGAGTTGTTGAAG---------------CCGGGAAAGGACACACGTTTA
VcSRV       AAGAATGGAATAGTTCCAGCGTCATATTTTACAGCTTGCCTTAAAGACGCTAGGATTCCTAATCGGAAAGTCTCTGTCCCAGGGAGTACTCGAGTT
CPMV        GCACACAGATTTGTTCCCGCTCTTGTTGGGATTGAATGTCCAAAAGATGAGAAGTTGCCTATGAGAAAGGTTTTTGATAAGCCTAAGACCAGGTGT
RTSV        ACAAGCGGTGTGAAACCTAGTATCATGACAATGGAGTGCCTCAAGGATGAAAGAAGAAAACTGGCAAAAATTTATGAGAAACCAGCAACTCGCACG
HAV         GAAAATTGTTCTGATTTGGATGTTGTTTTTACTACTTGCCCAAAAGATGAATTGAGACCATTGGAGAAGGTGTTGGAATCA---AAAACAAGAGCT
MCDV        AAACAGGGTGTAGCTCCCATTGTGGTTACTGTGGAGTGCCCAAAAGATGAACGCAGAAAGTTAAGTAAGATCTACGAACAACCAGCCACCAGGACT
Nasonia     GAACATCTAGTTGTACCTGCAACATATTTTACATCTTGTCTCAAGGACGCAAGGATACCTATTTATAAAATTCCTGAACCAGGTAAAACTAGGGTT
VDV         AAGAAGGGAATGAAGCCTCACACTATATTCACGGATTGTTTGAAAGATACATGTTTGCCTGTGGAAAAAATGCAGAATACCTGGTAAGACTAGAATA

     ;
End;
```

```
Taxon-deletion status changed:
  2 taxa restored
  Total number of taxa now deleted = 6
  Number of nondeleted taxa = 10

Exhaustive search settings:
  Optimality criterion = parsimony
    Character-status summary:
      Of 951 total characters:
        All characters are of type 'unord'
        All characters have equal weight
        133 characters are constant
        145 variable characters are parsimony-uninformative
        Number of parsimony-informative characters = 673
    Gaps are treated as "missing"
  Initial 'MaxTrees' setting = 200 (will be auto-increased by 100)
  Branches collapsed (creating polytomies) if maximum branch length is zero
  'MulTrees' option in effect
  Topological constraints not enforced
  Trees are unrooted

Exhaustive search completed:
   Number of trees evaluated = 2027025
   Score of best tree found = 2666
   Score of worst tree found = 3178
   Number of trees retained = 1
   Time used = 7.20 sec

Frequency distribution of tree scores:

      mean=3017.115089 sd=75.456801 g1=-1.099522 g2=0.874534
      ----------------------------------------------------------
2666 | (1)
2667 | (0)
2668 | (1)
2669 | (3)
2670 | (1)
2671 | (1)
2672 | (0)
2673 | (2)
2674 | (2)
2675 | (1)
2676 | (4)
2677 | (1)
2678 | (5)
2679 | (1)
2680 | (3)
2681 | (8)
2682 | (7)
2683 | (9)
2684 | (9)
2685 | (9)
2686 | (11)
2687 | (10)
2688 | (17)
2689 | (14)
2690 | (17)
2691 | (15)
2692 | (10)
2693 | (12)
2694 | (19)
2695 | (20)
2696 | (23)
2697 | (36)
2698 | (20)
2699 | (30)
2700 | (23)
2701 | (18)
2702 | (44)
2703 | (55)
2704 | (32)
2705 | (34)
2706 | (45)
2707 | (38)
2708 | (44)
2709 | (53)
2710 | (48)
2711 | (55)
```

1. Exhaustive Search



```
#NEXUS
Begin data;
     Dimensions ntax=16 nchar=951;
     Format datatype=DNA gap=-;
   Matrix
RhPV        CTTAATGGTACTCGGCCAGAAATTGTTTGGACAGATACACTTAAAGATCAGAAGATAGCAATAGCTAAAGCGAATGCAGGA---AAAACGCGTTTG
BQCV        AGACAAGGGAAGATATTGGACCATTATTTTATAGATACATTAAAGGACGAAAGGAAG------CCTAAACATAAGGCTCAT---AAGTCTCGAATG
ABPV        AAGAATGGTGTGAGAACACCTGTAATGTGGGTGGATACATTAAAGGATGAGAGAAGACCAATTGAAAAAGTAGATCAACTG---AAAACACGTGTG
DWV         AAAAAGGGAATAAAACCTCACACTATATTCACGGATTGTTTGAAAGATACTTGTTTGCCTGTTGAAAAATGTAGAATACCTGGTAAGACTAGAATA
KBV         CGTGAAGGAAAACGATTACCTGTGATGTGGGTCGACACATTGAAAGATGAACGTCGACCTATTGAAAAAGTTAATCAATTG---AAAACACGAGTA
DCV         GCTAGCGGAAAAATATCTAATGTCATTTTTGTAGATACATTGAAAGATGAGCGACGCGATATAGCTAAAGTAAATGTAGGT---AAAACGCGTGTA
IFV         ACTAATGGGATAAGGCCACCGTGTGTTTATTGGGCGCATCTTAAGGATGAATTGAGACCAAGTGAGAAATTACGAAGATTTGGAGGGACTCGTGTT
CrPV        GCGAAAGGCATTATTAAGGATGTAGTATTTGTTGACACACTTAAAGATGAAAGGCGACCCATTGAGAAGGTAGATGCTGGA---AAGACAAGAGTT
PnPV        AAACAAGCCACCCTCCCACCTACAACGTTTATGGACTTCCCAAAGGATGAGTTGTTGAAG--------------CCGGGAAAGGACACACGTTTA
VcSRV       AAGAATGGAATAGTTCCAGCGTCATATTTTACAGCTTGCCTTAAAGACGCTAGGATTCCTAATCGGAAAGTCTCTGTCCCAGGGAGTACTCGAGTT
CPMV        GCACACAGATTTGTTCCCGCTCTTGTTGGGATTGAATGTCCAAAAGATGAGAAGTTGCCTATGAGAAAGGTTTTTGATAAGCCTAAGACCAGGTGT
RTSV        ACAAGCGGTGTGAAACCTAGTATCATGACAATGGAGTGCCTCAAGGATGAAAGAAGAAAACTGGCAAAAATTTATGAGAAACCAGCAACTCGCACG
HAV         GAAAATTGTTCTGATTTGGATGTTGTTTTTACTACTTGCCCAAAAGATGAATTGAGACCATTGGAGAAGGTGTTGGAATCA---AAAACAAGAGCT
MCDV        AAACAGGGTGTAGCTCCCATTGTGGTTACTGTGGAGTGCCCAAAAGATGAACGCAGAAAGTTAAGTAAGATCTACGAACAACCAGCCACCAGGACT
Nasonia     GAACATCTAGTTGTACCTGCAACATATTTTACATCTTGTCTCAAGGACGCAAGGATACCTATTTATAAAATTCCTGAACCAGGTAAAACTAGGGTT
VDV         AAGAAGGGAATGAAGCCTCACACTATATTCACGGATTGTTTGAAAGATACATGTTTGCCTGTGGAAAAATGCAGAATACCTGGTAAGACTAGAATA

      ;
End;
```

```
Total memory available: 199049K            Largest free block: 198995K

Taxon-deletion status changed:
  2 taxa restored
  Total number of taxa now deleted = 6
  Number of nondeleted taxa = 10

Exhaustive search settings:
  Optimality criterion = parsimony
    Character-status summary:
      Of 951 total characters:
        All characters are of type 'unord'
        All characters have equal weight
        133 characters are constant
        145 variable characters are parsimony-uninformative
        Number of parsimony-informative characters = 673
    Gaps are treated as "missing"
  Initial 'MaxTrees' setting = 200 (will be auto-increased by 100)
  Branches collapsed (creating polytomies) if maximum branch length is zero
  'MulTrees' option in effect
  Topological constraints not enforced
  Trees are unrooted

Exhaustive search completed:
   Number of trees evaluated = 2027025
   Score of best tree found = 2666
   Score of worst tree found = 3178
   Number of trees retained = 1
   Time used = 7.20 sec

Frequency distribution of tree scores:

      mean=3017.115089 sd=75.456801 g1=-1.099522 g2=0.874534

2666 | (1)
2667 | (0)
2668 | (1)
2669 | (3)
2670 | (1)
2671 | (1)
2672 | (0)
2673 | (2)
2674 | (2)
2675 | (1)
2676 | (4)
2677 | (1)
2678 | (5)
2679 | (1)
2680 | (3)
2681 | (8)
2682 | (7)
2683 | (9)
2684 | (9)
2685 | (9)
2686 | (11)
2687 | (10)
2688 | (17)
2689 | (14)
2690 | (17)
2691 | (15)
2692 | (10)
2693 | (12)
2694 | (19)
2695 | (20)
2696 | (23)
2697 | (36)
2698 | (20)
2699 | (30)
2700 | (23)
2701 | (18)
2702 | (44)
2703 | (55)
2704 | (32)
2705 | (34)
2706 | (45)
2707 | (38)
2708 | (44)
2709 | (53)
2710 | (48)
2711 | (55)
```

```
Total memory available: 199049K            Largest free block: 198995K

3024 | (11742)
3025 | (11591)
3026 | (11961)
3027 | (12122)
3028 | (12348)
3029 | (12381)
3030 | (12529)
3031 | (12514)
3032 | (12771)
3033 | (12971)
3034 | (12876)
3035 | (13149)
3036 | (13339)
3037 | (13376)
3038 | (13642)
3039 | (13527)
3040 | (13893)
3041 | (13949)
3042 | (14087)
3043 | (14087)
3044 | (14155)
3045 | (14158)
3046 | (14624)
3047 | (14536)
3048 | (14451)
3049 | (14794)
3050 | (14797)
3051 | (15028)
3052 | (15001)
3053 | (14959)
3054 | (15100)
3055 | (15231)
3056 | (15467)
3057 | (15309)
3058 | (15246)
3059 | (15353)
3060 | (15499)
3061 | (15441)
3062 | (15547)
3063 | (15439)
3064 | (15359)
3065 | (15057)
3066 | (15284)
3067 | (15311)
3068 | (15159)
3069 | (14979)
3070 | (14789)
3071 | (14656)
3072 | (15001)
3073 | (14678)
3074 | (14078)
3075 | (14213)
3076 | (14091)
3077 | (13837)
3078 | (13775)
3079 | (13464)
3080 | (13194)
3081 | (13256)
3082 | (13028)
3083 | (12913)
3084 | (12492)
3085 | (12189)
3086 | (12039)
3087 | (11971)
3088 | (11640)
3089 | (11502)
3090 | (11205)
3091 | (11057)
3092 | (10604)
3093 | (10353)
3094 | (10221)
3095 | (9847)
3096 | (9714)
3097 | (9400)
3098 | (9034)
3099 | (8902)
3100 | (8623)
3101 | (8356)
```

2. Branch & Bound

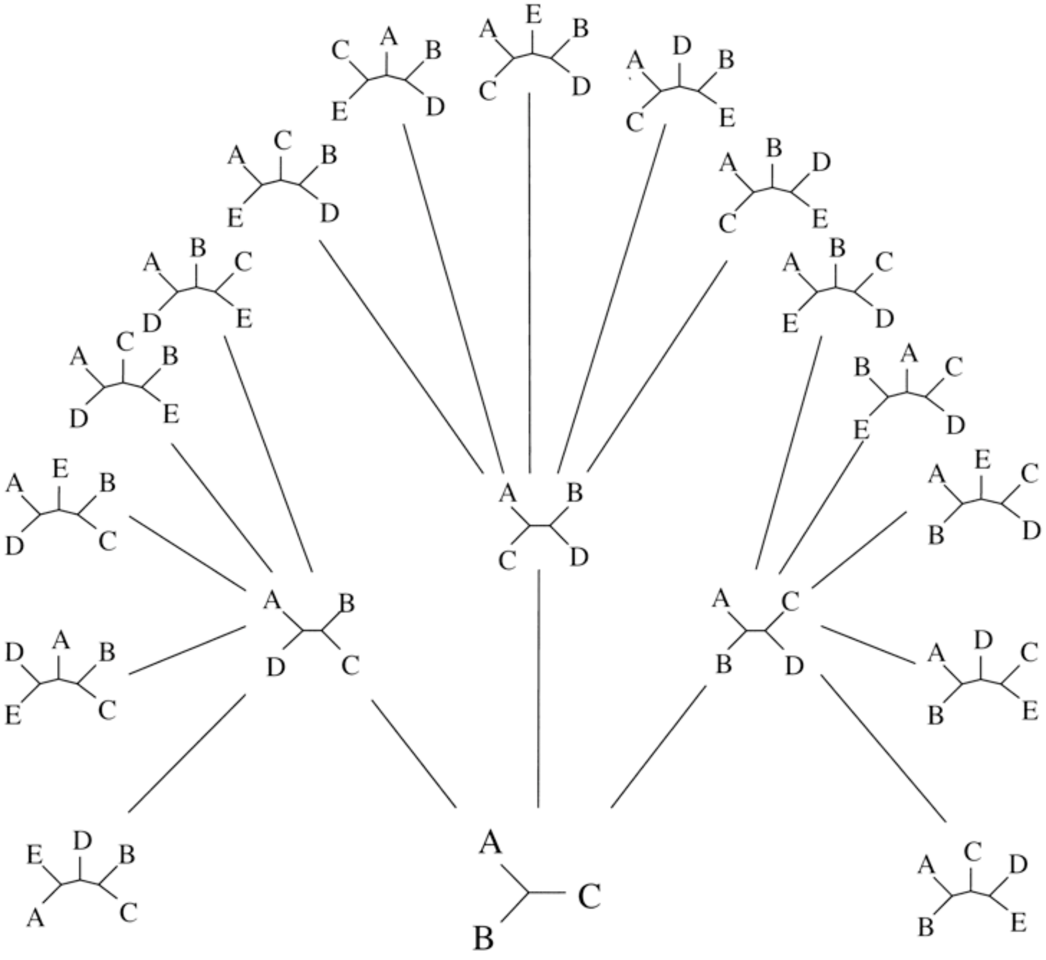


Figure 5.3: Search tree for most parsimonious tree in a five-species case.

2. Branch & Bound

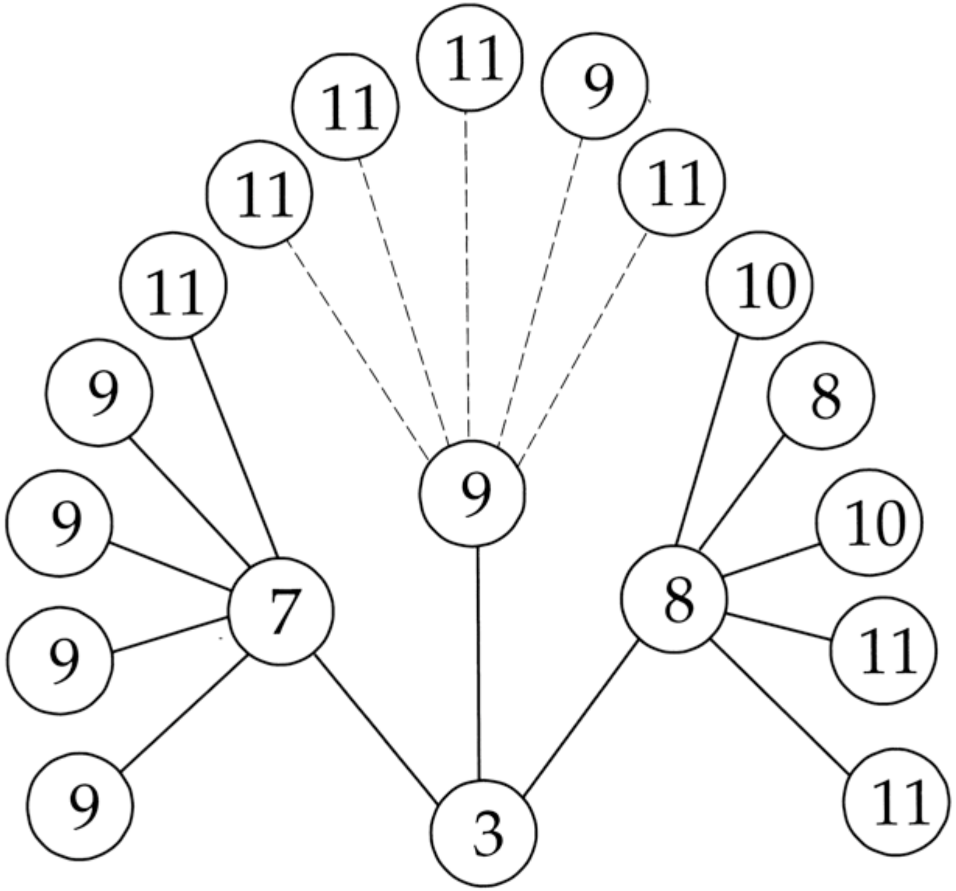
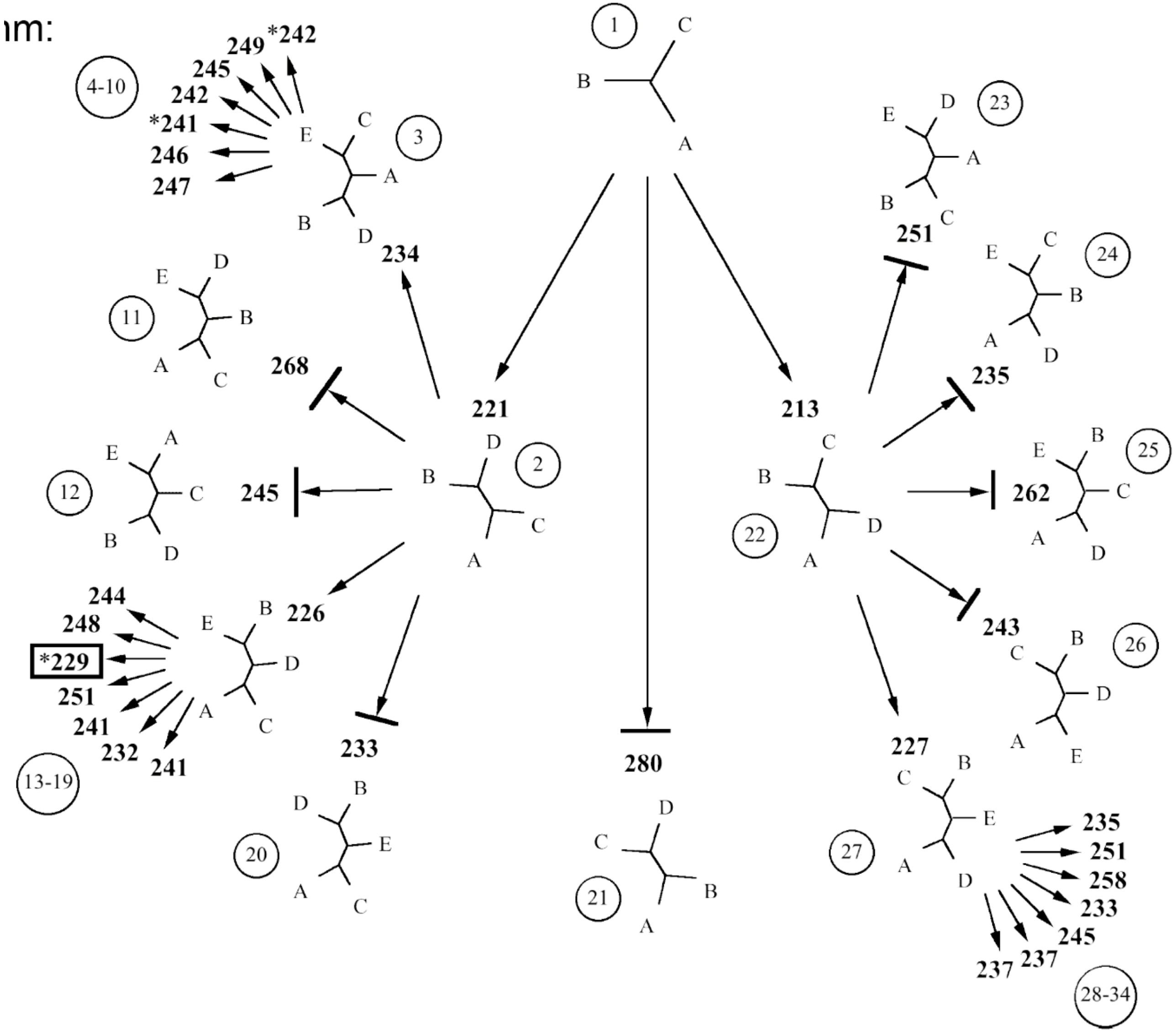
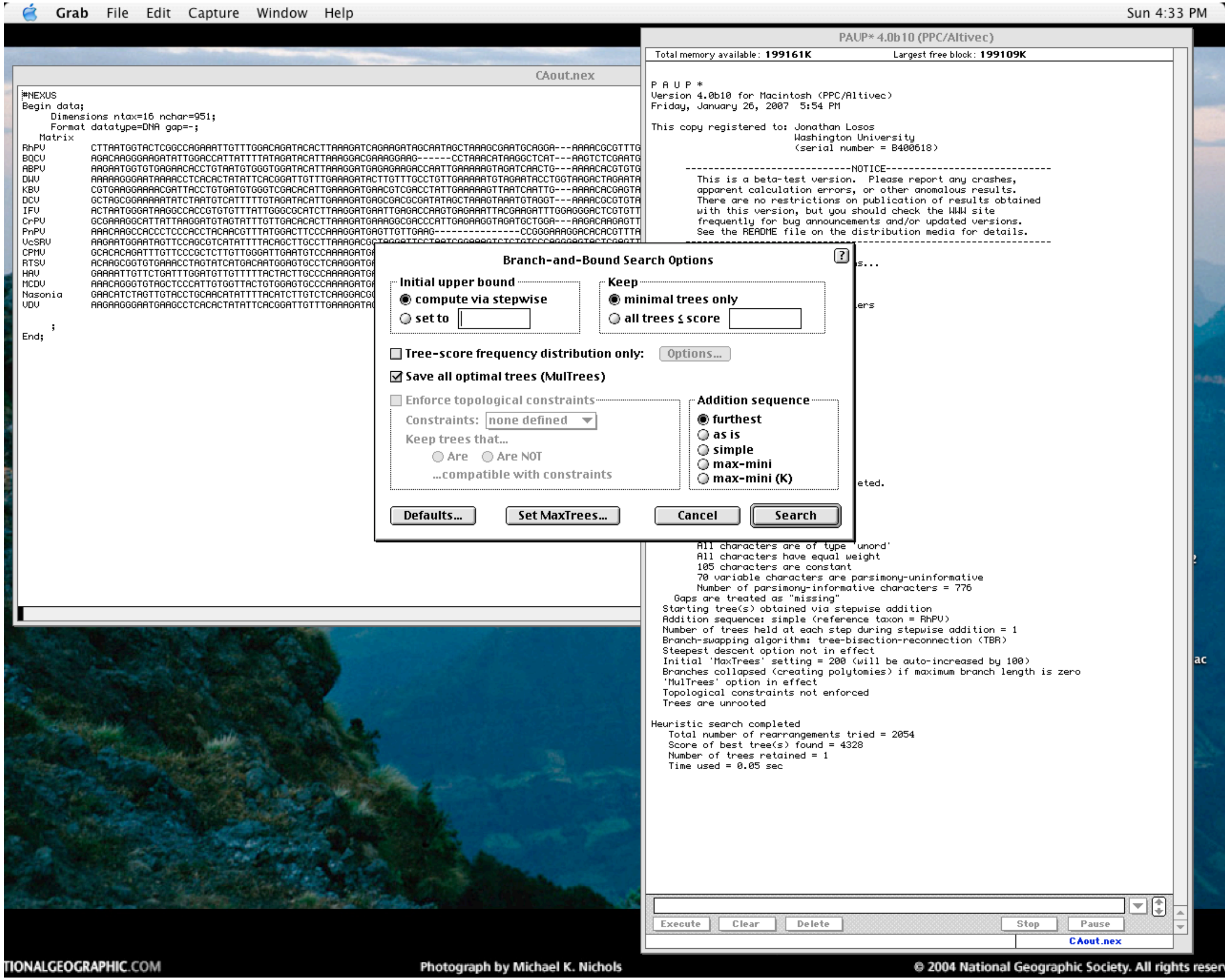


Figure 5.4: Search tree for most parsimonious tree for five species, using the data of Table 1.1. Trees are shown in Figure 5.3. Dashed lines are those not traversed by a branch and bound method. The species names in the data set correspond to labels A through E in Figure 5.3.

2. Branch & Bound



2. Branch & Bound



```
#NEXUS
Begin data;
     Dimensions ntax=16 nchar=951;
     Format datatype=DNA gap=-;
   Matrix
RhPV        CTTAATGGTACTCGGCCAGAAATTGTTTGGACAGATACACTTAAAGATCAGAAGATAGCAATAGCTAAAGCGAATGCAGGA---AAAACGCGTTTG
BQCV        AGACAAGGGAAGATATTGGACCATTATTTTATAGATACATTAAAGGACGAAAGGAAG------CCTAAACATAAGGCTCAT---AAGTCTCGAATG
ABPV        AAGAATGGTGTGAGAACACCTGTAATGTGGGTGGATACATTAAAGGATGAGAGAAGACCAATTGAAAAAGTAGATCAACTG---AAAACACGTGTG
DWV         AAAAAGGGAATAAAACCTCACACTATATTCACGGATTGTTTGAAAGATACTTGTTTGCCTGTTGAAAAATGTAGAATACCTGGTAAGACTAGAATA
KBV         CGTGAAGGAAAACGATTACCTGTGATGTGGGTCGACACATTGAAAGATGAACGTCGACCTATTGAAAAAGTTAATCAATTG---AAAACACGAGTA
DCV         GCTAGCGGAAAAATATCTAATGTCATTTTTGTAGATACATTGAAAGATGAGCGACGCGATATAGCTAAAGTAAATGTAGGT---AAAACGCGTGTA
IFV         ACTAATGGGATAAGGCCACCGTGTGTTTATTGGGCGCATCTTAAGGATGAATTGAGACCAAGTGAGAAATTACGAAGATTTGGAGGGACTCGTGTT
CrPV        GCGAAAGGCATTATTAAGGATGTAGTATTTGTTGACACACTTAAAGATGAAAGGCGACCCATTGAGAAGGTAGATGCTGGA---AAGACAAGAGTT
PnPV        AAACAAGCCACCCTCCCACCTACAACGTTTATGGACTTCCCAAAGGATGAGTTGTTGAAG---------------CCGGGAAAGGACACACGTTTA
VcSRV       AAGAATGGAATAGTTCCAGCGTCATATTTTACAGCTTGCCTTAAAGACGCTAGGATTCCTAATCGGAAAGTCTCTGTCCCAGGGAGTACTCGAGTT
CPMV        GCACACAGATTTGTTCCCGCTCTTGTTGGGATTGAATGTCCAAAAGATGA
RTSV        ACAAGCGGTGTGAAACCTAGTATCATGACAATGGAGTGCCTCAAGGATGA
HAV         GAAAATTGTTCTGATTTGGATGTTGTTTTTACTACTTGCCCAAAAGATGA
MCDV        AAACAGGGTGTAGCTCCCATTGTGGTTACTGTGGAGTGCCCAAAAGATGA
Nasonia     GAACATCTAGTTGTACCTGCAACATATTTTACATCTTGTCTCAAGGACG
VDV         AAGAAGGGAATGAAGCCTCACACTATATTCACGGATTGTTTGAAAGATA

     ;
End;
```

Total memory available: **199161K** Largest free block: **199109K**

```
P A U P *
Version 4.0b10 for Macintosh (PPC/Altivec)
Friday, January 26, 2007  5:54 PM

This copy registered to: Jonathan Losos
                         Washington University
                         (serial number = B400618)

      -------------------------------NOTICE-------------------------------
        This is a beta-test version.  Please report any crashes,
        apparent calculation errors, or other anomalous results.
        There are no restrictions on publication of results obtained
        with this version, but you should check the WWW site
        frequently for bug announcements and/or updated versions.
        See the README file on the distribution media for details.
      --------------------------------------------------------------------

                                                             s...


                                                             ers








                                                             eted.



        All characters are of type 'unord'
        All characters have equal weight
        105 characters are constant
        70 variable characters are parsimony-uninformative
        Number of parsimony-informative characters = 776
    Gaps are treated as "missing"
  Starting tree(s) obtained via stepwise addition
  Addition sequence: simple (reference taxon = RhPV)
  Number of trees held at each step during stepwise addition = 1
  Branch-swapping algorithm: tree-bisection-reconnection (TBR)
  Steepest descent option not in effect
  Initial 'MaxTrees' setting = 200 (will be auto-increased by 100)
  Branches collapsed (creating polytomies) if maximum branch length is zero
  'MulTrees' option in effect
  Topological constraints not enforced
  Trees are unrooted

Heuristic search completed
   Total number of rearrangements tried = 2054
   Score of best tree(s) found = 4328
   Number of trees retained = 1
   Time used = 0.05 sec
```

Branch-and-Bound Search Options

Initial upper bound
- ● compute via stepwise
- ○ set to

Keep
- ● minimal trees only
- ○ all trees ≤ score

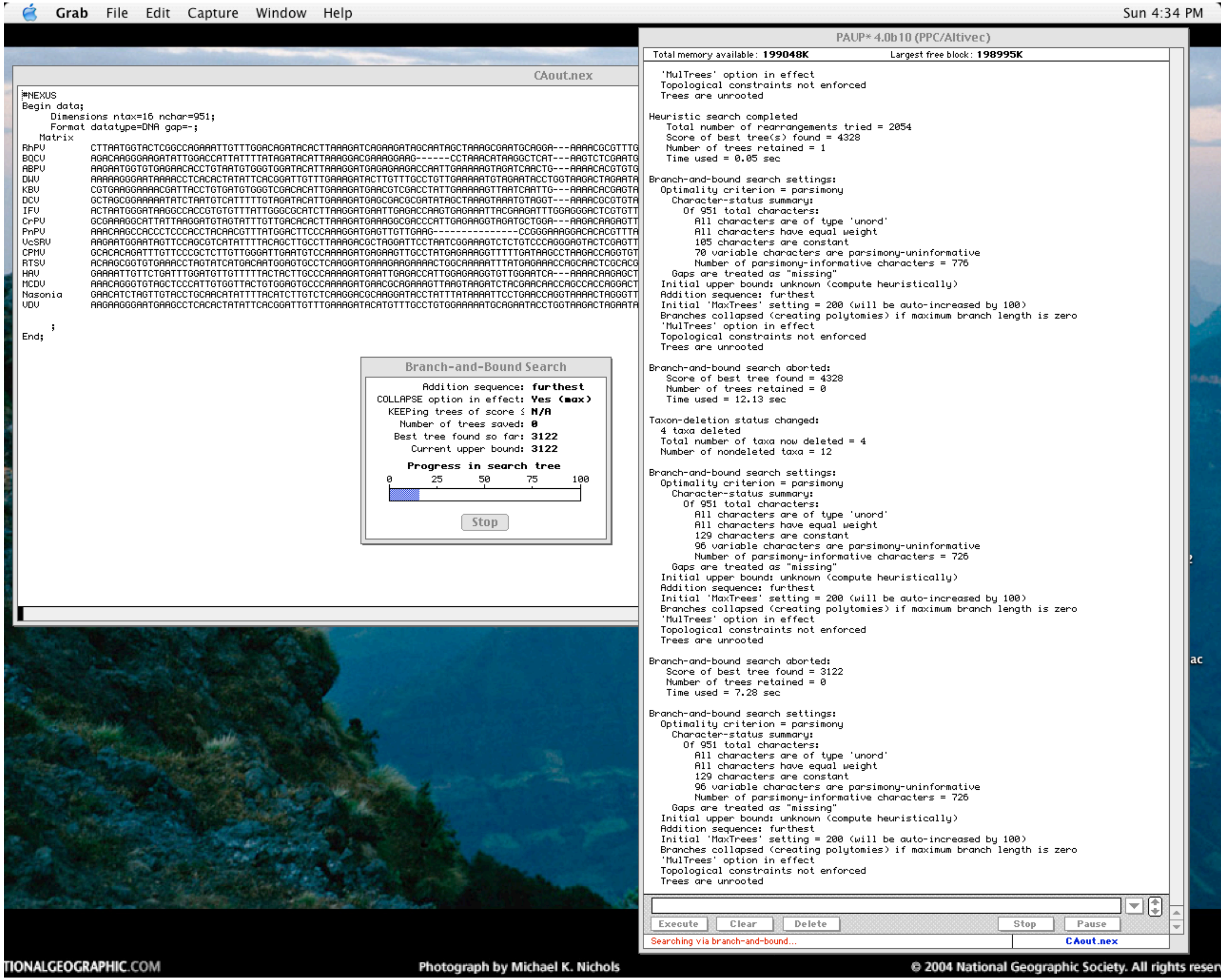
☐ Tree-score frequency distribution only:

☑ Save all optimal trees (MulTrees)

☐ Enforce topological constraints
- Constraints: none defined
- Keep trees that...
  - ○ Are ○ Are NOT
  - ...compatible with constraints

Addition sequence
- ● furthest
- ○ as is
- ○ simple
- ○ max-mini
- ○ max-mini (K)

2. Branch & Bound



```
#NEXUS
Begin data;
     Dimensions ntax=16 nchar=951;
     Format datatype=DNA gap=-;
   Matrix
RhPV        CTTAATGGTACTCGGCCAGAAATTGTTTGGACAGATACACTTAAAGATCAGAAGATAGCAATAGCTAAAGCGAATGCAGGA---AAAACGCGTTTG
BQCV        AGACAAGGGAAGATATTGGACCATTATTTTATAGATACATTAAAGGACGAAAGGAAG------CCTAAACATAAGGCTCAT---AAGTCTCGAATG
ABPV        AAGAATGGTGTGAGAACACCTGTAATGTGGGTGGATACATTAAAGGATGAGAGAAGACCAATTGAAAAAGTAGATCAACTG---AAAACACGTGTG
DWV         AAAAAGGGAATAAAACCTCACACTATATTCACGGATTGTTTGAAAGATACTTGTTTGCCTGTTGAAAAATGTAGAATACCTGGTAAGACTAGAATA
KBV         CGTGAAGGAAAACGATTACCTGTGATGTGGGTCGACACATTGAAAGATGAACGTCGACCTATTGAAAAAGTTAATCAATTG---AAAACACGAGTA
DCV         GCTAGCGGAAAAATATCTAATGTCATTTTTGTAGATACATTGAAAGATGAGCGACGCGATATAGCTAAAGTAAATGTAGGT---AAAACGCGTGTA
IFV         ACTAATGGGATAAGGCCACCGTGTGTTTATTGGGCGCATCTTAAGGATGAATTGAGACCAAGTGAGAAATTACGAAGATTTGGAGGGACTCGTGTT
CrPV        GCGAAAGGCATTATTAAGGATGTAGTATTTGTTGACACACTTAAAGATGAAAGGCGACCCATTGAGAAGGTAGATGCTGGA---AAGACAAGAGTT
PnPV        AAACAAGCCACCCTCCCACCTACAACGTTTATGGACTTCCCAAAGGATGAGTTGTTGAAG--------------CCGGGAAAGGACACACGTTTA
VcSRV       AAGAATGGAATAGTTCCAGCGTCATATTTTACAGCTTGCCTTAAAGACGCTAGGATTCCTAATCGGAAAGTCTCTGTCCCAGGGAGTACTCGAGTT
CPMV        GCACACAGATTTGTTCCCGCTCTTGTTGGGATTGAATGTCCAAAAGATGAGAAGTTGCCTATGAGAAAGGTTTTTGATAAGCCTAAGACCAGGTGT
RTSV        ACAAGCGGTGTGAAACCTAGTATCATGACAATGGAGTGCCTCAAGGATGAAAGAAGAAAACTGGCAAAAATTTATGAGAAACCAGCAACTCGCACG
HAV         GAAAATTGTTCTGATTTGGATGTTGTTTTTACTACTTGCCCAAAAGATGAATTGAGACCATTGGAGAAGGTGTTGGAATCA---AAAACAAGAGCT
MCDV        AAACAGGGTGTAGCTCCCATTGTGGTTACTGTGGAGTGCCCAAAAGATGAACGCAGAAAGTTAAGTAAGATCTACGAACAACCAGCCACCAGGACT
Nasonia     GAACATCTAGTTGTACCTGCAACATATTTTACATCTTGTCTCAAGGACGCAAGGATACCTATTTATAAAATTCCTGAACCAGGTAAAACTAGGGTT
VDV         AAGAAGGGAATGAAGCCTCACACTATATTCACGGATTGTTTGAAAGATACATGTTTGCCTGTGGAAAAATGCAGAATACCTGGTAAGACTAGAATA

      ;
End;
```

```
         Addition sequence: furthest
COLLAPSE option in effect: Yes (max)
   KEEPing trees of score ≤ N/A
     Number of trees saved: 0
    Best tree found so far: 3122
      Current upper bound: 3122

       Progress in search tree
  0     25     50     75     100
```

```
Total memory available: 199048K          Largest free block: 198995K

  'MulTrees' option in effect
  Topological constraints not enforced
  Trees are unrooted

Heuristic search completed
   Total number of rearrangements tried = 2054
   Score of best tree(s) found = 4328
   Number of trees retained = 1
   Time used = 0.05 sec

Branch-and-bound search settings:
  Optimality criterion = parsimony
    Character-status summary:
      Of 951 total characters:
        All characters are of type 'unord'
        All characters have equal weight
        105 characters are constant
        70 variable characters are parsimony-uninformative
        Number of parsimony-informative characters = 776
    Gaps are treated as "missing"
  Initial upper bound: unknown (compute heuristically)
  Addition sequence: furthest
  Initial 'MaxTrees' setting = 200 (will be auto-increased by 100)
  Branches collapsed (creating polytomies) if maximum branch length is zero
  'MulTrees' option in effect
  Topological constraints not enforced
  Trees are unrooted

Branch-and-bound search aborted:
   Score of best tree found = 4328
   Number of trees retained = 0
   Time used = 12.13 sec

Taxon-deletion status changed:
  4 taxa deleted
  Total number of taxa now deleted = 4
  Number of nondeleted taxa = 12

Branch-and-bound search settings:
  Optimality criterion = parsimony
    Character-status summary:
      Of 951 total characters:
        All characters are of type 'unord'
        All characters have equal weight
        129 characters are constant
        96 variable characters are parsimony-uninformative
        Number of parsimony-informative characters = 726
    Gaps are treated as "missing"
  Initial upper bound: unknown (compute heuristically)
  Addition sequence: furthest
  Initial 'MaxTrees' setting = 200 (will be auto-increased by 100)
  Branches collapsed (creating polytomies) if maximum branch length is zero
  'MulTrees' option in effect
  Topological constraints not enforced
  Trees are unrooted

Branch-and-bound search aborted:
   Score of best tree found = 3122
   Number of trees retained = 0
   Time used = 7.28 sec

Branch-and-bound search settings:
  Optimality criterion = parsimony
    Character-status summary:
      Of 951 total characters:
        All characters are of type 'unord'
        All characters have equal weight
        129 characters are constant
        96 variable characters are parsimony-uninformative
        Number of parsimony-informative characters = 726
    Gaps are treated as "missing"
  Initial upper bound: unknown (compute heuristically)
  Addition sequence: furthest
  Initial 'MaxTrees' setting = 200 (will be auto-increased by 100)
  Branches collapsed (creating polytomies) if maximum branch length is zero
  'MulTrees' option in effect
  Topological constraints not enforced
  Trees are unrooted
```

```
Searching via branch-and-bound...
```

3. Heuristic Search

The greedy search algorithm

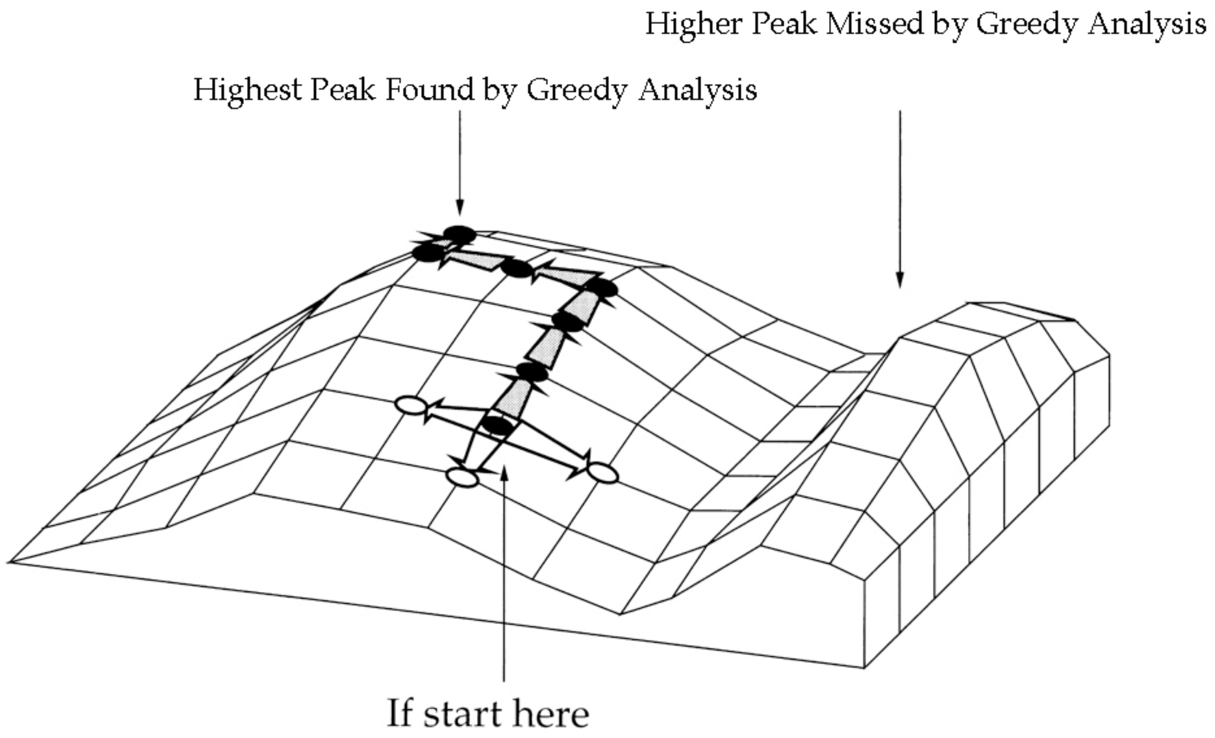


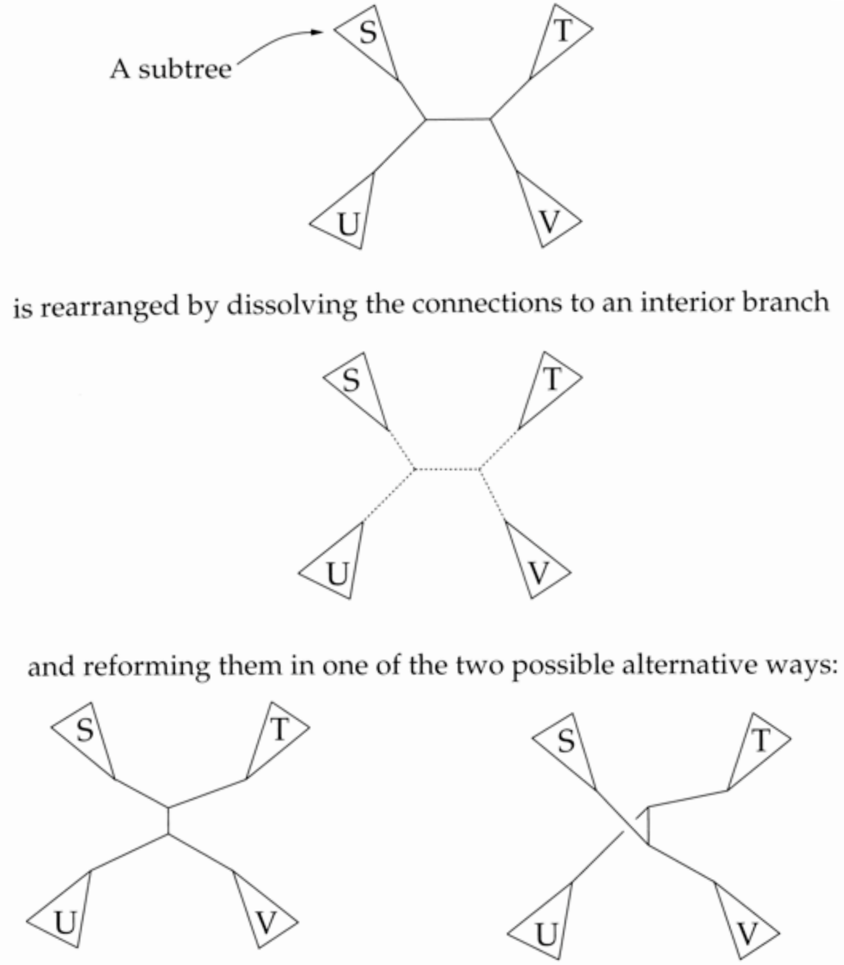
Figure 4.1: A surface rising above a two-dimensional plain (or plane). The process of climbing uphill on the surface is illustrated, as well as the failure to find a higher peak by this "greedy" method.

3. Heuristic Search

1. Rearranging the topology
    - Nearest neighbor interchange (NNI).
    - Subtree pruning and regrafting (SPR).
    - Tree bisection and reconnection (TBR).
2. Sequence of taxon addition
    - Multiple sequence addition replicates required to avoid non-optimal peaks.

3. Heuristic Search

- **Nearest Neighbor Interchange (NNI)**
- Dissolve an internal branch and reform in the two other possible ways.



is rearranged by dissolving the connections to an interior branch

and reforming them in one of the two possible alternative ways:

Figure 4.2: The process of nearest-neighbor interchange. An interior branch is dissolved and the four subtrees connected to it are isolated. These then can be reconnected in two other ways.

3. Heuristic Search

- **Nearest Neighbor Interchange (NNI)**
- Dissolve an internal branch and reform in the two other possible ways.

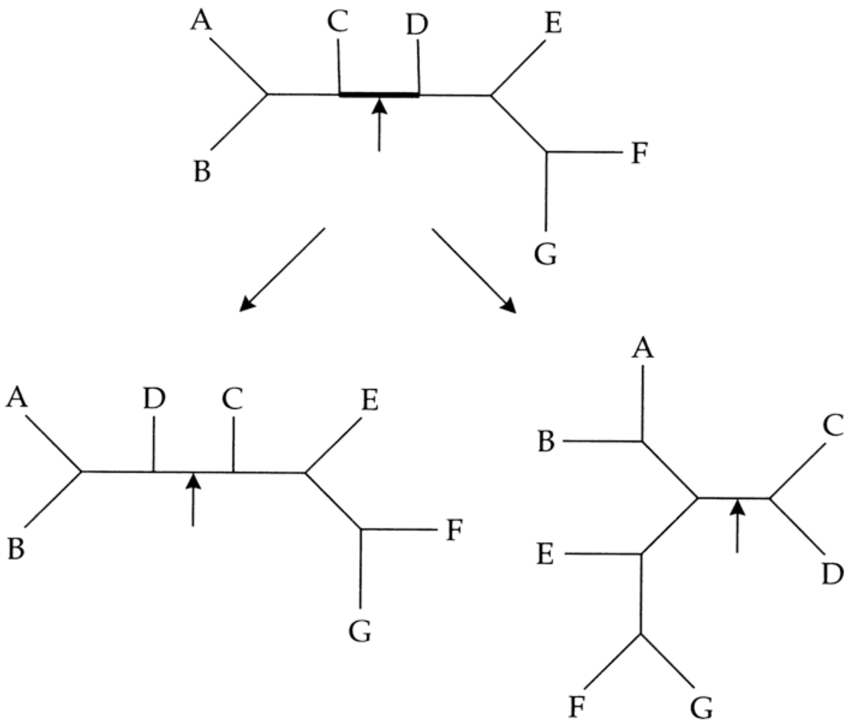
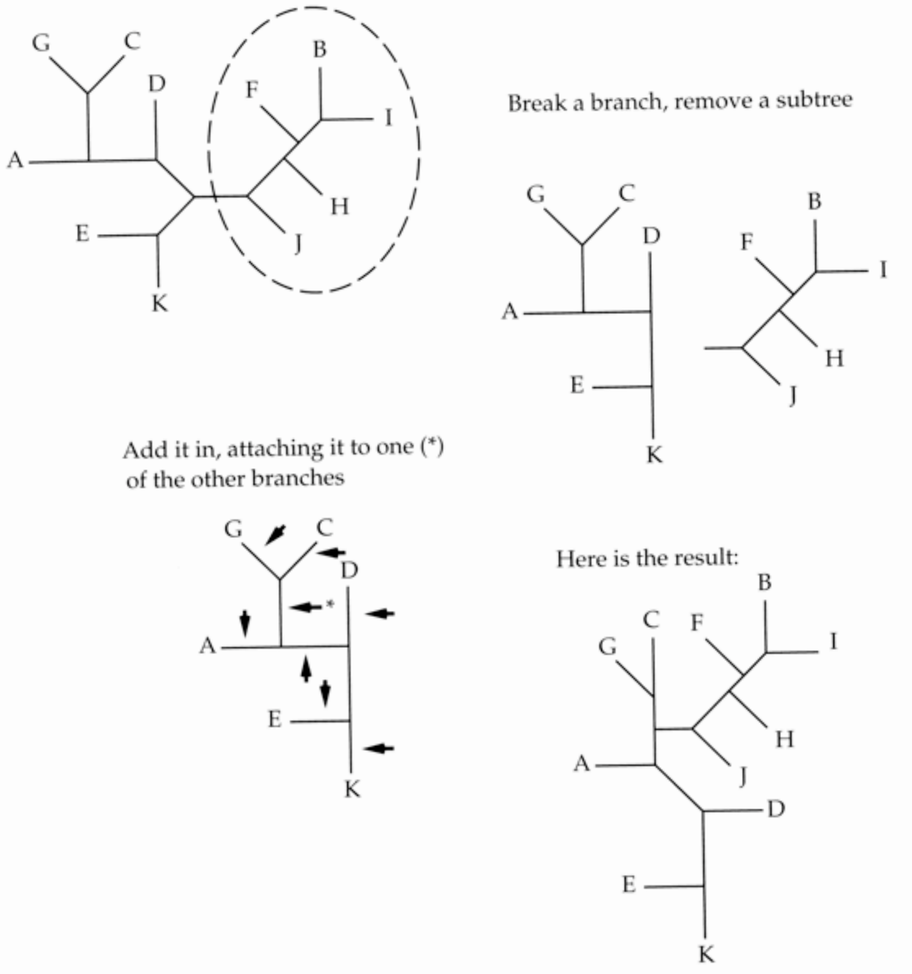


Figure 26 Branch swapping by nearest-neighbor interchanges (NNIs). Each interior branch of the tree defines a local region of four subtrees connected by the interior branch. Interchanging a subtree on one side of the branch with one from the other constitutes an NNI. Two such rearrangements are possible for each interior branch.

3. Heuristic Search

- **Subtree Pruning and Regrafting (SPR)**
- Remove some chunk of the tree and try placing it in every possible position on the remainder of the original tree.


Break a branch, remove a subtree

Add it in, attaching it to one (*) of the other branches

Here is the result:

Figure 4.5: Subtree pruning and regrafting (SPR) rearrangement. The places where the subtree could be reinserted are shown by arrows. The result of one of these reinsertions (at the branch that separates B and C from the other species) is shown.

3. Heuristic Search

- **Tree Bisection and Reconnection (TBR).**
- More elaborate form of SPR in which the two possible subtrees are merged in every possible way.

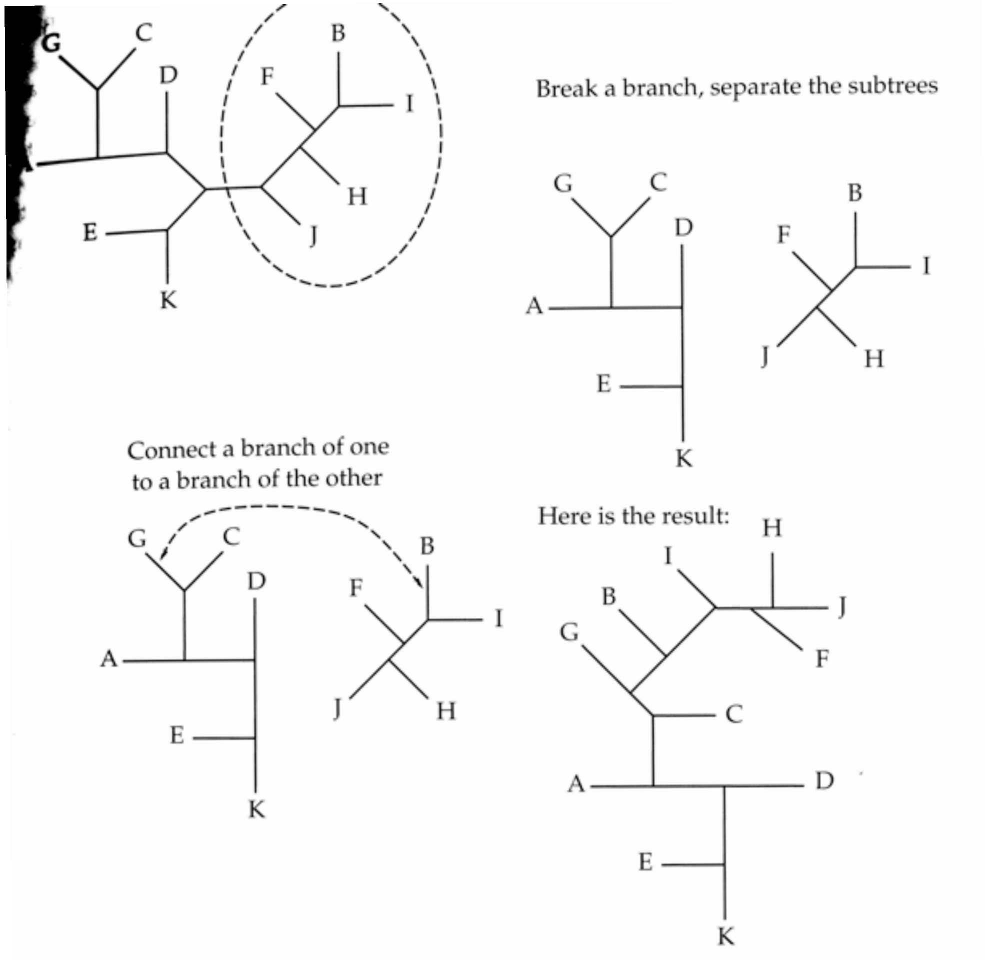
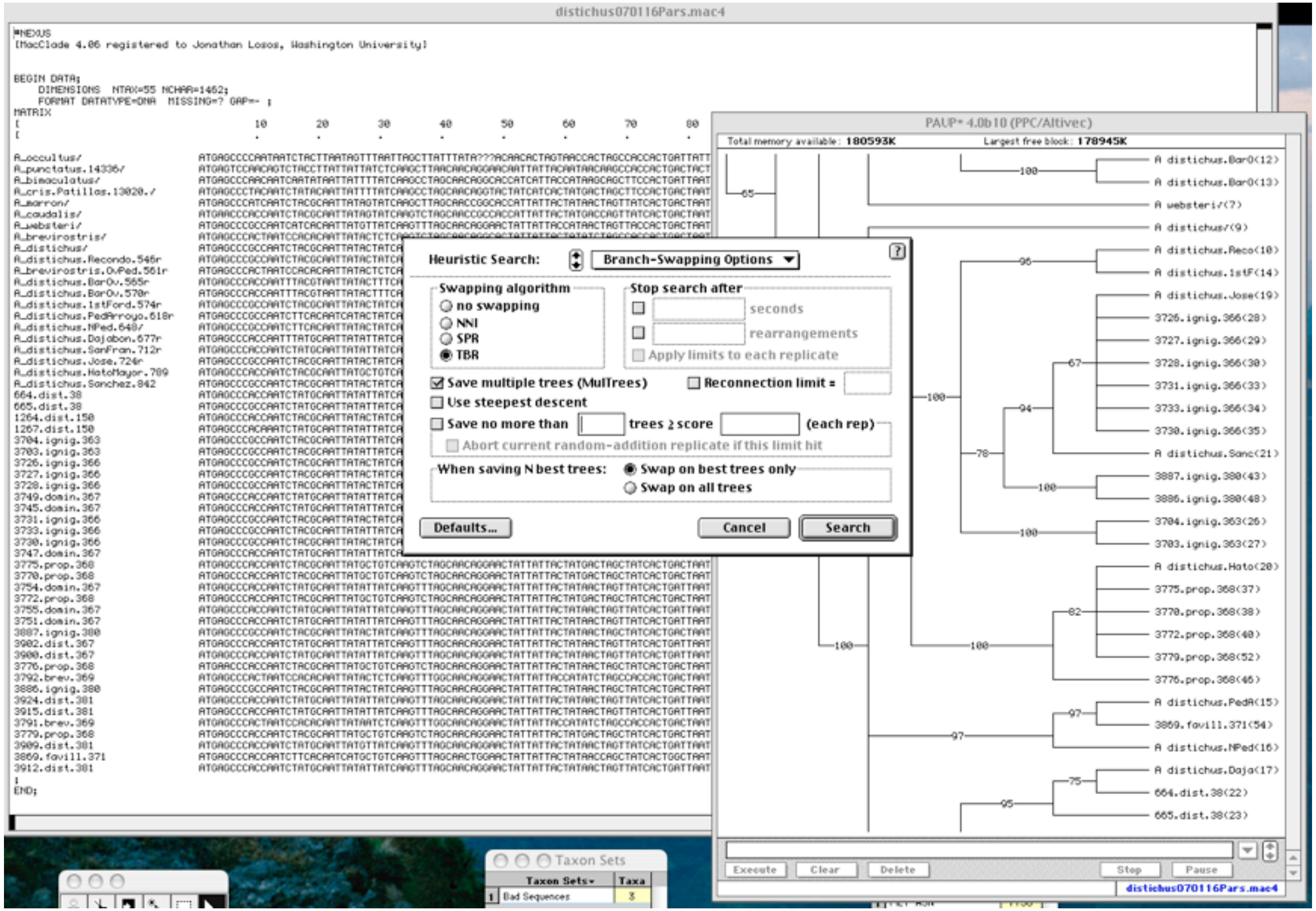


Figure 4.6: Tree bisection and reconnection (TBR). A branch is broken and the two tree fragments are reconnected by putting in branches between all possible branches in one and all possible branches in the other. One of these reconnections and its result are shown here.

3. Heuristic Search



```
#NEXUS
[MacClade 4.06 registered to Jonathan Losos, Washington University]

BEGIN DATA;
    DIMENSIONS  NTAX=55 NCHAR=1462;
    FORMAT DATATYPE=DNA  MISSING=? GAP=- ;
MATRIX
```

Heuristic Search: Branch-Swapping Options

Swapping algorithm
- ○ no swapping
- ○ NNI
- ○ SPR
- ● TBR

Stop search after
- ☐ seconds
- ☐ rearrangements
- ☐ Apply limits to each replicate

☑ Save multiple trees (MulTrees)

☐ Reconnection limit =

☐ Use steepest descent

☐ Save no more than trees ≥ score (each rep)

☐ Abort current random-addition replicate if this limit hit

When saving N best trees:
- ● Swap on best trees only
- ○ Swap on all trees

Defaults... Cancel Search

3. Heuristic Search

- Taxon addition sequence.
    - Manner in which taxa are added to the analysis can influence the results of a heuristic analysis.

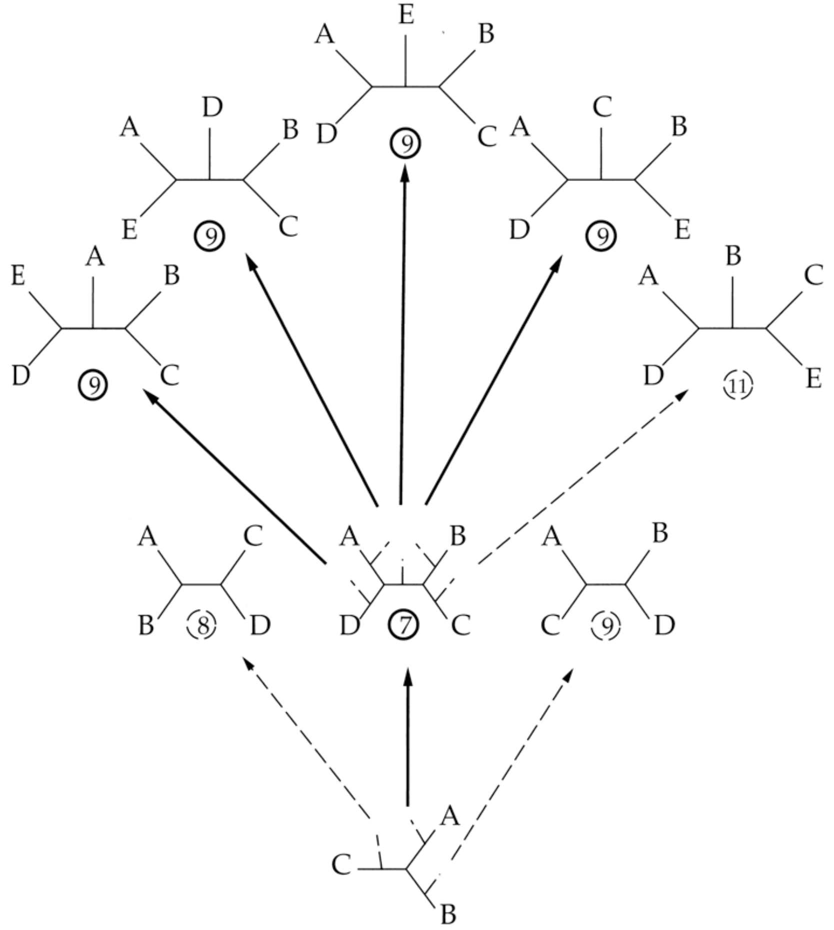


Figure 4.8: Sequential addition carried out on the data in Table 1.1. The species names A–E correspond, respectively, to the names Alpha through Epsilon in Table 1.1. Sequential addition ends up with four trees tied for best. None of these is actually the most parsimonious tree.

3. Heuristic Search

- Taxon addition sequence.
  - Manner in which taxa are added to the analysis can influence the results of a heuristic analysis.
  - Always run multiple sequence addition replicates.

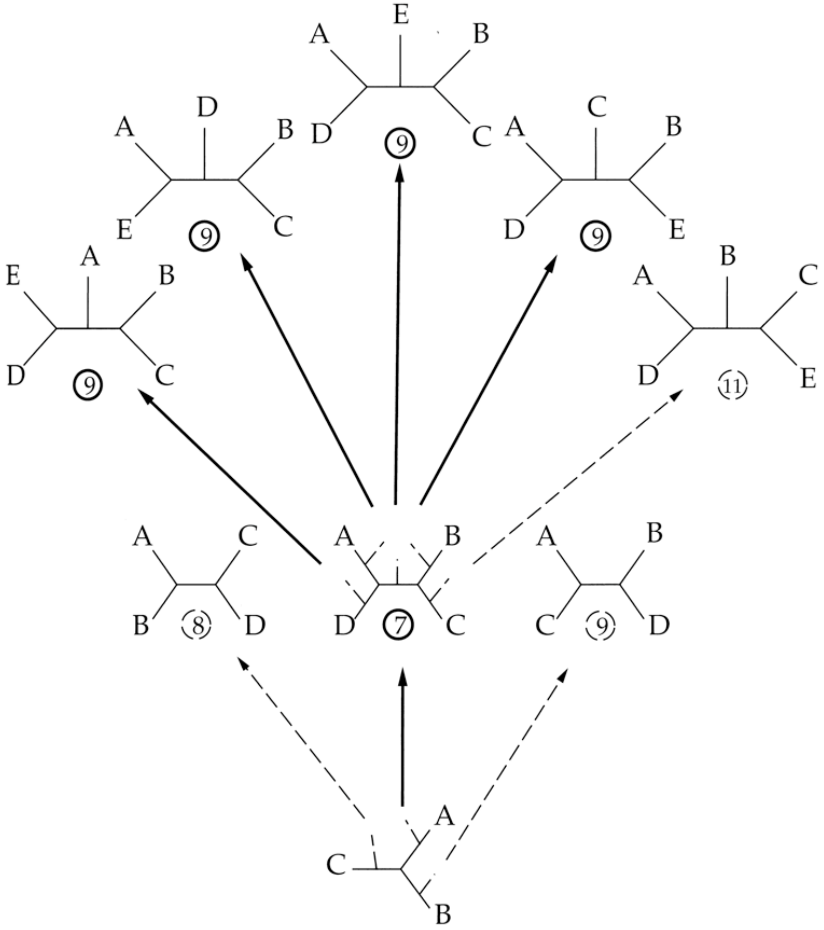
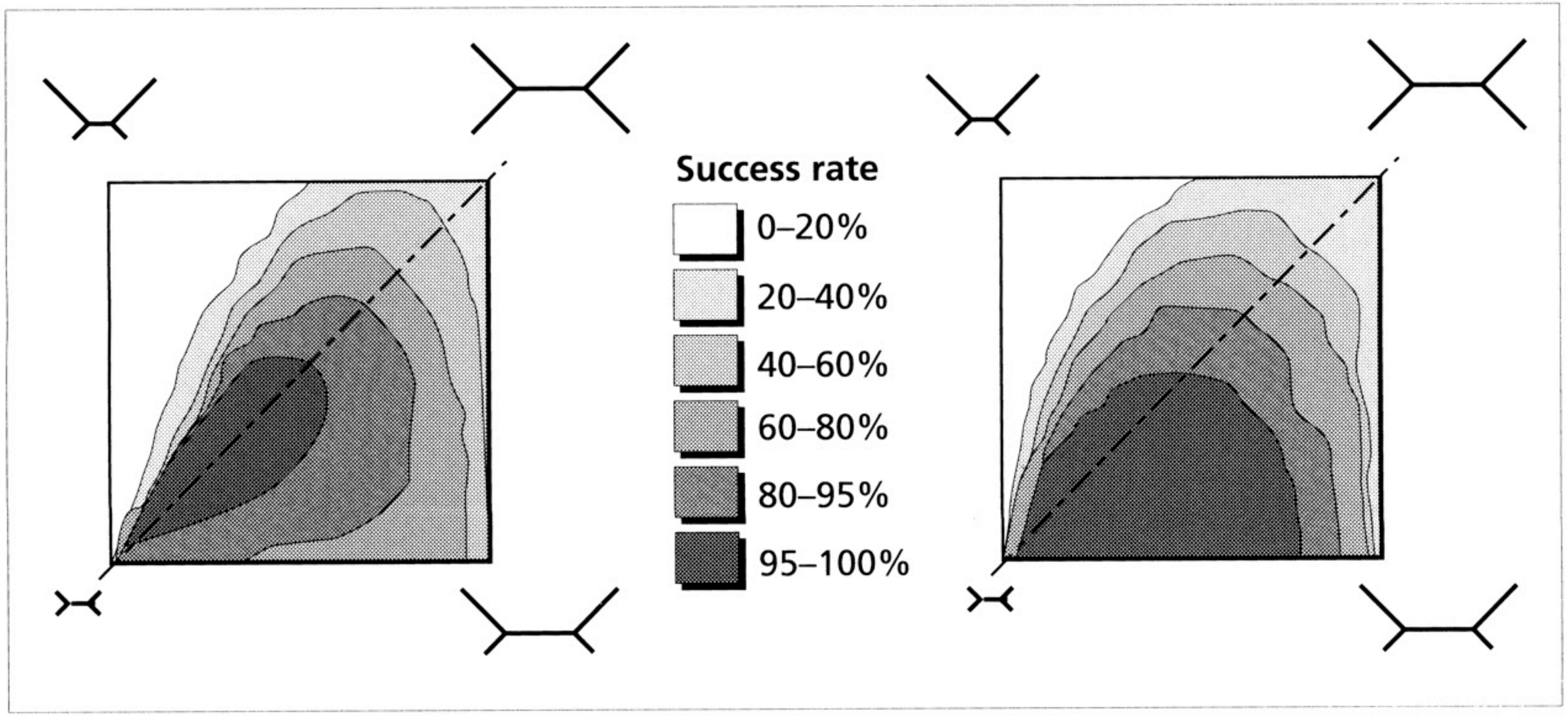
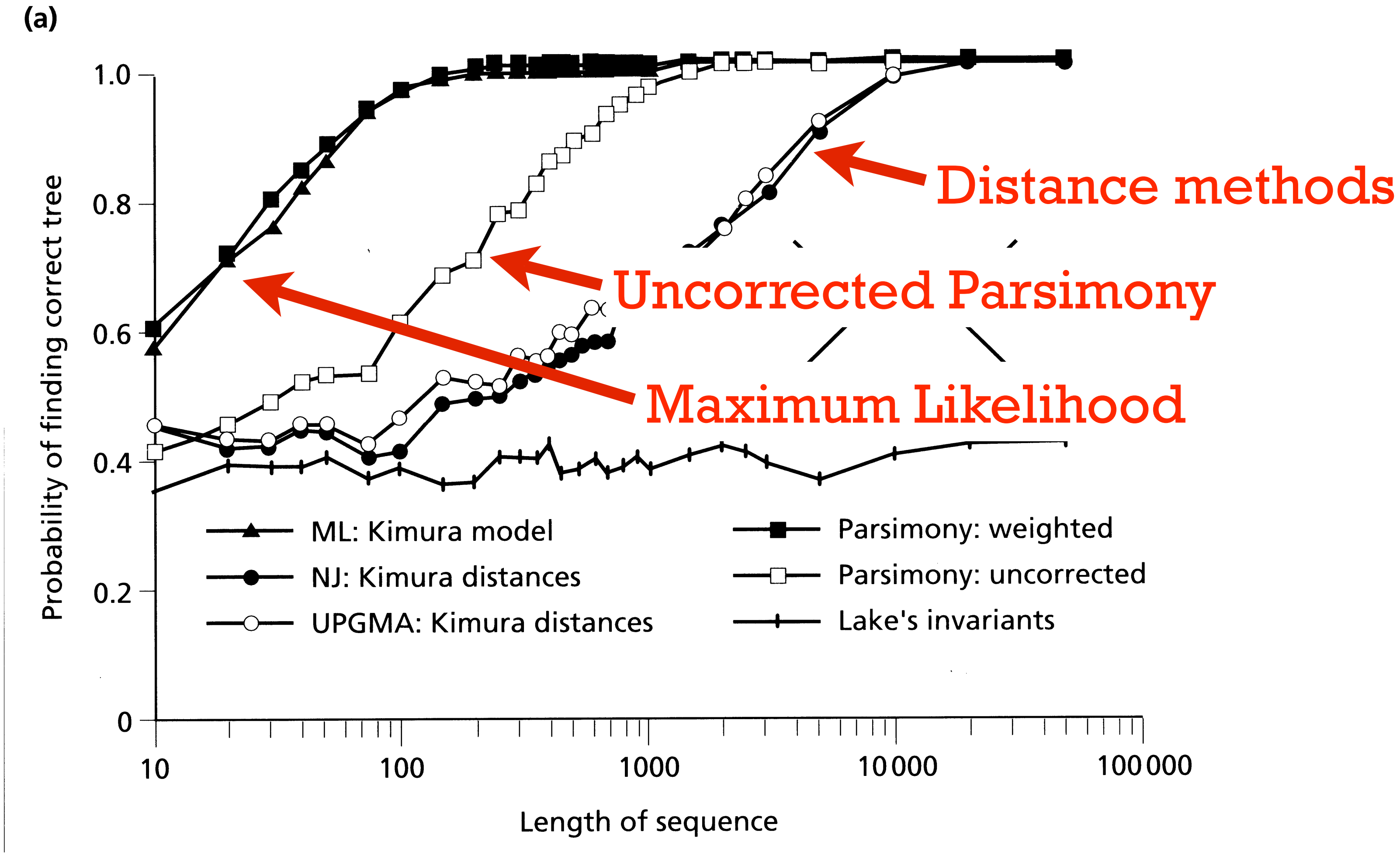


Figure 4.8: Sequential addition carried out on the data in Table 1.1. The species names A–E correspond, respectively, to the names Alpha through Epsilon in Table 1.1. Sequential addition ends up with four trees tied for best. None of these is actually the most parsimonious tree.

III. Performance of Parsimony

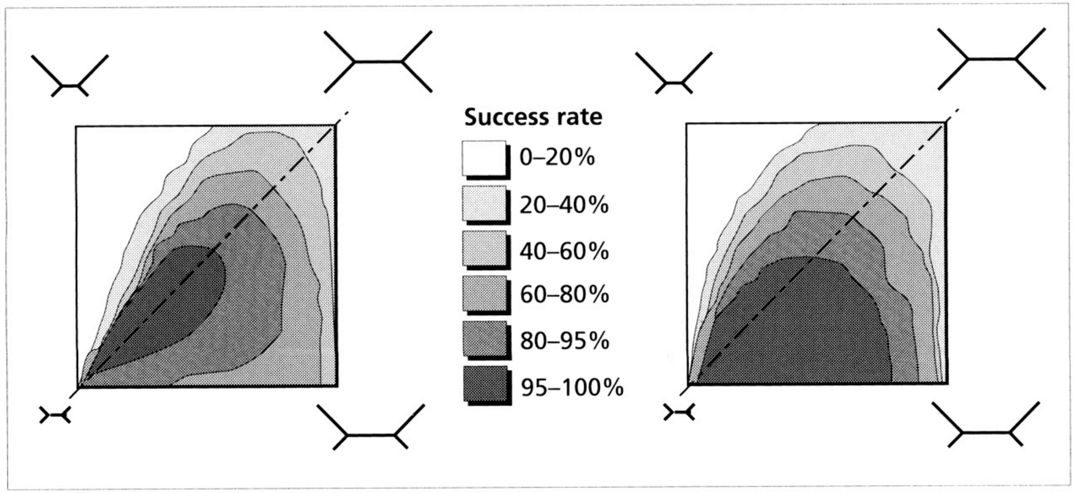


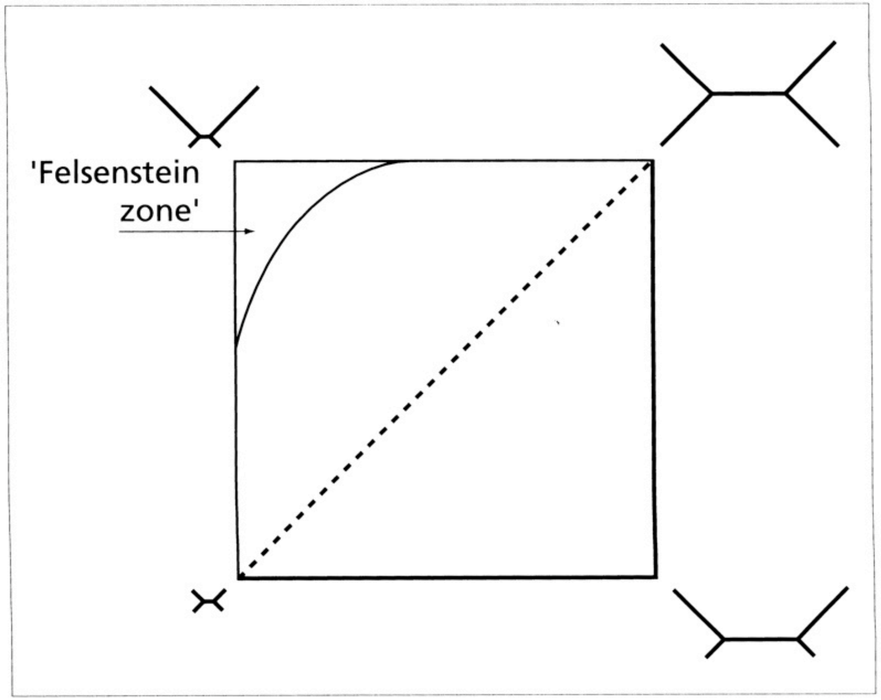
III. Performance of Parsimony



IV. Problems with Parsimony

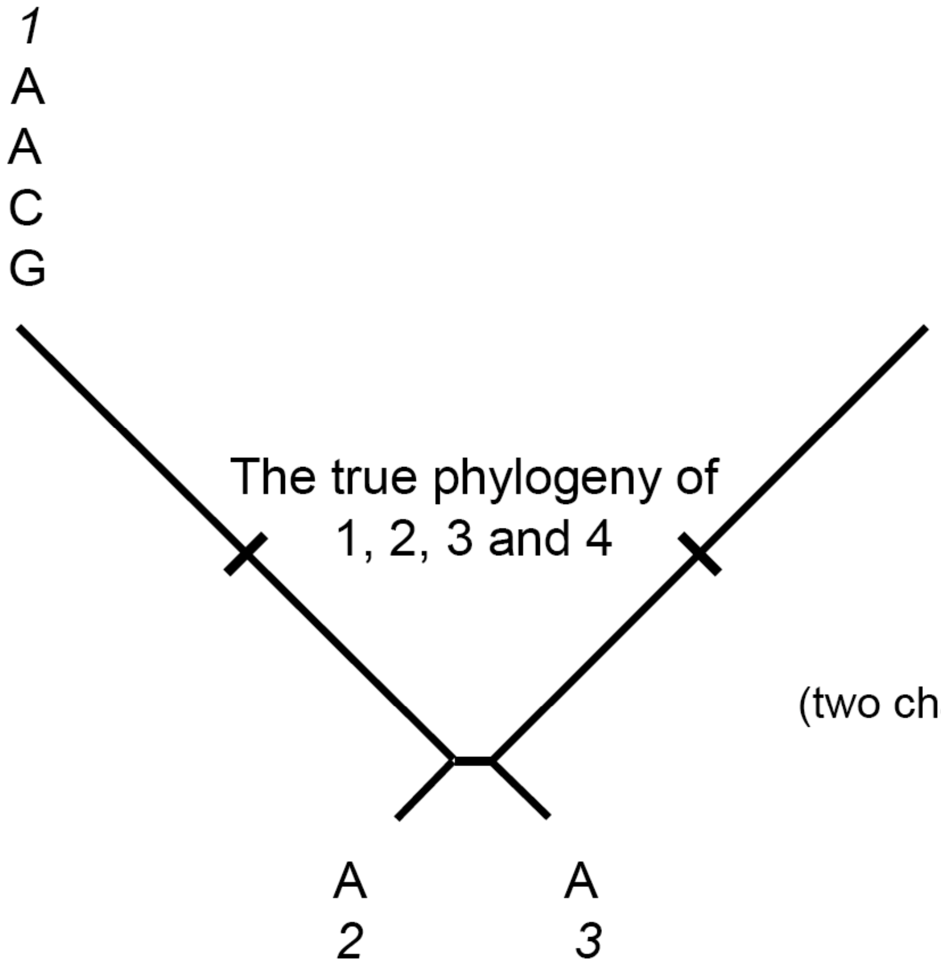
- Substitutional saturation.
- Long-branch attraction & the dreaded “Felsenstein Zone”





In the Felsenstein Zone

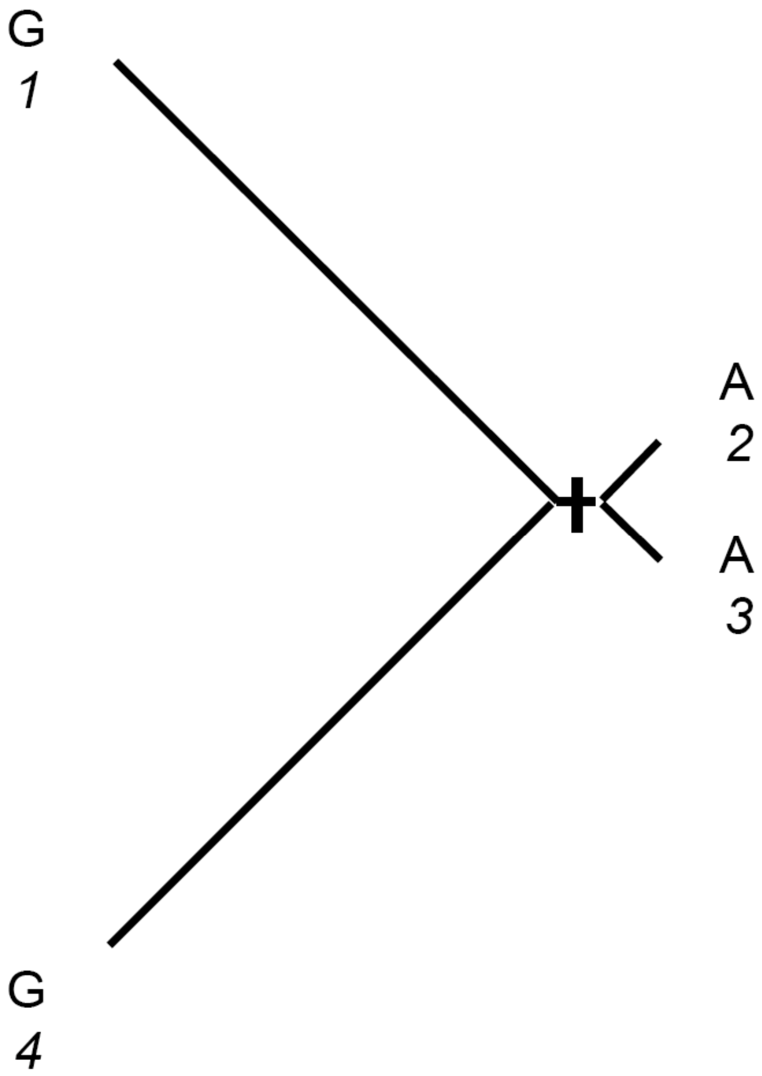
Pattern type



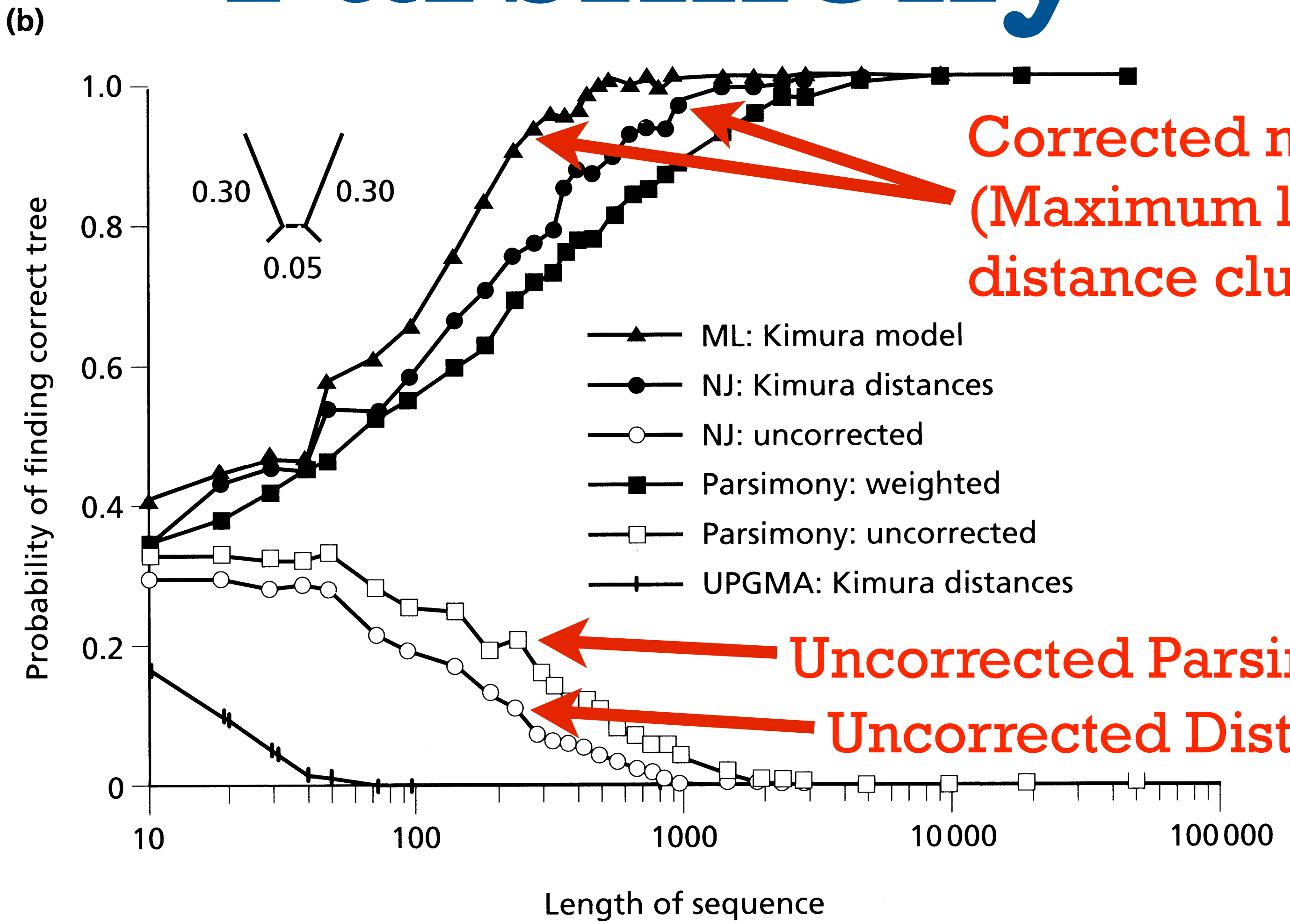
The true phylogeny of
1, 2, 3 and 4

(two changes required on true tree)

In the Felsenstein Zone



IV. Problems with Parsimony



Corrected methods (Maximum likelihood & distance clustering)

Uncorrected Parsimony

Uncorrected Distances

Models of DNA Evolution

- DNA distances.
- Using models of evolution to ‘correct’ DNA distance estimates.

Uncorrected ‘p’

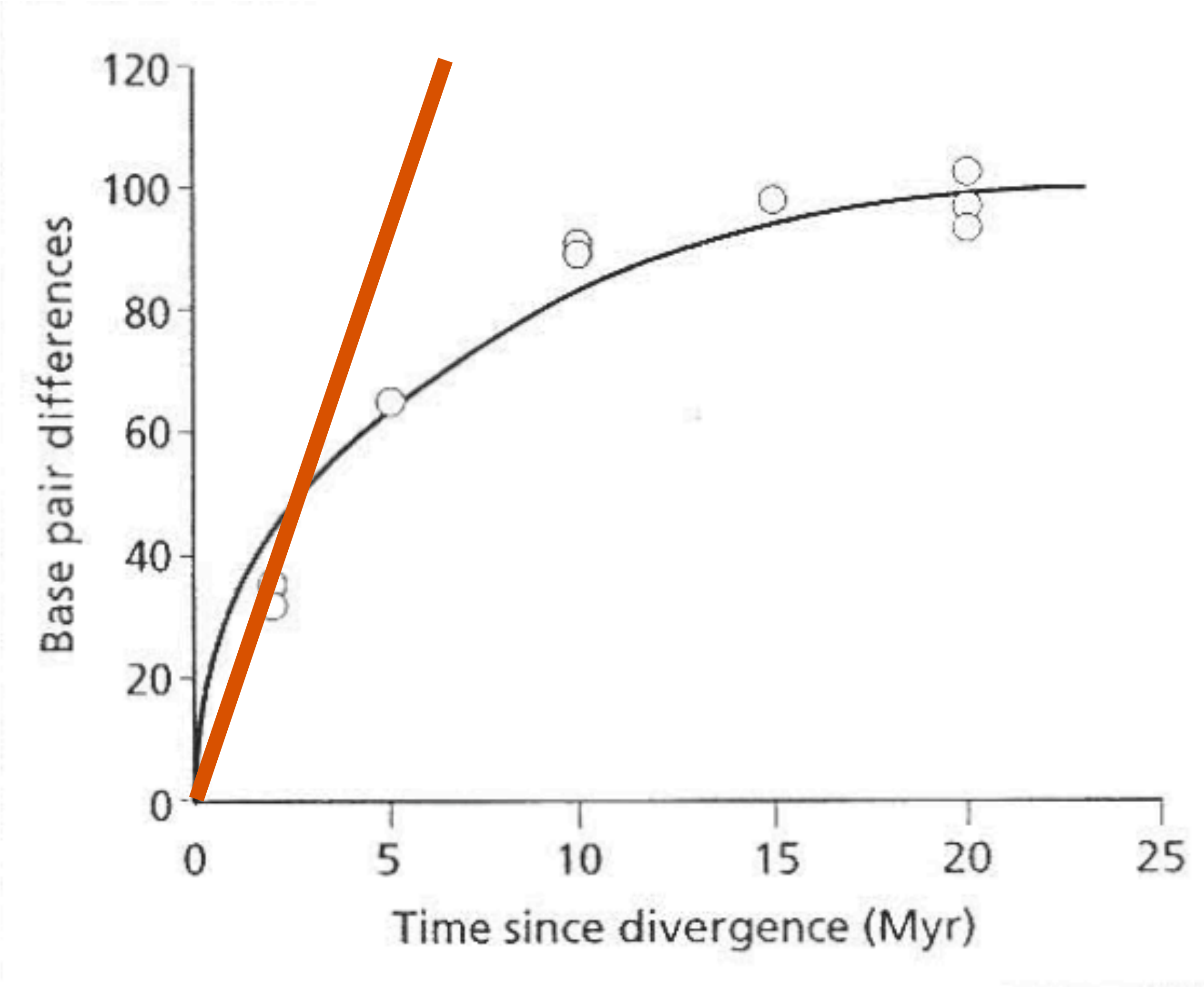
$$p = n_d / n$$

p = proportion of nucleotides that differ between two sequences.

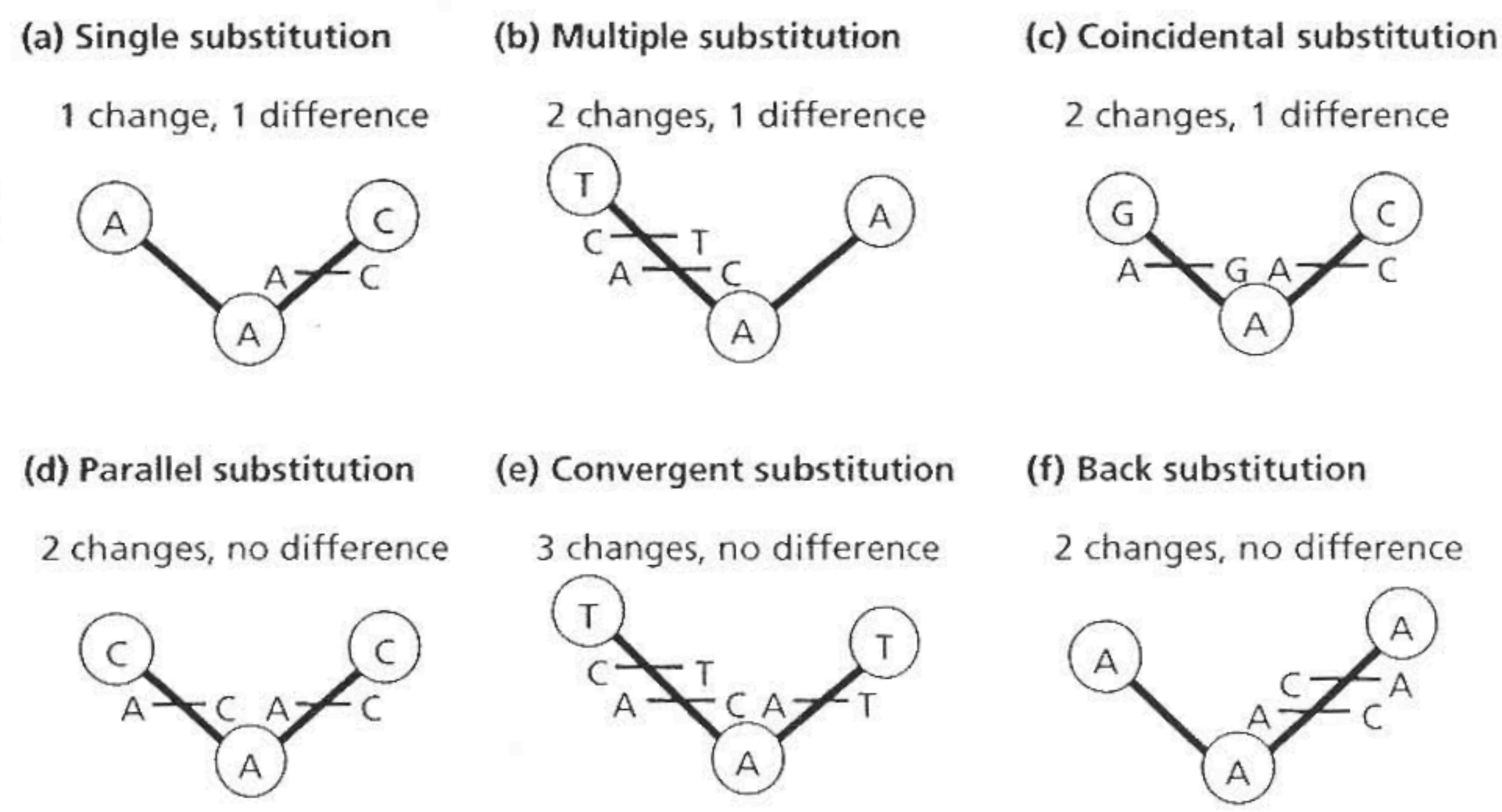
nd = number of differences.

n = total sequence length.

Saturation Curve



Types of Changes



(a) Single substitution

1 change, 1 difference

(b) Multiple substitution

2 changes, 1 difference

(c) Coincidental substitution

2 changes, 1 difference

(d) Parallel substitution

2 changes, no difference

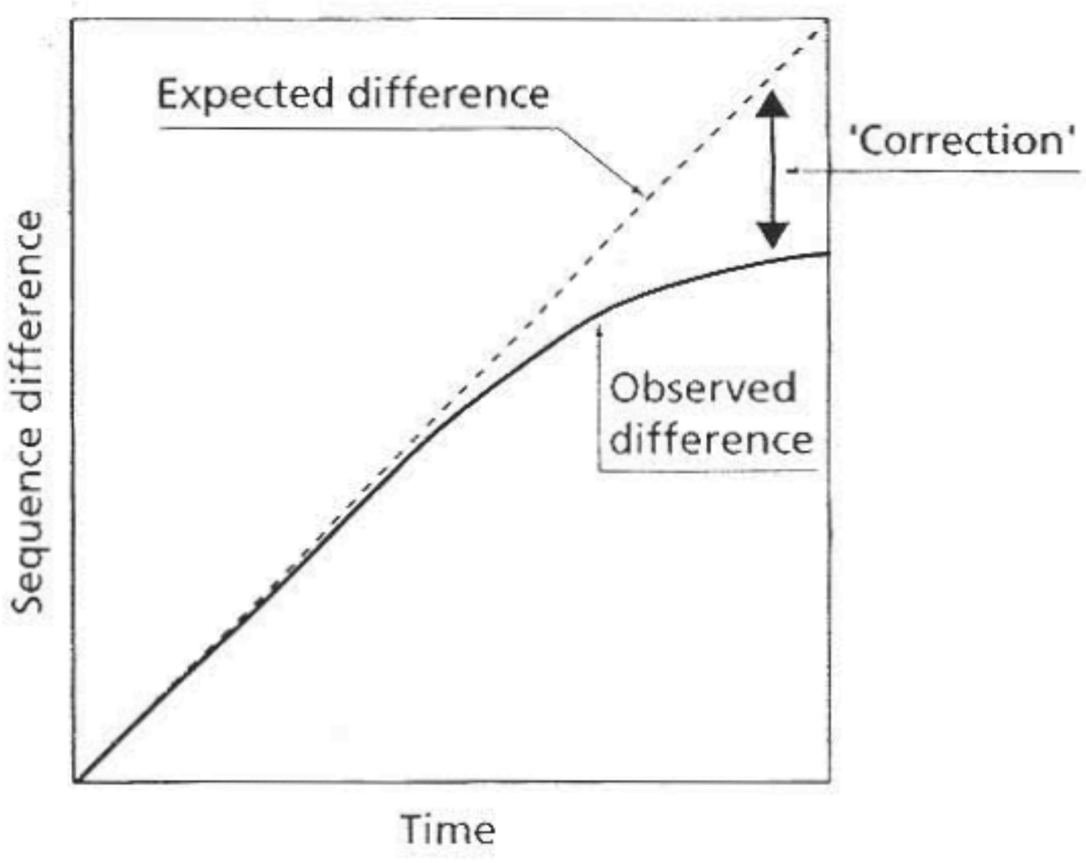
(e) Convergent substitution

3 changes, no difference

(f) Back substitution

2 changes, no difference

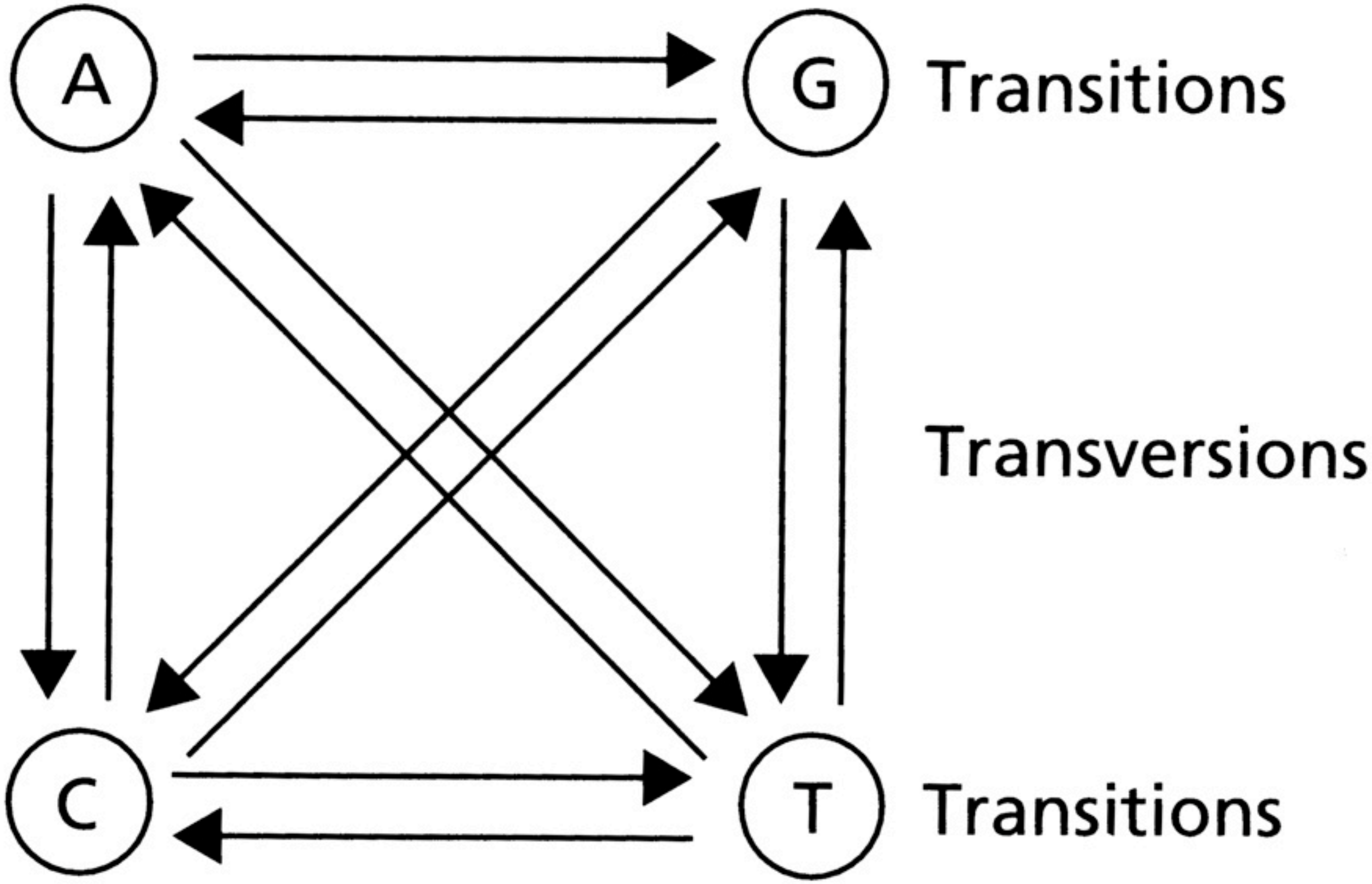
Correction of Distances



Models of DNA Evolution

1. Different types of substitutions (transitions, transversions, etc.).
2. Different nucleotide frequencies.
3. Rate variation among sites.
4. Proportion of sites lacking polymorphism.

Models of DNA Evolution

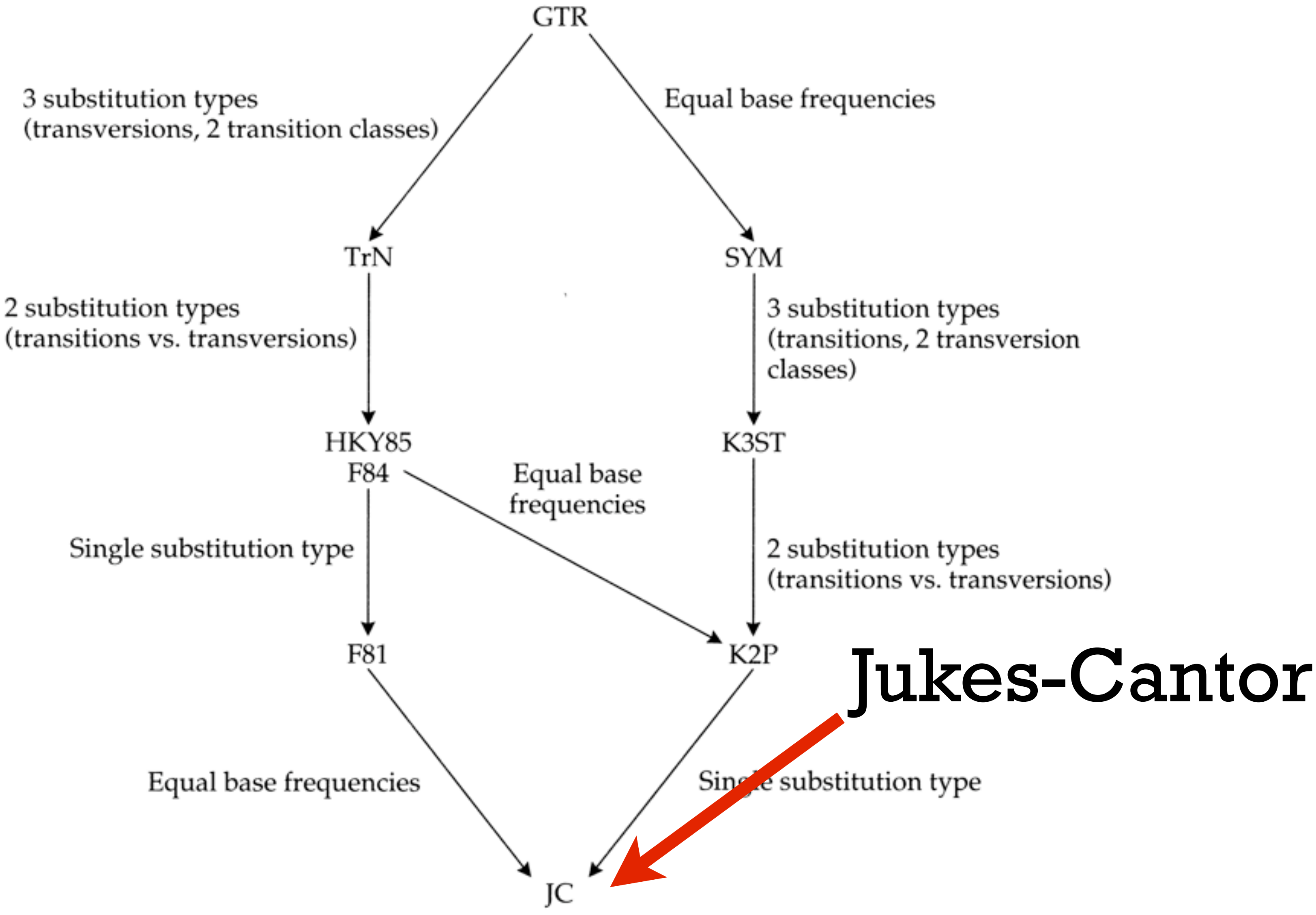


Transitions

Transversions

Transitions

Models of DNA Evolution



GTR

3 substitution types
(transversions, 2 transition classes)

Equal base frequencies

TrN

SYM

2 substitution types
(transitions vs. transversions)

3 substitution types
(transitions, 2 transversion classes)

HKY85
F84

K3ST

Equal base frequencies

Single substitution type

2 substitution types
(transitions vs. transversions)

F81

K2P

Equal base frequencies

Single substitution type

JC

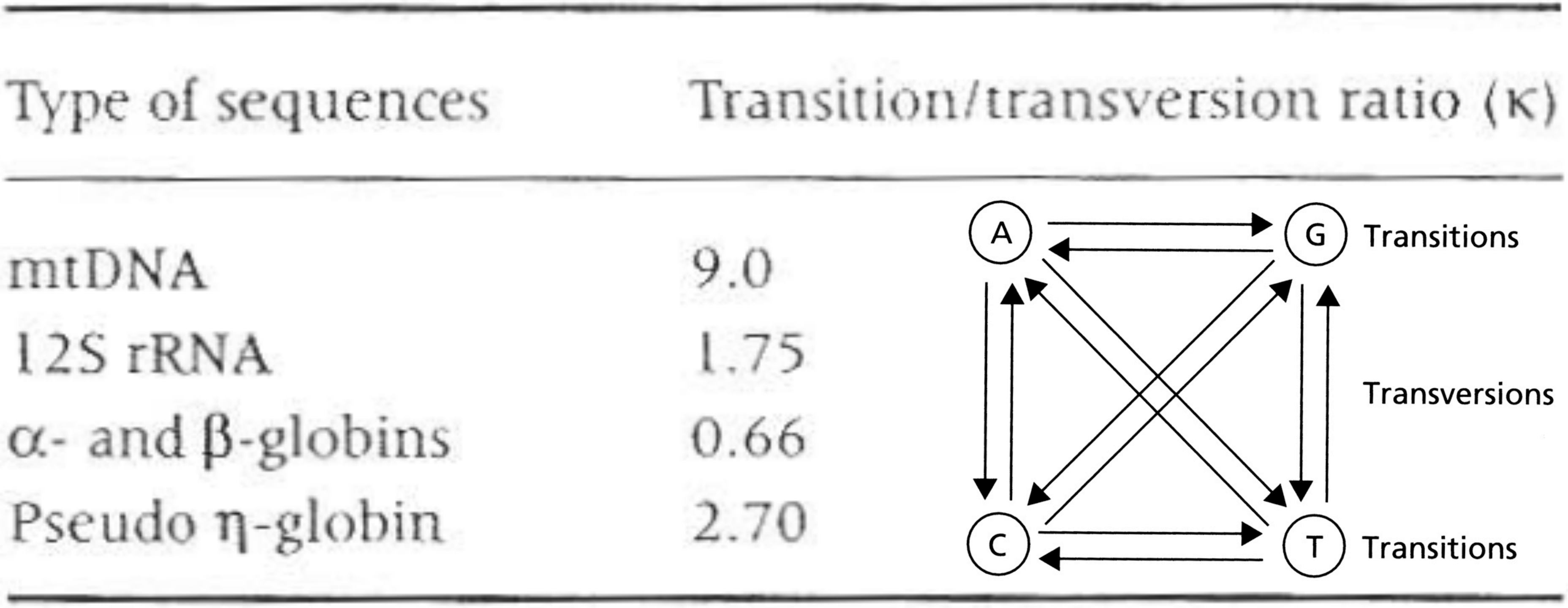
Jukes-Cantor

Jukes-Cantor

$$\mathbf{P}_t = \begin{bmatrix} \cdot & \alpha & \alpha & \alpha \\ \alpha & \cdot & \alpha & \alpha \\ \alpha & \alpha & \cdot & \alpha \\ \alpha & \alpha & \alpha & \cdot \end{bmatrix}, \qquad \mathbf{f} = [\tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}]$$

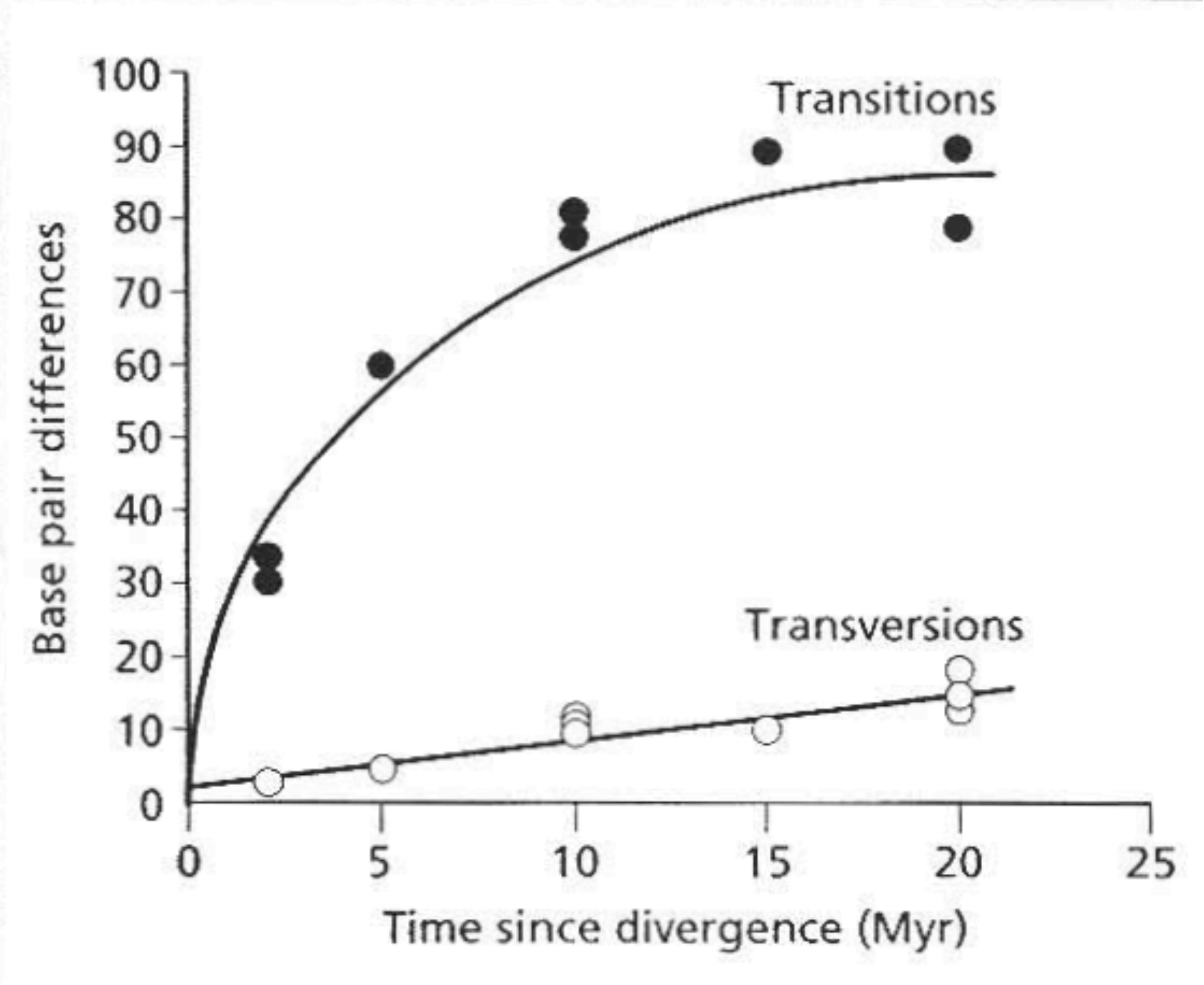
1. All types of substitutions equally likely.
2. All nucleotides present at the same frequency.
3. No rate variation among sites.
4. All sites potentially variable.

Transition Bias

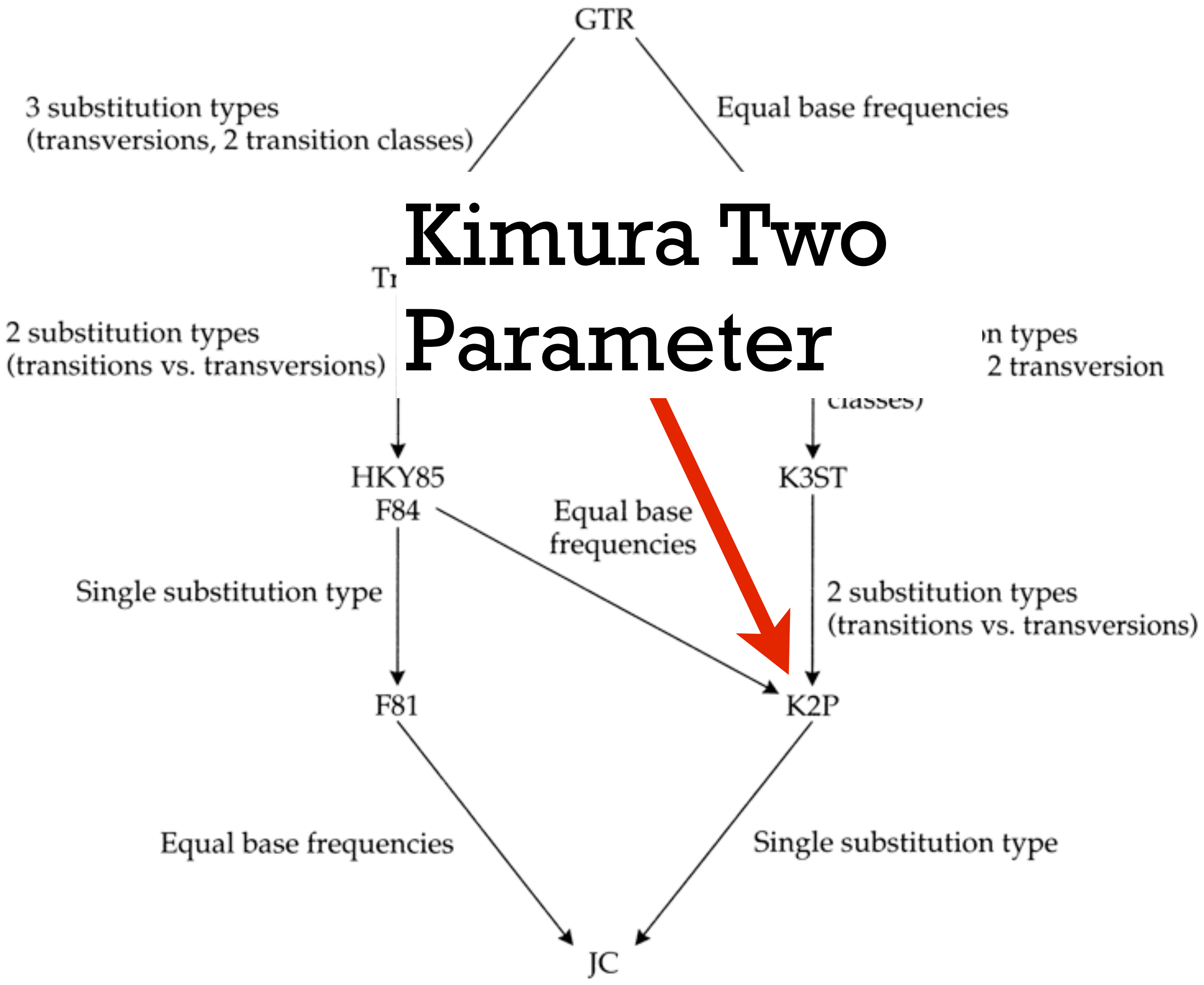


Type of sequences	Transition/transversion ratio (κ)
mtDNA	9.0
12S rRNA	1.75
α- and β-globins	0.66
Pseudo η-globin	2.70

Transition Bias



Models of DNA Evolution



GTR

3 substitution types
(transversions, 2 transition classes)

Equal base frequencies

Tr

2 substitution types
(transitions vs. transversions)

n types
2 transversion
classes)

HKY85
F84

K3ST

Equal base
frequencies

Single substitution type

2 substitution types
(transitions vs. transversions)

F81

K2P

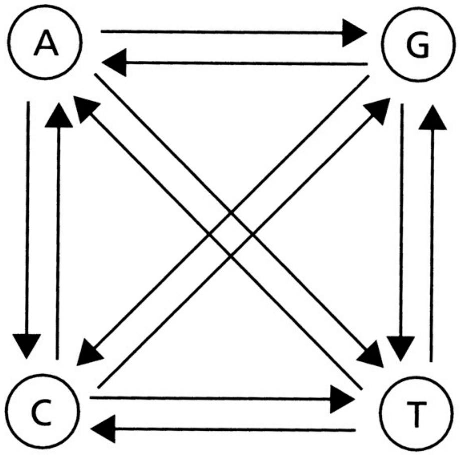
Equal base frequencies

Single substitution type

JC

Kimura Two
Parameter

Kimura Two Parameter



Transitions α

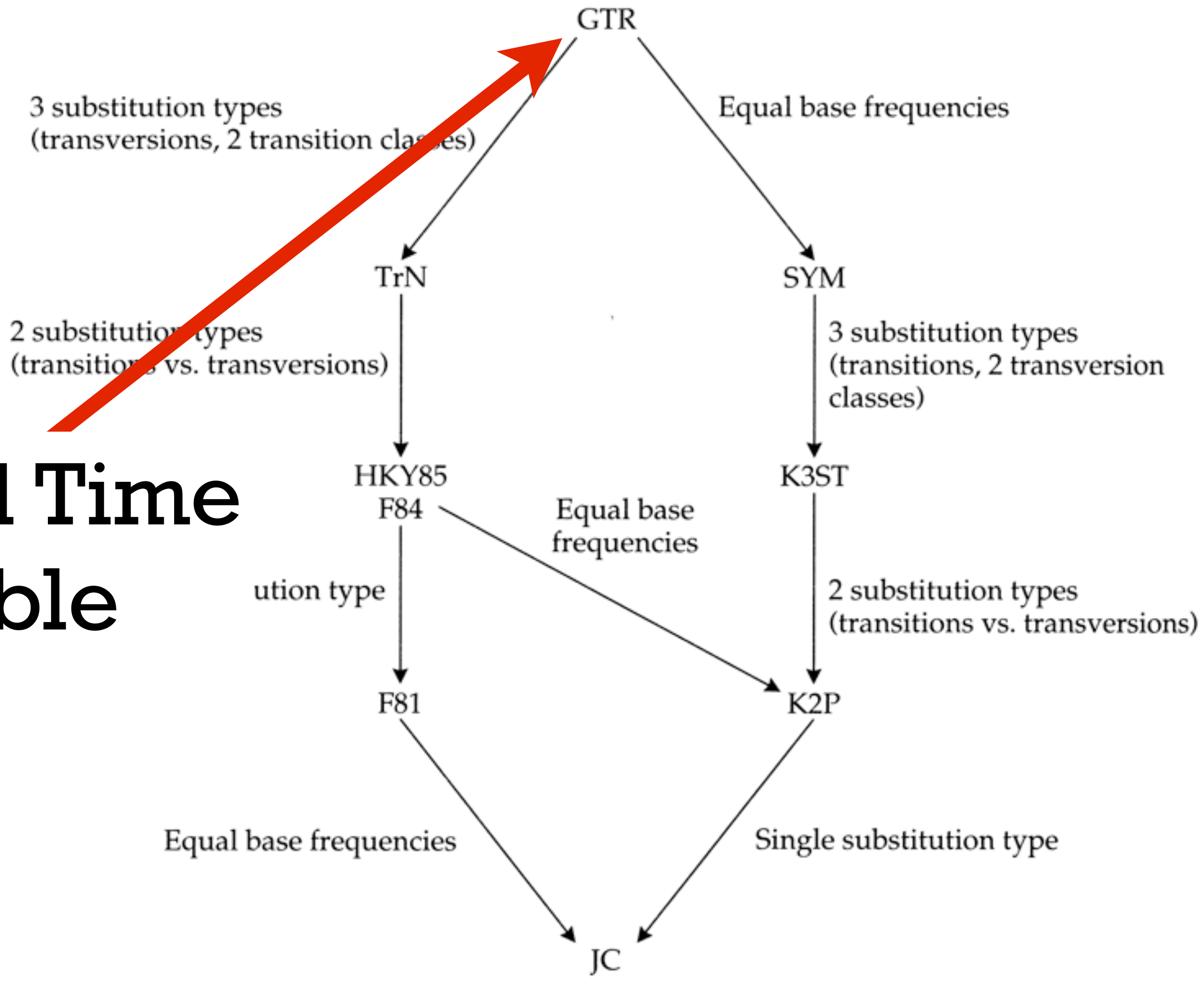
Transversions β

Transitions α

$$\mathbf{P}_t = \begin{array}{c} \begin{matrix} A & C & G & T \end{matrix} \\ \begin{bmatrix} . & \beta & \alpha & \beta \\ \beta & . & \beta & \alpha \\ \alpha & \beta & . & \beta \\ \beta & \alpha & \beta & . \end{bmatrix} \end{array}, \qquad \mathbf{f} = [\tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}].$$

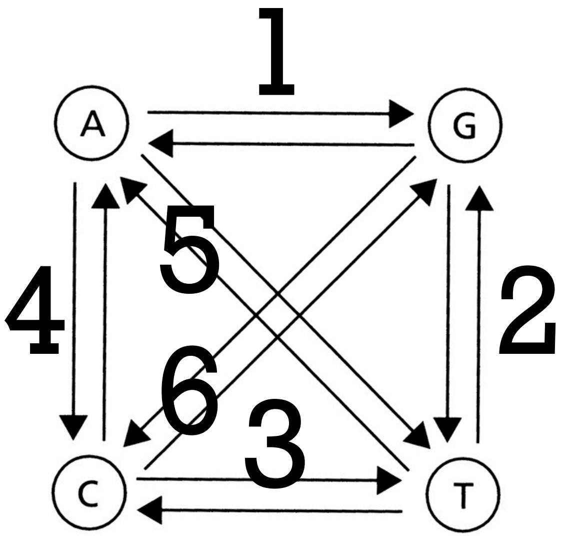
1. Transitions & transversions assigned different rates.
2. All nucleotides present at the same frequency.
3. No rate variation among sites.
4. All sites potentially variable.

GTR or REV



General Time Reversible

GTR or REV



1. Each type of reversible substitution with its own rate.
2. Nucleotides frequencies vary.
3. No rate variation among sites.
4. All sites potentially variable.

Rate Parameters for ML

$$\mathbf{Q} = \begin{pmatrix} - & \alpha & \alpha & \alpha \\ \alpha & - & \alpha & \alpha \\ \alpha & \alpha & - & \alpha \\ \alpha & \alpha & \alpha & - \end{pmatrix}$$

Jukes-Cantor (1969)

$$\mathbf{Q} = \begin{pmatrix} - & \beta & \alpha & \beta \\ \beta & - & \beta & \alpha \\ \alpha & \beta & - & \beta \\ \beta & \alpha & \beta & - \end{pmatrix}$$

Kimura (1980) “2-parameter”

$$\mathbf{Q} = \begin{pmatrix} - & \pi_{\mathrm{C}}\beta & \pi_{\mathrm{G}}\alpha & \pi_{\mathrm{T}}\beta \\ \pi_{\mathrm{A}}\beta & - & \pi_{\mathrm{G}}\beta & \pi_{\mathrm{T}}\alpha \\ \pi_{\mathrm{A}}\alpha & \pi_{\mathrm{C}}\beta & - & \pi_{\mathrm{T}}\beta \\ \pi_{\mathrm{A}}\beta & \pi_{\mathrm{C}}\alpha & \pi_{\mathrm{G}}\beta & - \end{pmatrix}$$

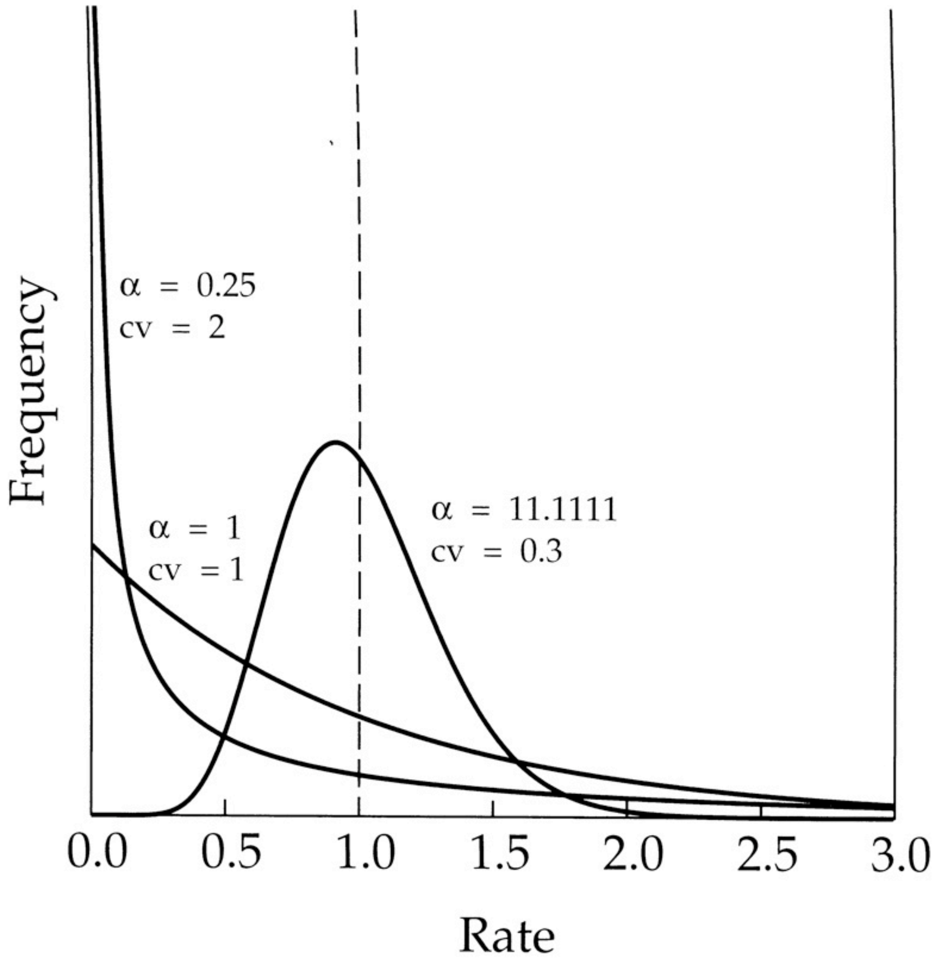
Hasegawa-Kishino-Yano (1985)

$$\mathbf{Q} = \begin{pmatrix} \pi_{\mathrm{A}} r_{\mathrm{AA}} & \pi_{\mathrm{C}} r_{\mathrm{AC}} & \pi_{\mathrm{G}} r_{\mathrm{AG}} & \pi_{\mathrm{T}} r_{\mathrm{AT}} \\ \pi_{\mathrm{A}} r_{\mathrm{CA}} & \pi_{\mathrm{C}} r_{\mathrm{CC}} & \pi_{\mathrm{G}} r_{\mathrm{CG}} & \pi_{\mathrm{T}} r_{\mathrm{CT}} \\ \pi_{\mathrm{A}} r_{\mathrm{GA}} & \pi_{\mathrm{C}} r_{\mathrm{GC}} & \pi_{\mathrm{G}} r_{\mathrm{GG}} & \pi_{\mathrm{T}} r_{\mathrm{GT}} \\ \pi_{\mathrm{A}} r_{\mathrm{TA}} & \pi_{\mathrm{C}} r_{\mathrm{TC}} & \pi_{\mathrm{G}} r_{\mathrm{TG}} & \pi_{\mathrm{T}} r_{\mathrm{TT}} \end{pmatrix}$$

General-Time Reversible

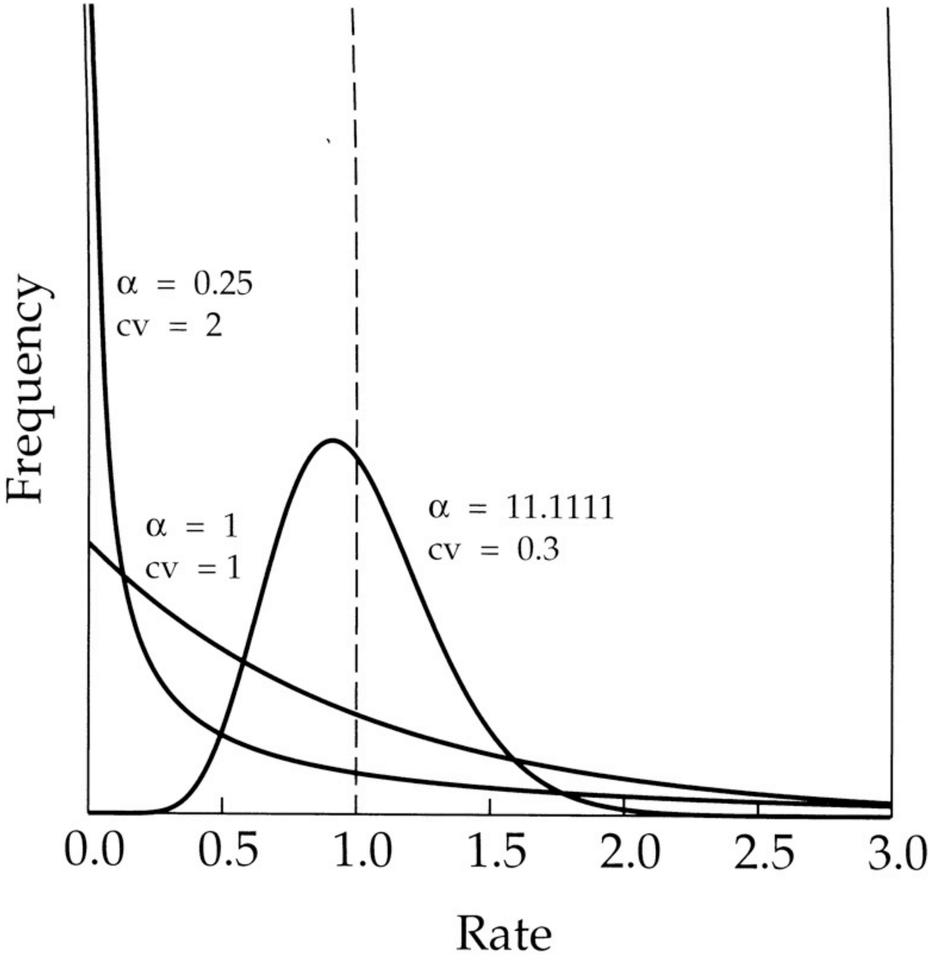
Rate Variation Among Sites

- **Gamma**
    - Shape parameter as the inverse of the squared coefficient of variation.
    - Small coefficient of variation results in rates normally distributed around 1.
    - Large coefficient results in many sites with a rate close to 0 and a few with very fast rates.

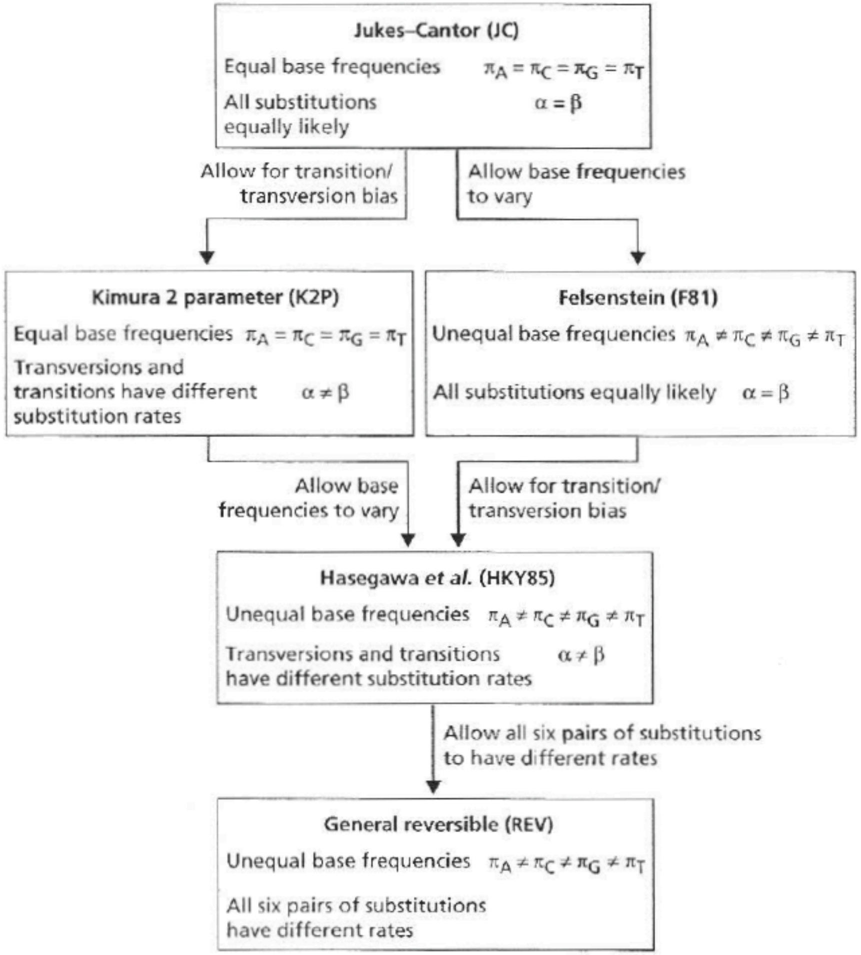


Rate Variation Among Sites

Type of sequences	α
Nuclear genes	
Albumin genes	1.05
Insulin genes	0.40
c-myc genes	0.47
Prolactin genes	1.37
16S-like rRNAs, stem region	0.29
16S-like rRNAs, loop region	0.58
ψη-globin pseudogenes	0.66
Viral genes	
Hepatitis B virus genomes	0.26
Mitochondrial genes	
12S rRNAs	0.16
Position 1 of four genes	0.18
Position 2 of four genes	0.08
Position 3 of four genes	1.58
D-loop region	0.17
Cytochrome *b*	0.44



Selecting a Model



Jukes–Cantor (JC)

Equal base frequencies $\pi_A = \pi_C = \pi_G = \pi_T$

All substitutions equally likely $\alpha = \beta$

Allow for transition/ transversion bias

Allow base frequencies to vary

Kimura 2 parameter (K2P)

Equal base frequencies $\pi_A = \pi_C = \pi_G = \pi_T$

Transversions and transitions have different substitution rates $\alpha \neq \beta$

Felsenstein (F81)

Unequal base frequencies $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$

All substitutions equally likely $\alpha = \beta$

Allow base frequencies to vary

Allow for transition/ transversion bias

Hasegawa *et al.* (HKY85)

Unequal base frequencies $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$

Transversions and transitions have different substitution rates $\alpha \neq \beta$

Allow all six pairs of substitutions to have different rates

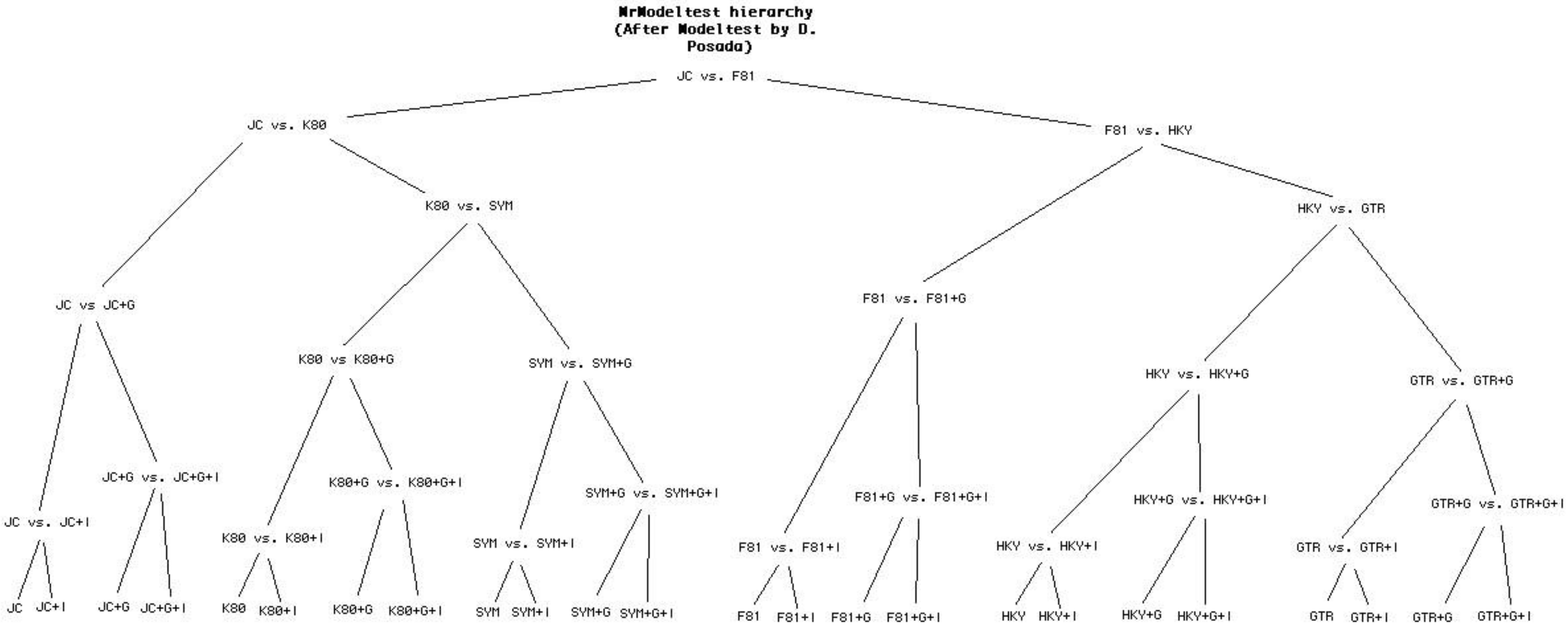
General reversible (REV)

Unequal base frequencies $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$

All six pairs of substitutions have different rates

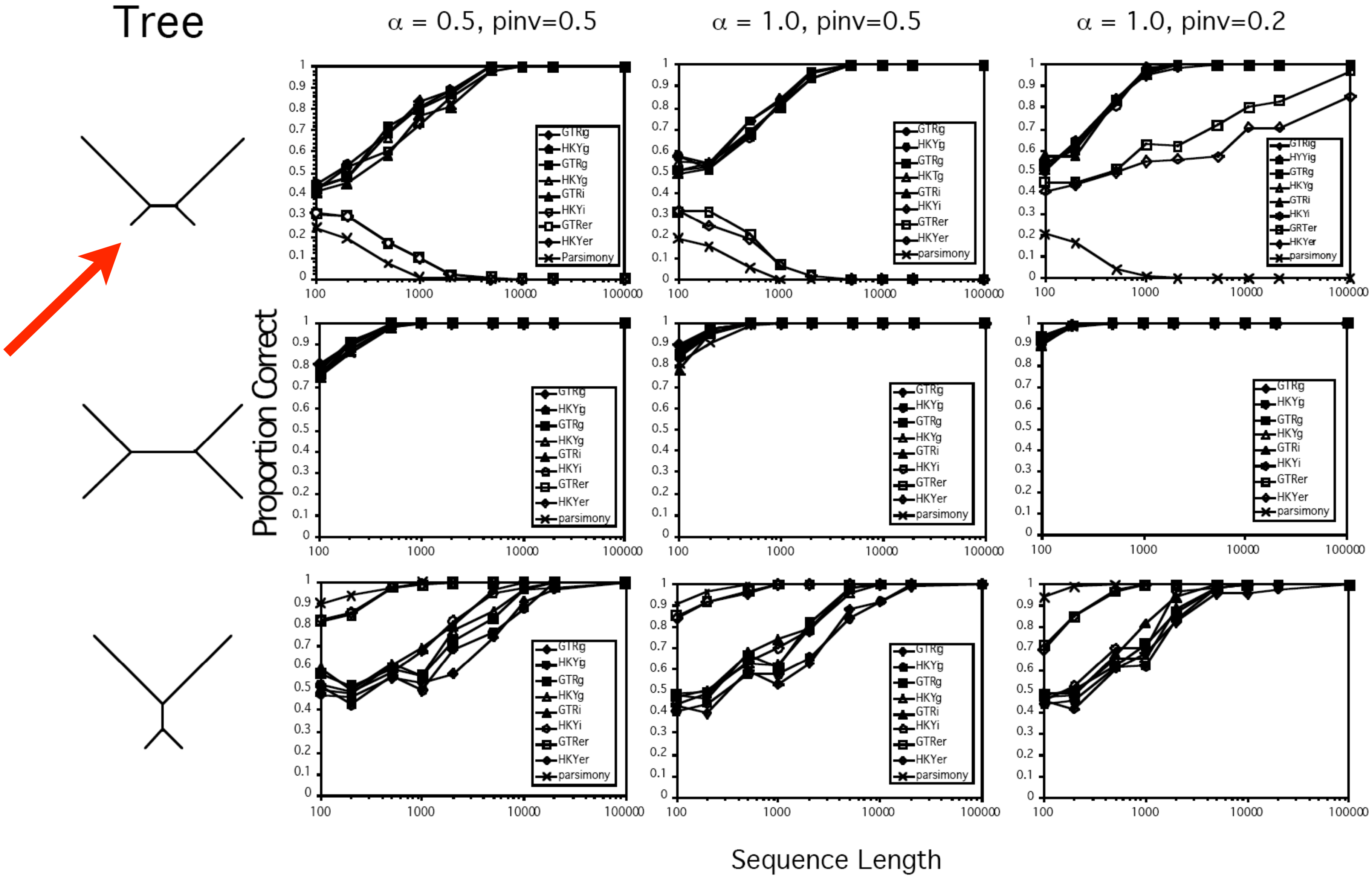
Selecting a Model

- Use of Likelihood ratio test to identify most appropriate model.

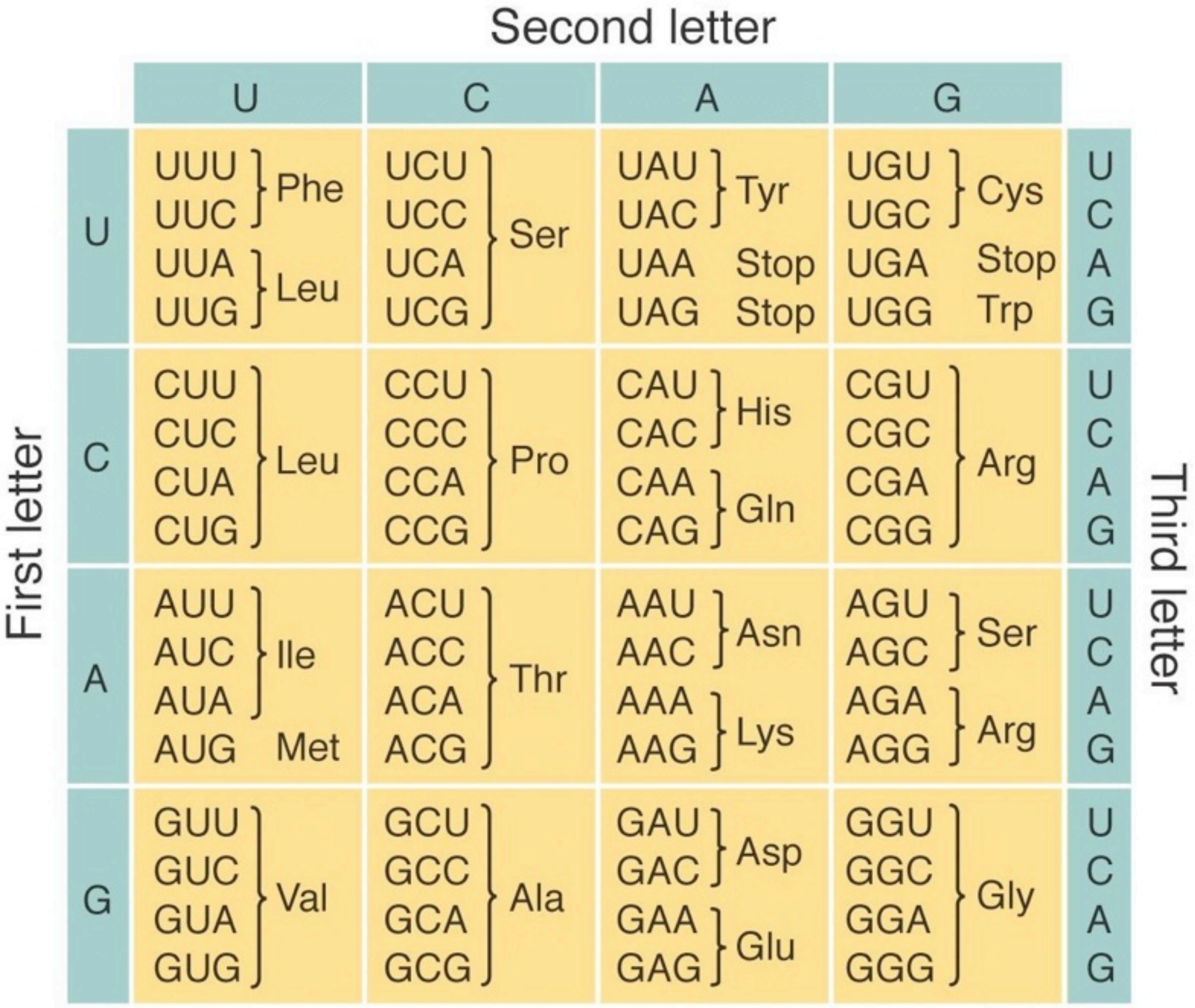


Importance of Model Selection

Expect Long Branch Attraction



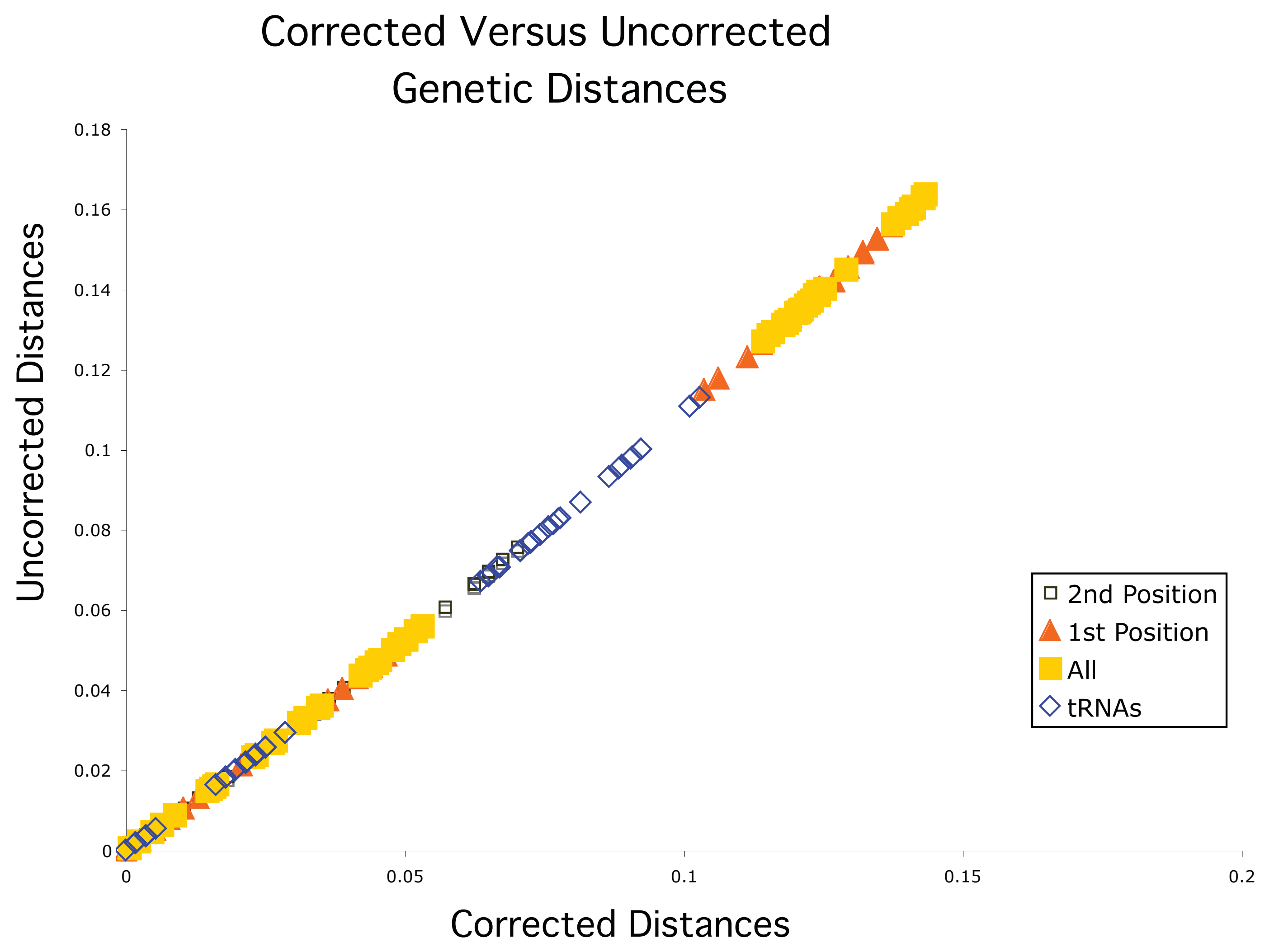
Synonymous v. Nonsynonymous



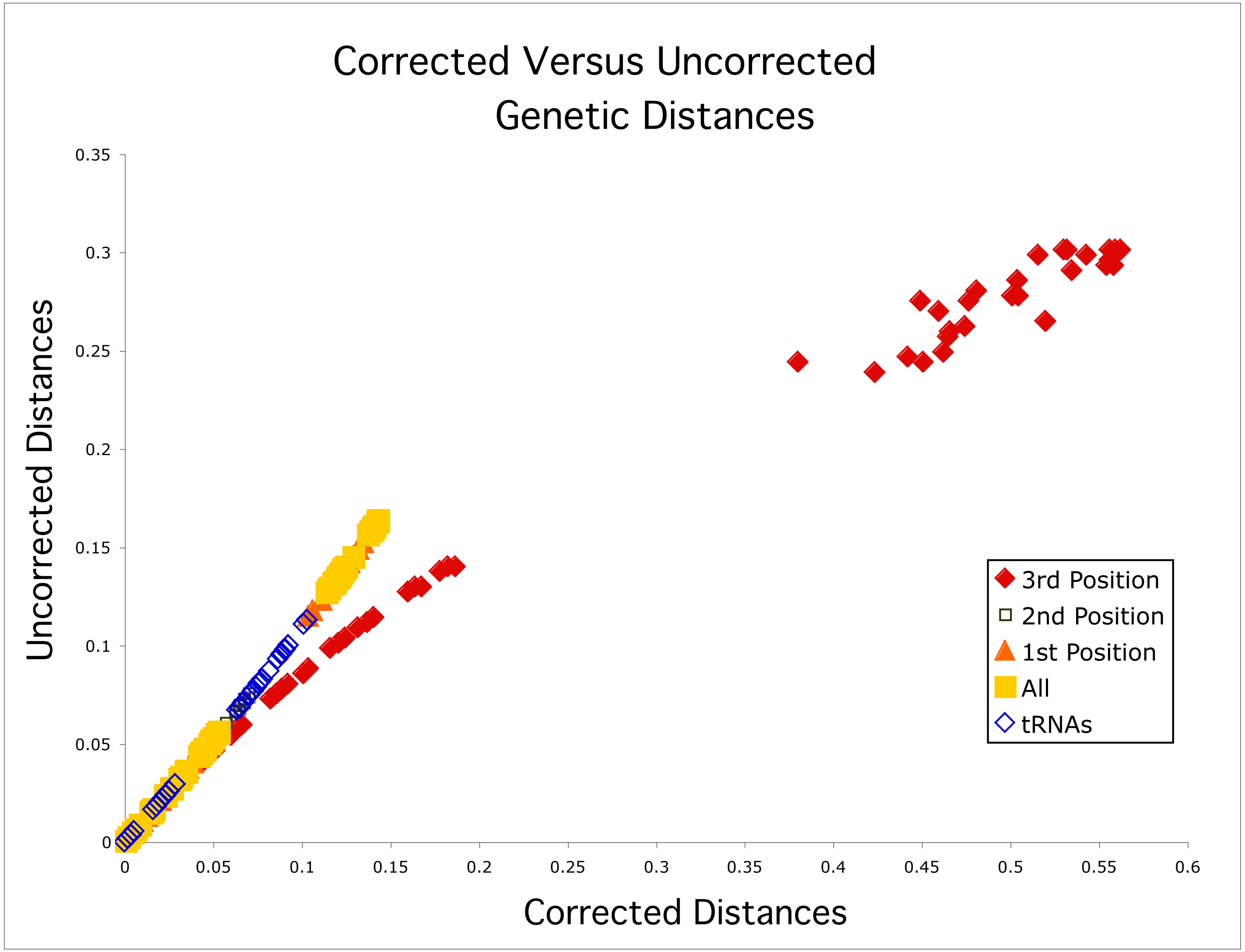
Second letter

First letter	U		C		A		G		Third letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

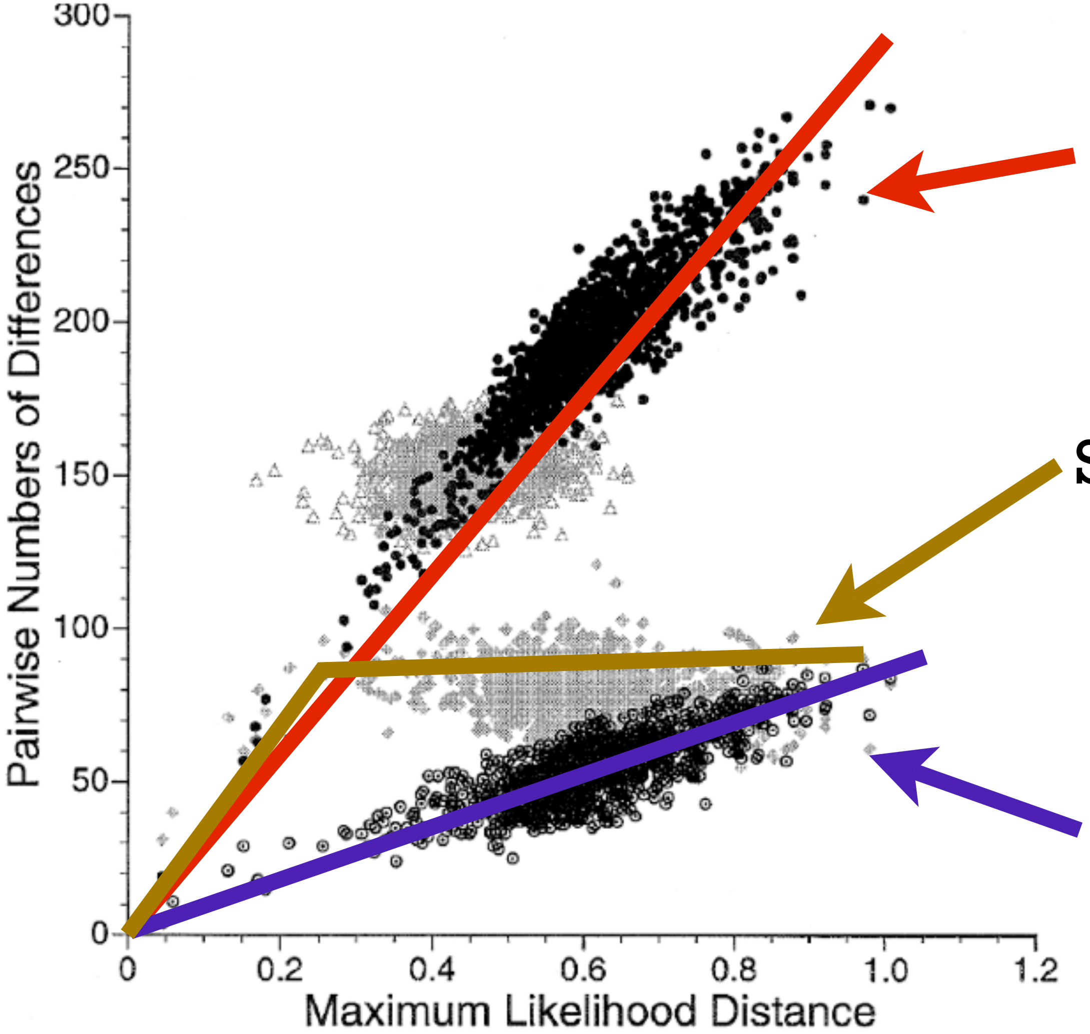
Corrected v. Uncorrected Distances



Corrected v. Uncorrected Distances



Saturation

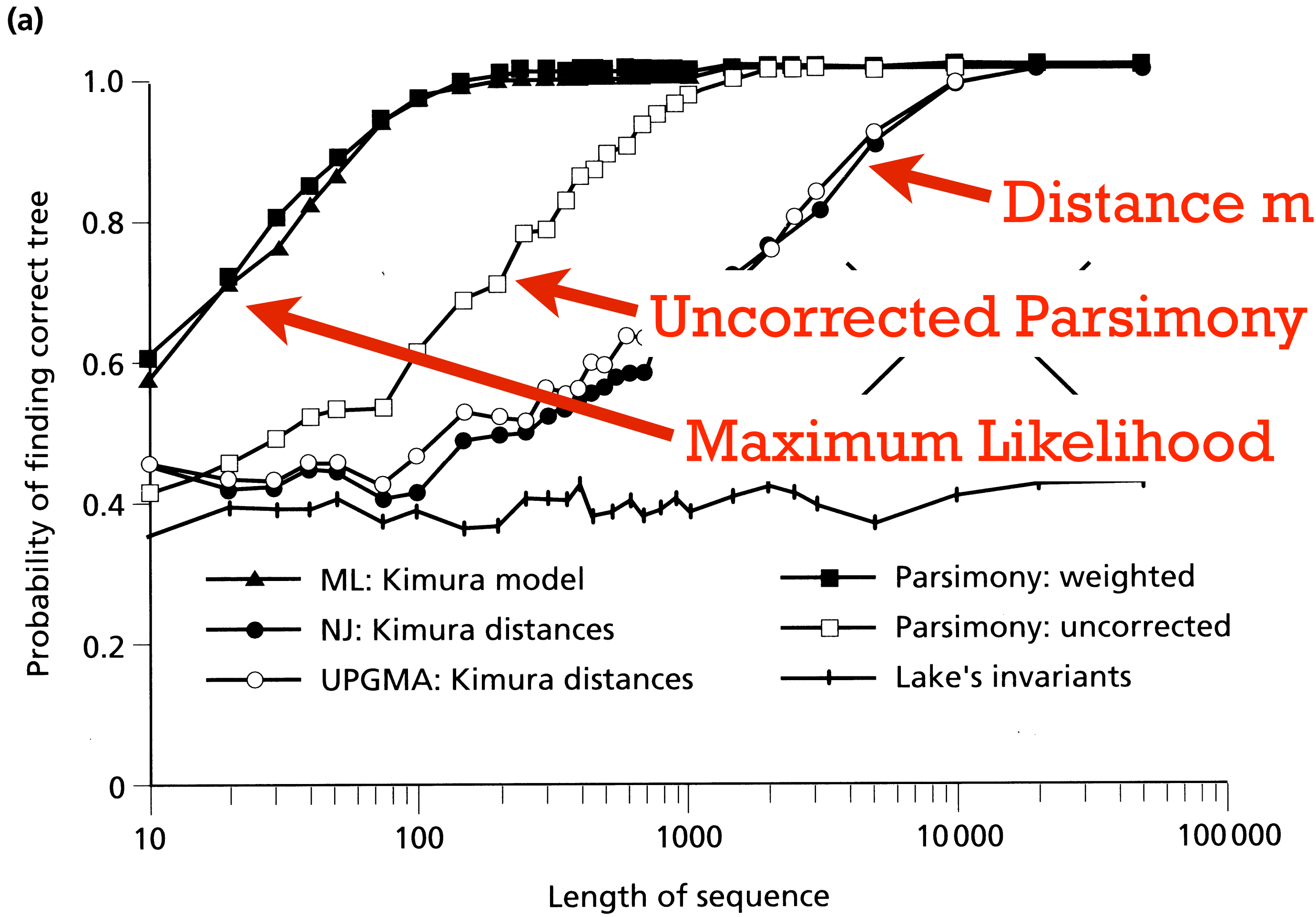


Non-synonymous
Substitutions

Silent Substitutions

tRNA Changes

Why Use Models?

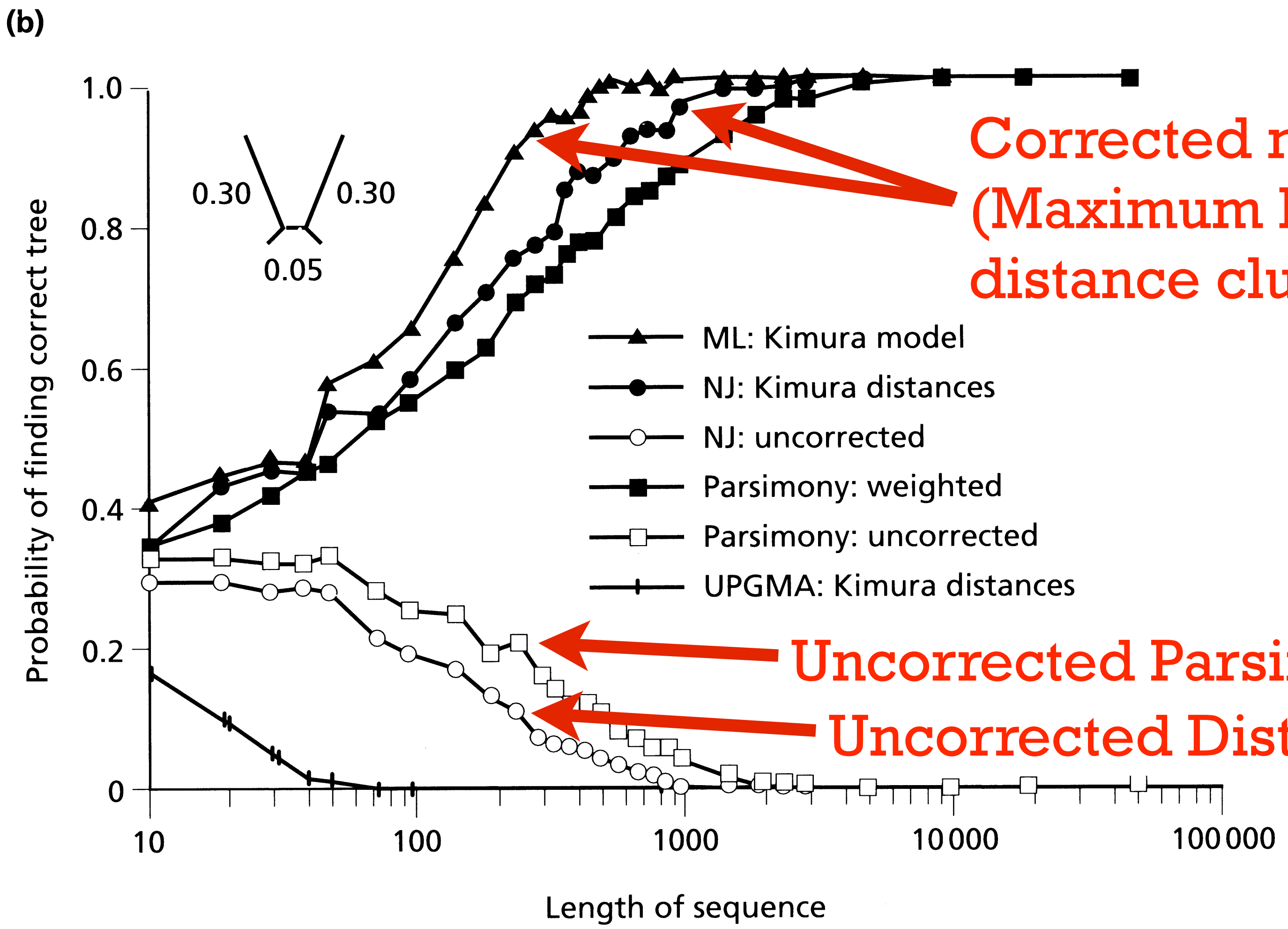


Distance methods

Uncorrected Parsimony

Maximum Likelihood

Why Use Models?



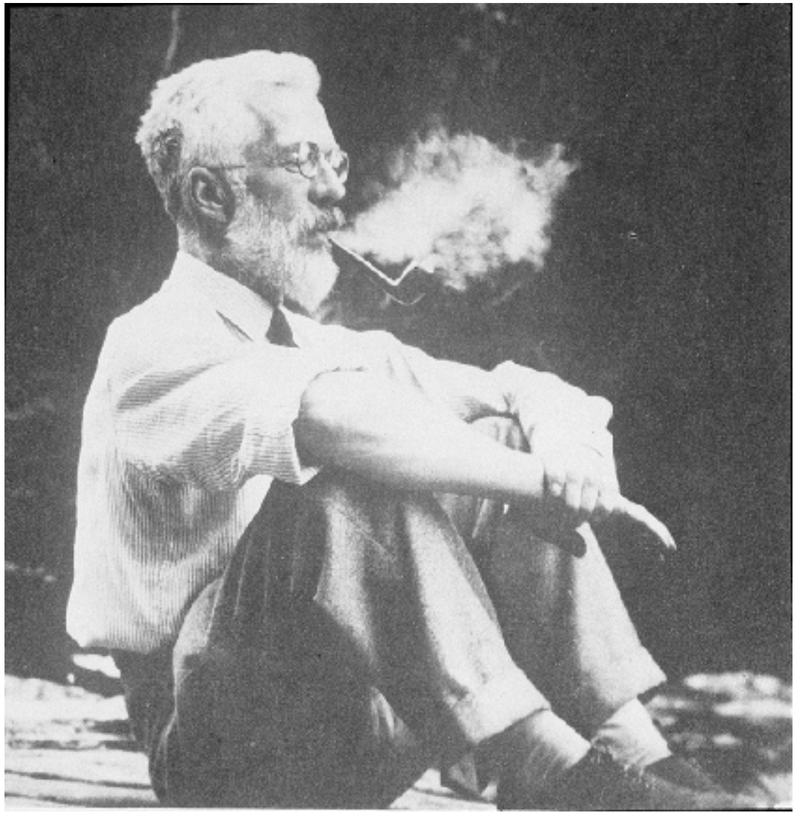
Corrected methods (Maximum likelihood & distance clustering)

Uncorrected Parsimony

Uncorrected Distances

Maximum Likelihood

- Invented by R. A. Fisher.
- Good properties:
  - Consistency (converges on correct parameters).
  - Efficiency (smallest possible variance around true parameter value).



R. A. Fisher

Maximum Likelihood

- Typical thinking
  - Prob(Hypothesis|Data)
  - Prob(Tree|Data,Model)
- Maximum Likelihood thinking
  - Prob(Data|Hypothesis)
  - Prob(observed sequences|tree, model of evolution) = Likelihood(tree, model of evolution)

Likelihood Estimation

- What is the probability of obtaining a heads (p) with the toss of an unknown coin?
- Data: HHTTHTHHTTT

Likelihood Estimation

- What is the probability of obtaining a heads (p) with the toss of an unknown coin?
- Data: HHTTHTHHTTT
- $L = \text{Prob}(D|p) = pp(1-p)(1-p)p(1-p)pp(1-p)(1-p)(1-p) = \mathbf{p^5(1-p)^6}$

Calculating Maximum Likelihood Values

HHTTHTHHTTT

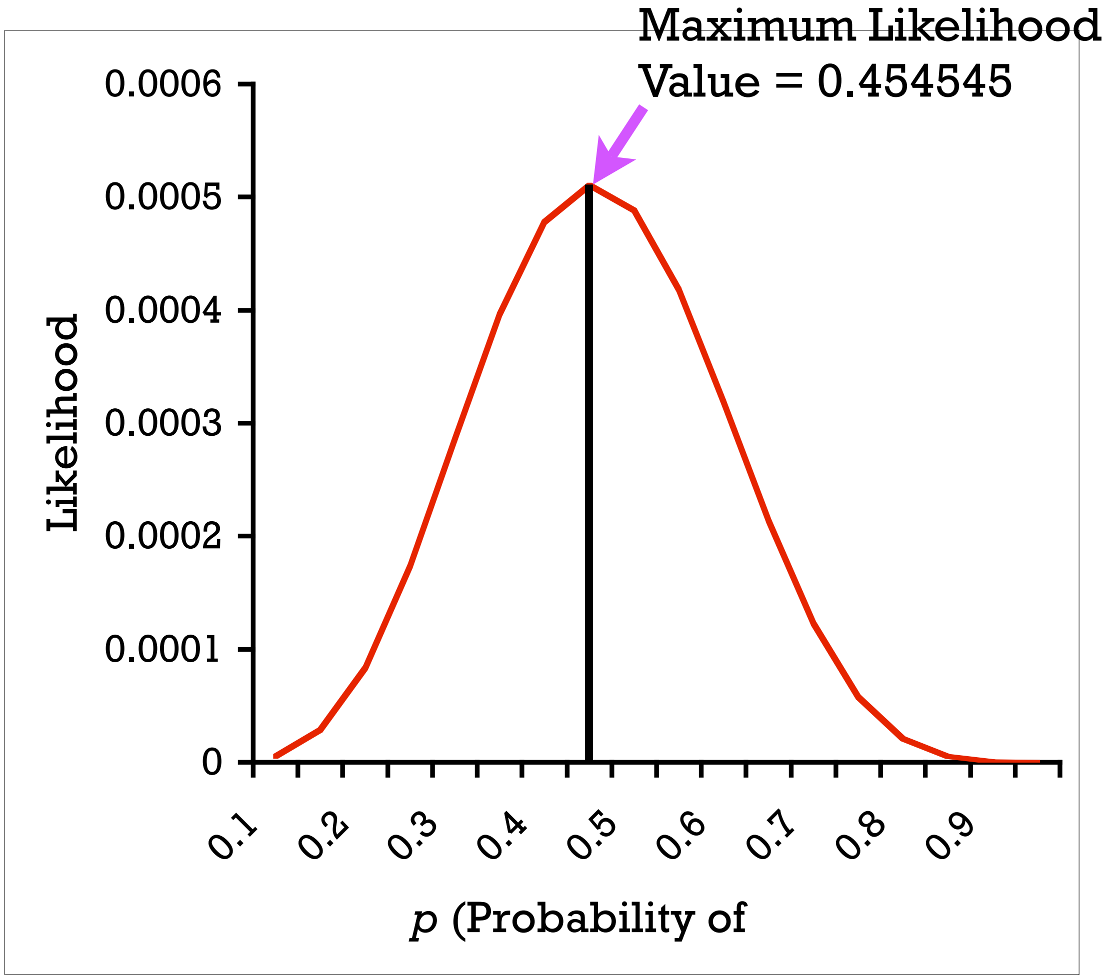
$L = p^5(1-p)^6$

If $p = 0.2$, $L = 8.4E-5$

If $p = 0.8$, $L = 2.1E-5$

If $p = 0.5$, $L = 4.9E-4$

5/11 = 0.454545



Calculating Maximum Likelihood Values

HHHTHHHHHHT

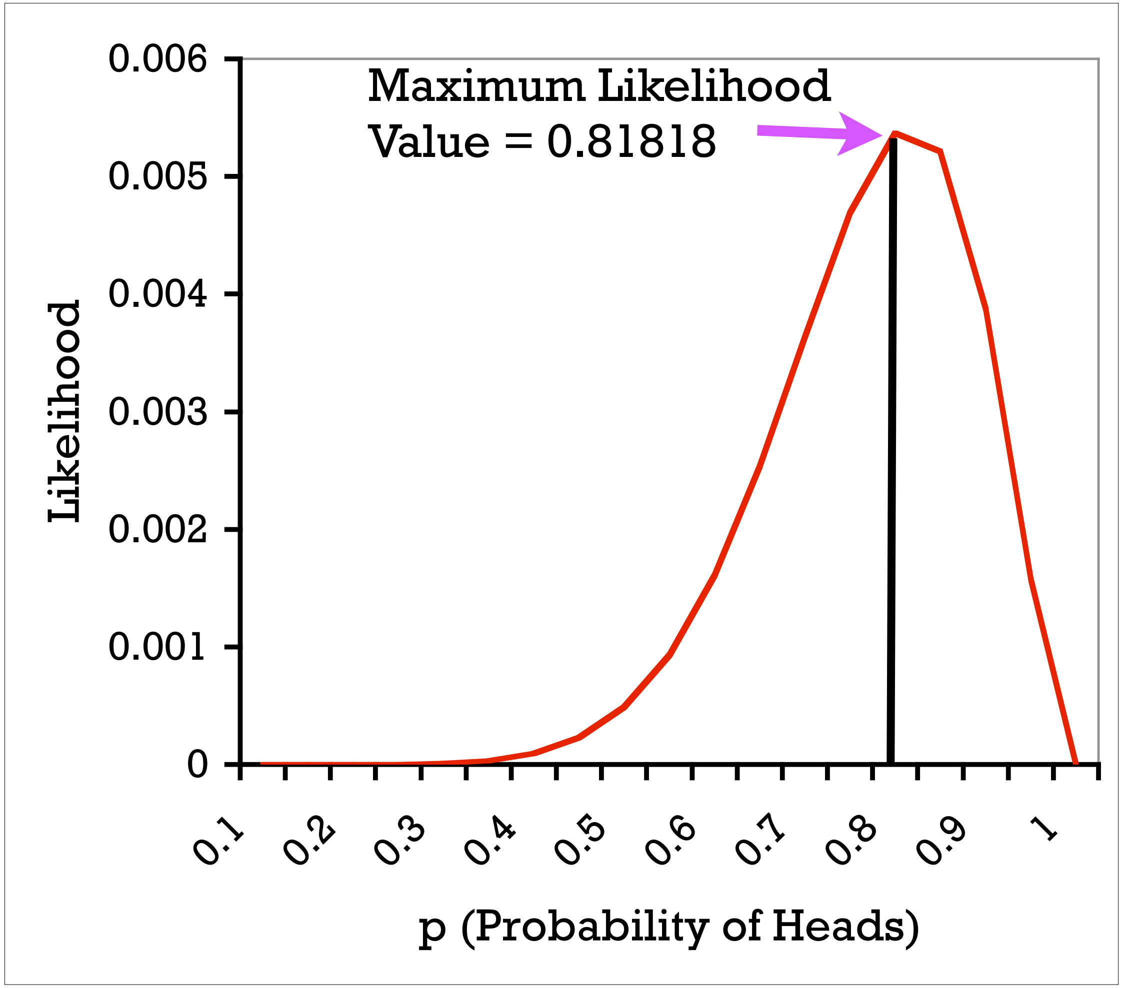
$L = p^9(1-p)^2$

If $p = 0.2$, $L = 3.3E-7$

If $p = 0.8$, $L = 0.005$

If $p = 0.5$, $L = 4.9E-4$

$9/11 = 0.818181$



Calculating a Trees ML Score

```
     1       j       N
(1)  C…GGACA…C…GTTTA…C
(2)  C…AGACA…C…CTCTA…C
(3)  C…GGATA…A…GTTAA…C
(4)  C…GGATA…G…CCTAG…C
```

$$L = \mathrm{Prob}(D|T) = \prod_{j=1}^{m} \mathrm{Prob}(D^{(j)}|T)$$

- Likelihood of tree equals product of likelihoods or that tree at each individual position.

Calculating a Trees ML Score

```
      1       j       N
(1)   C…GGACA…C…GTTTA…C
(2)   C…AGACA…C…CTCTA…C
(3)   C…GGATA…A…GTTAA…C
(4)   C…GGATA…G…CCTAG…C
```

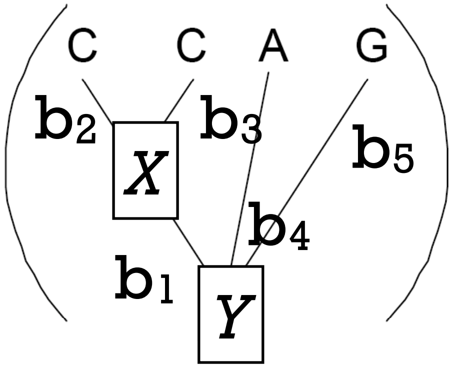
$$L = \mathrm{Prob}(D|T) = \prod_{j=1}^{m} \mathrm{Prob}(D^{(j)}|T)$$

- Likelihood of tree equals product of likelihoods or that tree at each individual position.

Calculating a Trees ML Score

Likelihood at site *j* =

Prob



Likelihood = $\text{Prob}(C,C,A,G,X,Y|T)$ =

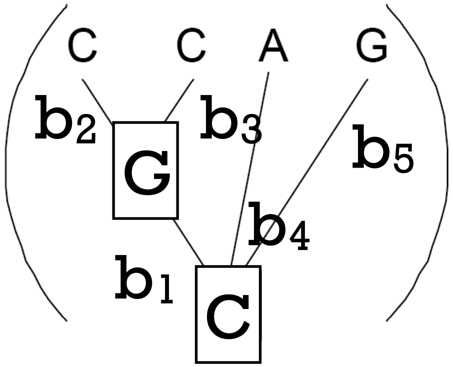
$\text{Prob}(Y)\quad \text{Prob}(X|Y,b_1)\quad \text{Prob}(C|X,b_2)\quad \text{Prob}(C|X,b_3)$

$\text{Prob}(A|Y,b_4)\quad \text{Prob}(G|Y,b_5)$

Calculating a Trees ML Score

Likelihood at site *j* =

Prob



$\mathrm{Prob}(\mathrm{C},\mathrm{C},\mathrm{A},\mathrm{G},G,C\,|\,\mathrm{T}) =$

$\mathrm{Prob}(\mathrm{C}) \quad \mathrm{Prob}(G\,|\,C,\mathrm{b}_1) \quad \mathrm{Prob}(\mathrm{C}\,|\,G,\mathrm{b}_2) \quad \mathrm{Prob}(\mathrm{C}\,|\,G,\mathrm{b}_3)$

$\mathrm{Prob}(\mathrm{A}\,|\,C,\mathrm{b}_4) \quad \mathrm{Prob}(\mathrm{G}\,|\,C,\mathrm{b}_5)$

Calculating a Trees ML Score

$\mathrm{Prob}(\mathrm{C},\mathrm{C},\mathrm{A},\mathrm{G},G,C|\mathrm{T}) =$

$\mathrm{Prob}(\mathrm{C})$ $\mathrm{Prob}(G|C,\mathrm{b}_1)$ $\mathrm{Prob}(\mathrm{C}|G,\mathrm{b}_2)$ $\mathrm{Prob}(\mathrm{C}|G,\mathrm{b}_3)$
$\mathrm{Prob}(\mathrm{A}|C,\mathrm{b}_4)$ $\mathrm{Prob}(\mathrm{G}|C,\mathrm{b}_5)$

Probabilities under Kimura 2 Parameter Model

1. Prob(transition|b) =

$$\frac{1}{4}-\frac{1}{2}\exp\left(-\frac{2R-1}{R+1}b\right)+\frac{1}{4}\exp\left(-\frac{2}{R+1}b\right)$$

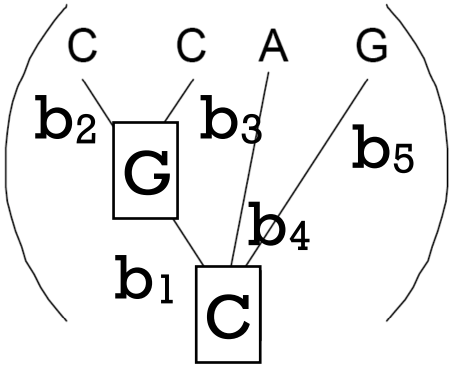
2. Prob(transversion|b) =

$$\frac{1}{2}-\frac{1}{2}\exp\left(-\frac{2}{R+1}b\right)$$

Calculating a Trees ML Score

Likelihood at site *j* =

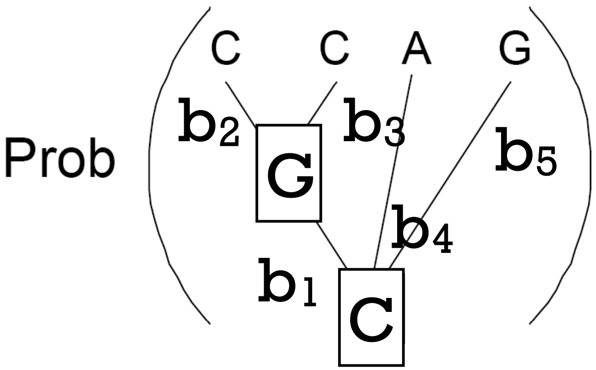
Prob



Prob(C,C,A,G,*G*,*C*|T) =
Prob(C) Prob(tranversion,b_1) Prob(tranversion,b_2) Prob(tranversion,b_3)
Prob(tranversion,b_4) Prob(tranversion,b_5)

Calculating a Trees ML Score

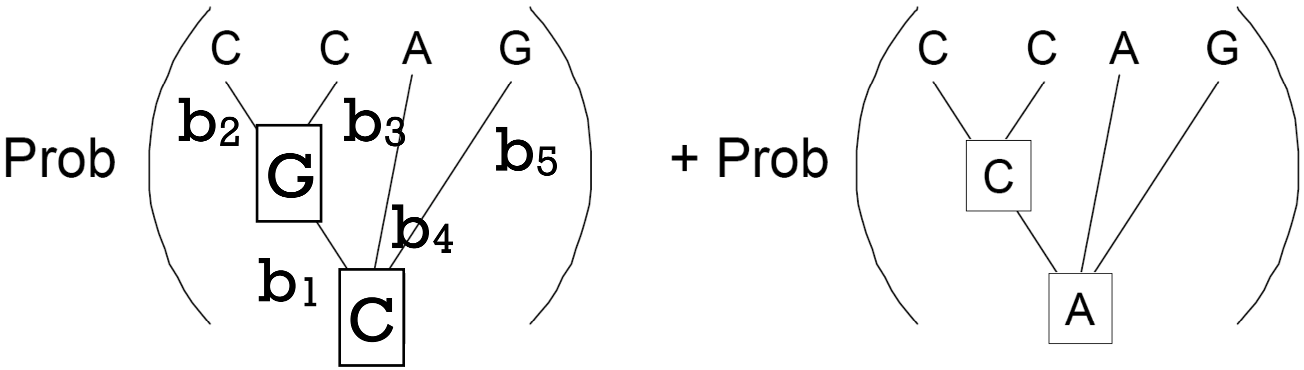
Likelihood at site *j* =



Prob(C,C,A,G,*G*,*C*|T) =
Prob(C) Prob(tranversion,b_1) Prob(tranversion,b_2) Prob(tranversion,b_3)
Prob(tranversion,b_4) Prob(tranversion,b_5)

Calculating a Trees ML Score

Likelihood at site *j* =



- n-1 nodes on tree with n species, each with one of 4 states
- 4^{n-1} terms
- For n=10 there are 262,144
- For n=20 there are 274,877,906,944

Calculating a Trees ML Score

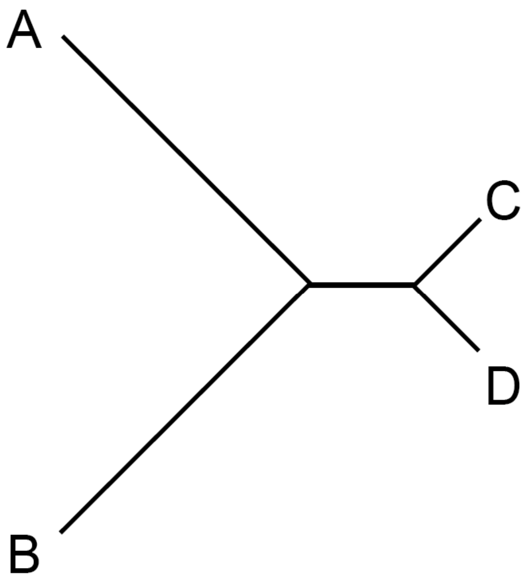
```
       1            j          N
(1)    C…GGACA…C…GTTTA…C
(2)    C…AGACA…C…CTCTA…C
(3)    C…GGATA…A…GTTAA…C
(4)    C…GGATA…G…CCTAG…C
```

$$L = \mathrm{Prob}(D|T) = \prod_{j=1}^{m} \mathrm{Prob}(D^{(j)}|T)$$

- Likelihood of tree equals product of likelihoods or that tree at each individual position.

Problems with Likelihood

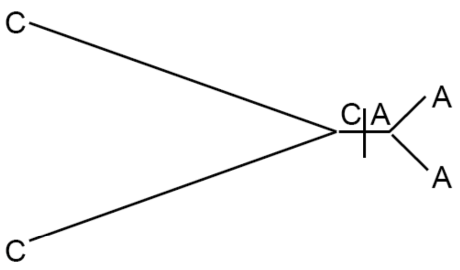
- Take a very long time.



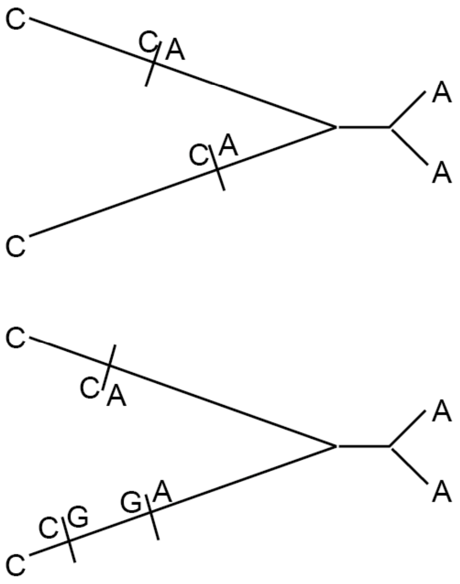
(Siddall, 1998)

- The "Farris Zone"

In the Farris Zone

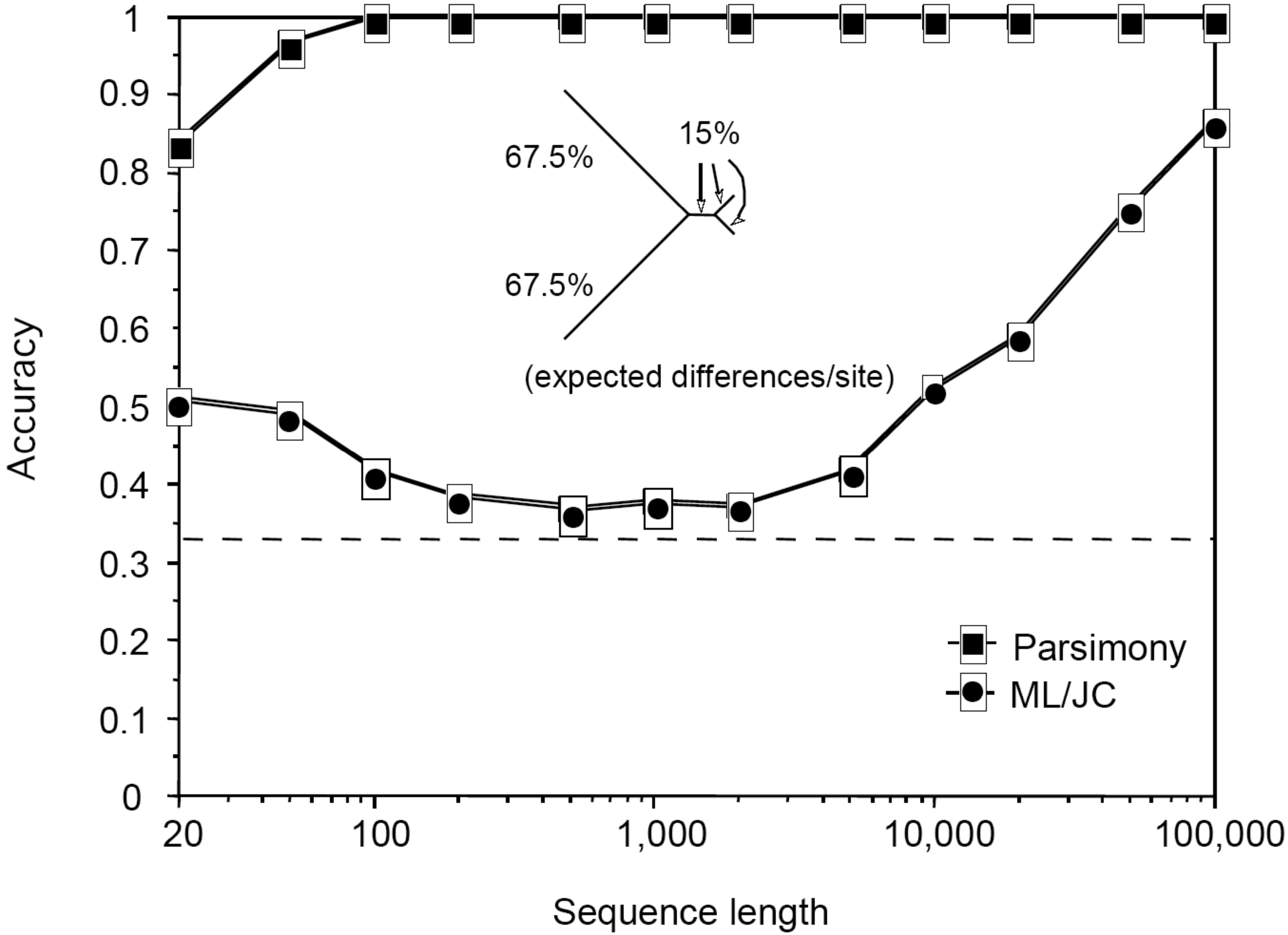


True synapomorphy

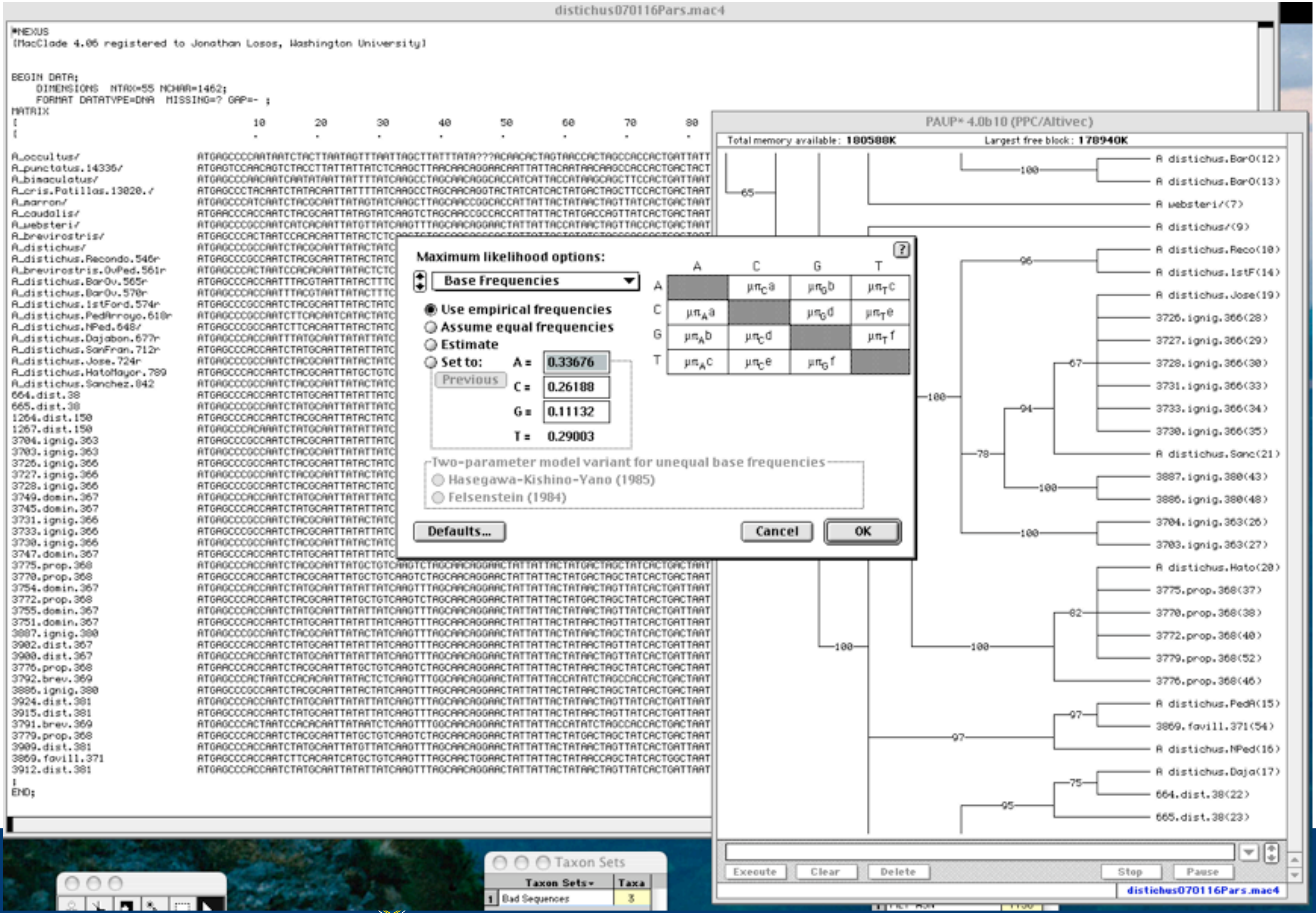


Apparent synapomorphies actually due to misinterpreted homoplasy

In the Farris Zone



Likelihood Analysis in PAUP



```
#NEXUS
[MacClade 4.06 registered to Jonathan Losos, Washington University]


BEGIN DATA;
     DIMENSIONS  NTAX=55 NCHAR=1462;
     FORMAT DATATYPE=DNA  MISSING=? GAP=- ;
MATRIX
[                                       10        20        30        40        50        60        70        80
[                                       .         .         .         .         .         .         .         .

A_occultus/
A_punctatus.14336/
A_bimaculatus/
A_cris.Patillas.13020./
A_marron/
A_caudalis/
A_websteri/
A_brevirostris/
A_distichus/
A_distichus.Recondo.546r
A_brevirostris.OvPed.561r
A_distichus.BarOv.565r
A_distichus.BarOv.570r
A_distichus.1stFord.574r
A_distichus.PedArroyo.610r
A_distichus.NPed.648/
A_distichus.Dajabon.677r
A_distichus.SanFran.712r
A_distichus.Jose.724r
A_distichus.HatoMayor.789
A_distichus.Sanchez.842
664.dist.38
665.dist.38
1264.dist.150
1267.dist.150
3704.ignig.363
3703.ignig.363
3726.ignig.366
3727.ignig.366
3728.ignig.366
3749.domin.367
3745.domin.367
3731.ignig.366
3733.ignig.366
3730.ignig.366
3747.domin.367
3775.prop.368
3770.prop.368
3754.domin.367
3772.prop.368
3755.domin.367
3751.domin.367
3887.ignig.380
3902.dist.367
3900.dist.367
3776.prop.368
3792.brev.369
3886.ignig.380
3924.dist.381
3915.dist.381
3791.brev.369
3779.prop.368
3909.dist.381
3869.favill.371
3912.dist.381
;
END;
```

Maximum likelihood options:

Base Frequencies

- ● Use empirical frequencies
- ○ Assume equal frequencies
- ○ Estimate
- ○ Set to:

A = 0.33676
C = 0.26188
G = 0.11132
T = 0.29003

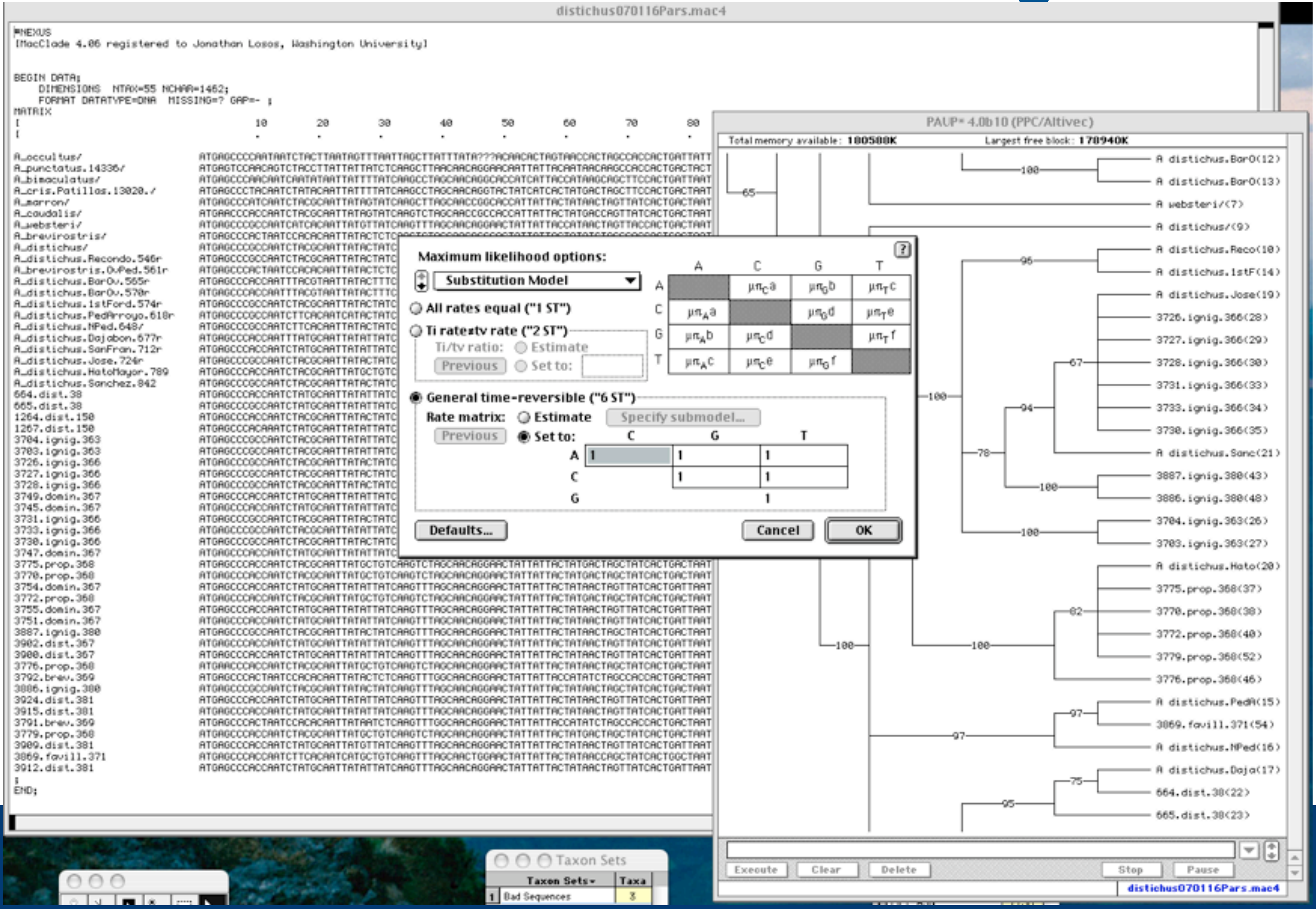
Two-parameter model variant for unequal base frequencies

- ○ Hasegawa-Kishino-Yano (1985)
- ○ Felsenstein (1984)

	A	C	G	T
A	▨	$\mu\pi_C a$	$\mu\pi_G b$	$\mu\pi_T c$
C	$\mu\pi_A a$	▨	$\mu\pi_G d$	$\mu\pi_T e$
G	$\mu\pi_A b$	$\mu\pi_C d$	▨	$\mu\pi_T f$
T	$\mu\pi_A c$	$\mu\pi_C e$	$\mu\pi_G f$	▨

Total memory available: **180588K**   Largest free block: **178940K**

Likelihood Analysis



distichus070116Pars.mac4

```
#NEXUS
[MacClade 4.06 registered to Jonathan Losos, Washington University]


BEGIN DATA;
    DIMENSIONS  NTAX=55 NCHAR=1462;
    FORMAT DATATYPE=DNA  MISSING=? GAP=- ;
MATRIX
[                                        10        20        30        40        50        60        70        80
[                                         .         .         .         .         .         .         .         .

A_occultus/                         ATGAGCCCCAATAATCTACTTAATAGTTTAATTAGCTTATTTATA???ACAACACTAGTAACCACTAGCCACCACTGATTATT
A_punctatus.14336/                  ATGAGTCCAACAGTCTACCTTATTATTATCTCAAGCTTAACAACAGGAACAATTATTACAATAACAAGCCACCACTGACTACT
A_bimaculatus/                      ATGAGCCCAACAATCAATATAATTATTTTATCAAGCCTAGCAACAGGCACCATCATTACCATAAGCAGCTTCCACTGATTAAT
A_cris.Patillas.13020./             ATGAGCCCTACAATCTATACAATTATTTTATCAAGCCTAGCAACAGGTACTATCATCACTATGACTAGCTTCCACTGACTAAT
A_marron/                           ATGAGCCCATCAATCTACGCAATTATAGTATCAAGCTTAGCAACCGGCACCATTATTACTATAACTAGTTATCACTGACTAAT
A_caudalis/                         ATGAACCCACCAATCTACGCAATTATAGTATCAAGTCTAGCAACCGCCACCATTATTACTATGACCAGTTATCACTGACTAAT
A_websteri/                         ATGAGCCCGCCAATCATCACAATTATGTTATCAAGTTTAGCAACAGGAACTATTATTACCATAACTAGTTACCACTGACTAAT
A_brevirostris/                     ATGAGCCCACTAATCCACACAATTATACTCTC
A_distichus/                        ATGAGCCCGCCAATCTACGCAATTATACTATC
A_distichus.Recondo.546r            ATGAGCCCGCCAATCTACGCAATTATACTATC
A_brevirostris.OvPed.561r           ATGAGCCCACTAATCCACACAATTATACTCTC
A_distichus.BarOv.565r              ATGAGCCCACCAATTTACGTAATTATACTTTC
A_distichus.BarOv.570r              ATGAGCCCACCAATTTACGTAATTATACTTTC
A_distichus.1stFord.574r            ATGAGCCCGCCAATCTACGCAATTATACTATC
A_distichus.PedArroyo.618r          ATGAGCCCGCCAATCTTCACAATCATACTATC
A_distichus.NPed.648/               ATGAGCCCGCCAATCTTCACAATTATACTATC
A_distichus.Dajabon.677r            ATGAGCCCACCAATTTATGCAATTATATTATC
A_distichus.SanFran.712r            ATGAGCCCACCAATCTATGCAATTATATTATC
A_distichus.Jose.724r               ATGAGCCCGCCAATCTACGCAATTATACTATC
A_distichus.HatoMayor.789           ATGAGCCCACCAATCTACGCAATTATGCTGTC
A_distichus.Sanchez.842             ATGAGCCCGCCAATCTACGCAATTATACTATC
664.dist.38                         ATGAGCCCACCAATCTATGCAATTATATTATC
665.dist.38                         ATGAGCCCGCCAATCTATGCAATTATATTATC
1264.dist.150                       ATGAGCCCACCAATCTACGCAATTATACTATC
1267.dist.150                       ATGAGCCCACAAATCTATGCAATTATATTATC
3704.ignig.363                      ATGAGCCCGCCAATCTACGCAATTATATTATC
3703.ignig.363                      ATGAGCCCGCCAATCTACGCAATTATATTATC
3726.ignig.366                      ATGAGCCCGCCAATCTACGCAATTATACTATC
3727.ignig.366                      ATGAGCCCGCCAATCTACGCAATTATACTATC
3728.ignig.366                      ATGAGCCCGCCAATCTACGCAATTATACTATC
3749.domin.367                      ATGAGCCCACCAATCTATGCAATTATATTATC
3745.domin.367                      ATGAGCCCACCAATCTATGCAATTATATTATC
3731.ignig.366                      ATGAGCCCGCCAATCTACGCAATTATACTATC
3733.ignig.366                      ATGAGCCCGCCAATCTACGCAATTATATTATC
3730.ignig.366                      ATGAGCCCGCCAATCTACGCAATTATACTATC
3747.domin.367                      ATGAGCCCACCAATCTATGCAATTATATTATC
3775.prop.368                       ATGAGCCCACCAATCTACGCAATTATGCTGTCAAGTCTAGCAACAGGAACTATTATTACTATGACTAGCTATCACTGACTAAT
3770.prop.368                       ATGAGCCCACCAATCTACGCAATTATGCTGTCAAGTCTAGCAACAGGAACTATTATTACTATGACTAGCTATCACTGACTAAT
3754.domin.367                      ATGAGCCCACCAATCTATGCAATTATATTATCAAGTTTAGCAACAGGAACTATTATTACTATAACTAGTTATCACTGATTAAT
3772.prop.368                       ATGAGCCCACCAATCTACGCAATTATGCTGTCAAGTCTAGCAACAGGAACTATTATTACTATGACTAGCTATCACTGACTAAT
3755.domin.367                      ATGAGCCCACCAATCTATGCAATTATATTATCAAGTTTAGCAACAGGAACTATTATTACTATAACTAGTTATCACTGATTAAT
3751.domin.367                      ATGAGCCCACCAATCTATGCAATTATATTATCAAGTTTAGCAACAGGAACTATTATTACTATAACTAGTTATCACTGATTAAT
3887.ignig.380                      ATGAGCCCGCCAATCTACGCAATTATACTATCAAGTTTAGCAACAGGAACTATTATTACTATAACTAGCTATCACTGACTAAT
3902.dist.367                       ATGAGCCCACCAATCTATGCAATTATATTATCAAGTTTAGCAACAGGAACTATTATTACTATAACTAGTTATCACTGATTAAT
3900.dist.367                       ATGAGCCCACCAATCTATGCAATTATATTATCAAGTTTAGCAACAGGAACTATTATTACTATAACTAGTTATCACTGATTAAT
3776.prop.368                       ATGAACCCACCAATCTACGCAATTATGCTGTCAAGTCTAGCAACAGGAACTATTATTACTATAACTAGCTATCACTGACTAAT
3792.brev.369                       ATGAGCCCACTAATCCACACAATTATACTCTCAAGTTTGGCAACAGGAACTATTATTACCATATCTAGCCACCACTGACTAAT
3886.ignig.380                      ATGAGCCCGCCAATCTACGCAATTATACTATCAAGTTTAGCAACAGGAACTATTATTACTATAACTAGCTATCACTGACTAAT
3924.dist.381                       ATGAGCCCACCAATCTATGCAATTATATTATCAAGTTTAGCAACAGGAACTATTATTACTATAACTAGTTATCACTGATTAAT
3915.dist.381                       ATGAGCCCACCAATCTATGCAATTATATTATCAAGTTTAGCAACAGGAACTATTATTACTATAACTAGTTATCACTGATTAAT
3791.brev.369                       ATGAGCCCACTAATCCACACAATTATAATCTCAAGTTTGGCAACAGGAACTATTATTACCATATCTAGCCACCACTGACTAAT
3779.prop.368                       ATGAGCCCACCAATCTACGCAATTATGCTGTCAAGTCTAGCAACAGGAACTATTATTACTATGACTAGCTATCACTGACTAAT
3909.dist.381                       ATGAGCCCACCAATCTATGCAATTATGTTATCAAGTTTAGCAACAGGAACTATTATTACTATAACTAGTTATCACTGATTAAT
3869.favill.371                     ATGAGCCCACCAATCTTCACAATCATGCTGTCAAGTTTAGCAACTGGAACTATTATTACTATAACCAGCTATCACTGGCTAAT
3912.dist.381                       ATGAGCCCACCAATCTATGCAATTATATTATCAAGTTTAGCAACAGGAACTATTATTACTATAACTAGTTATCACTGATTAAT
;
END;
```

Maximum likelihood options:

Substitution Model

○ All rates equal ("1 ST")

○ Ti rate≠tv rate ("2 ST")

Ti/tv ratio: ○ Estimate

Previous ○ Set to:

● General time-reversible ("6 ST")

Rate matrix: ○ Estimate Specify submodel...

Previous ● Set to:

	A	C	G	T
A	▨	$\mu\pi_C a$	$\mu\pi_G b$	$\mu\pi_T c$
C	$\mu\pi_A a$	▨	$\mu\pi_G d$	$\mu\pi_T e$
G	$\mu\pi_A b$	$\mu\pi_C d$	▨	$\mu\pi_T f$
T	$\mu\pi_A c$	$\mu\pi_C e$	$\mu\pi_G f$	▨

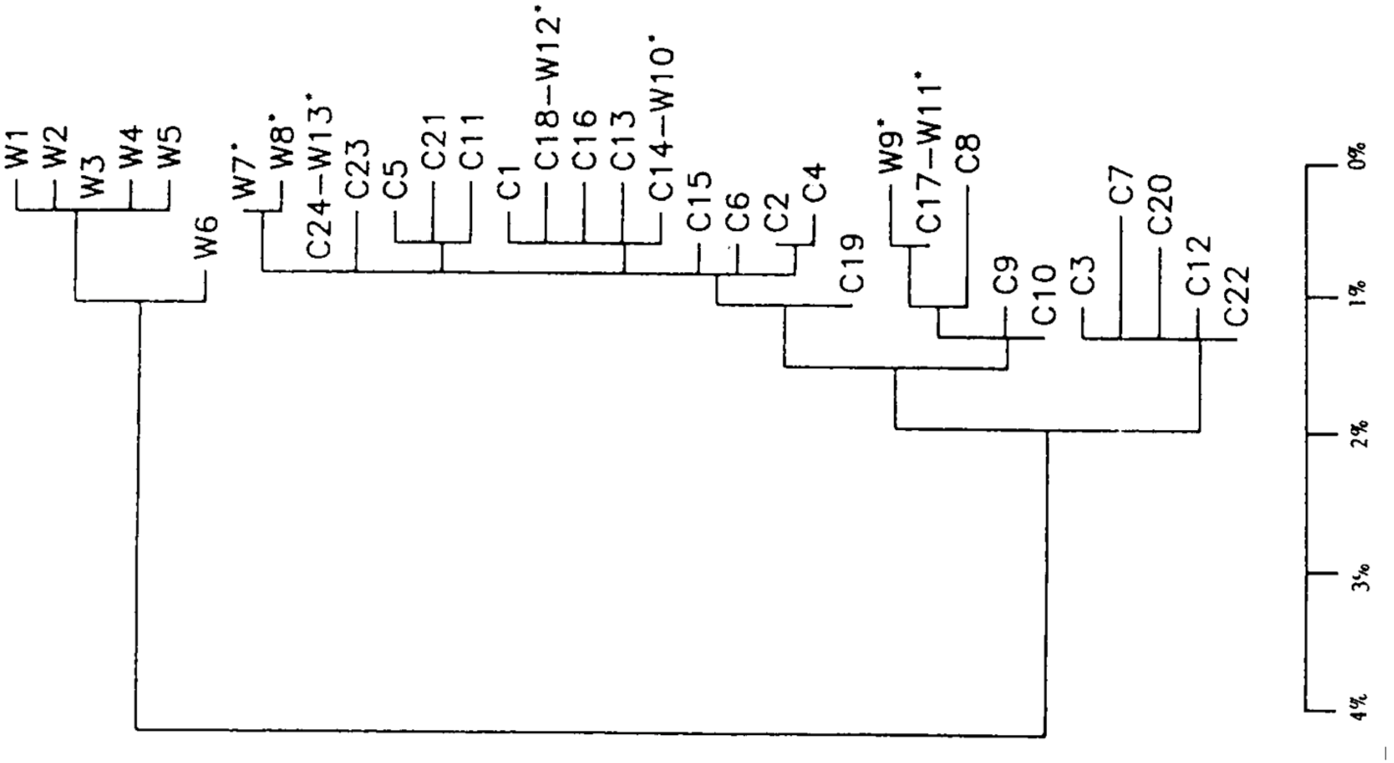
	C	G	T
A	1	1	1
C		1	1
G			1

Defaults... Cancel OK

PAUP* 4.0b10 (PPC/Altivec)

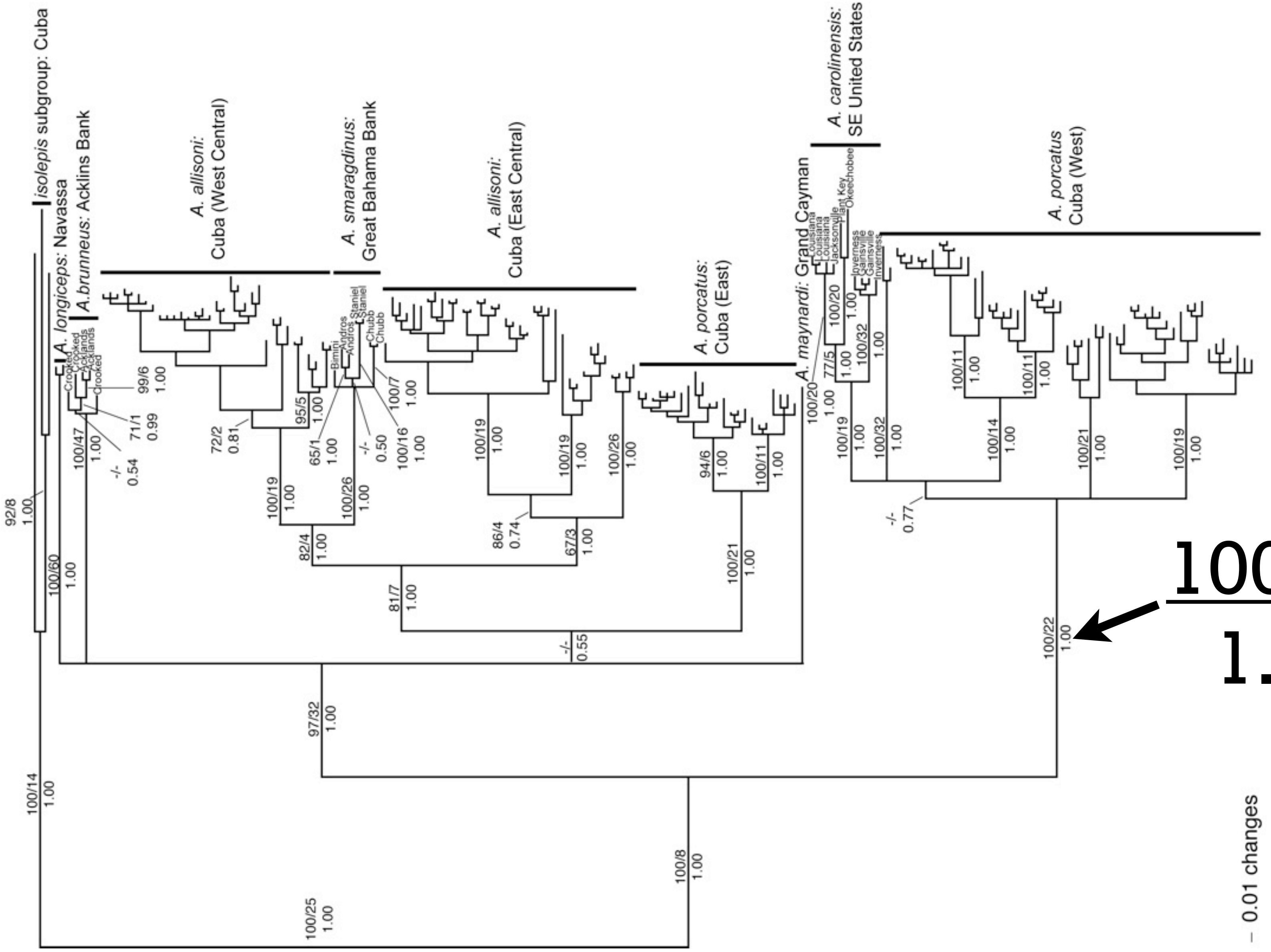
Total memory available: **180508K** Largest free block: **178940K**

Tree Support



Tree without support = crap

Tree Support

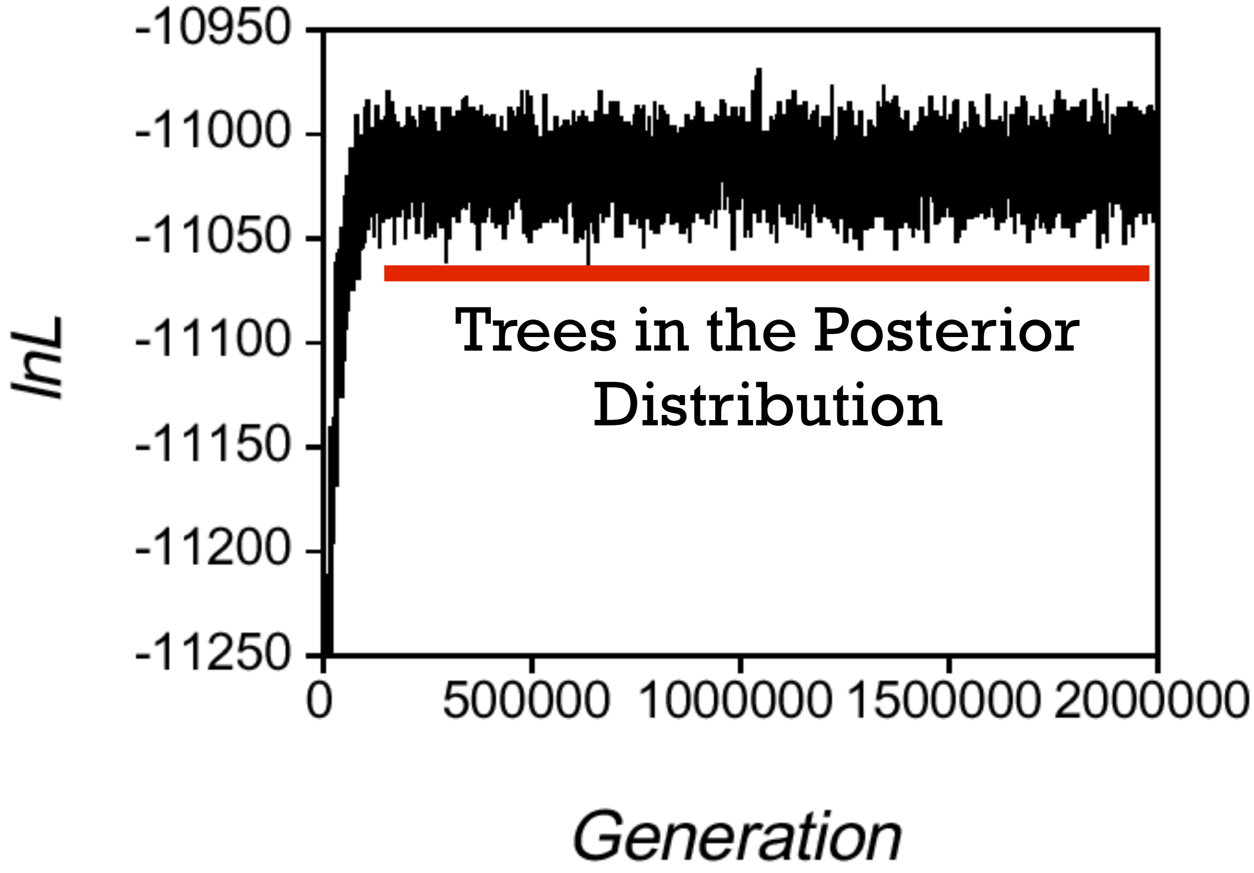


$$\frac{100/22}{1.00}$$

Measures of Tree Support

- Heuristic Measures of Support
  - Majority rule consensus trees
  - Posterior Probability (Bayesian only)
  - The Bootstrap
  - Decay Index
- Hypothesis Testing
  - Paired-sites tests (e.g., Templeton Test, KH Test)
  - Parametric bootstrap

Bayesian Posterior Probability



Consensus Trees

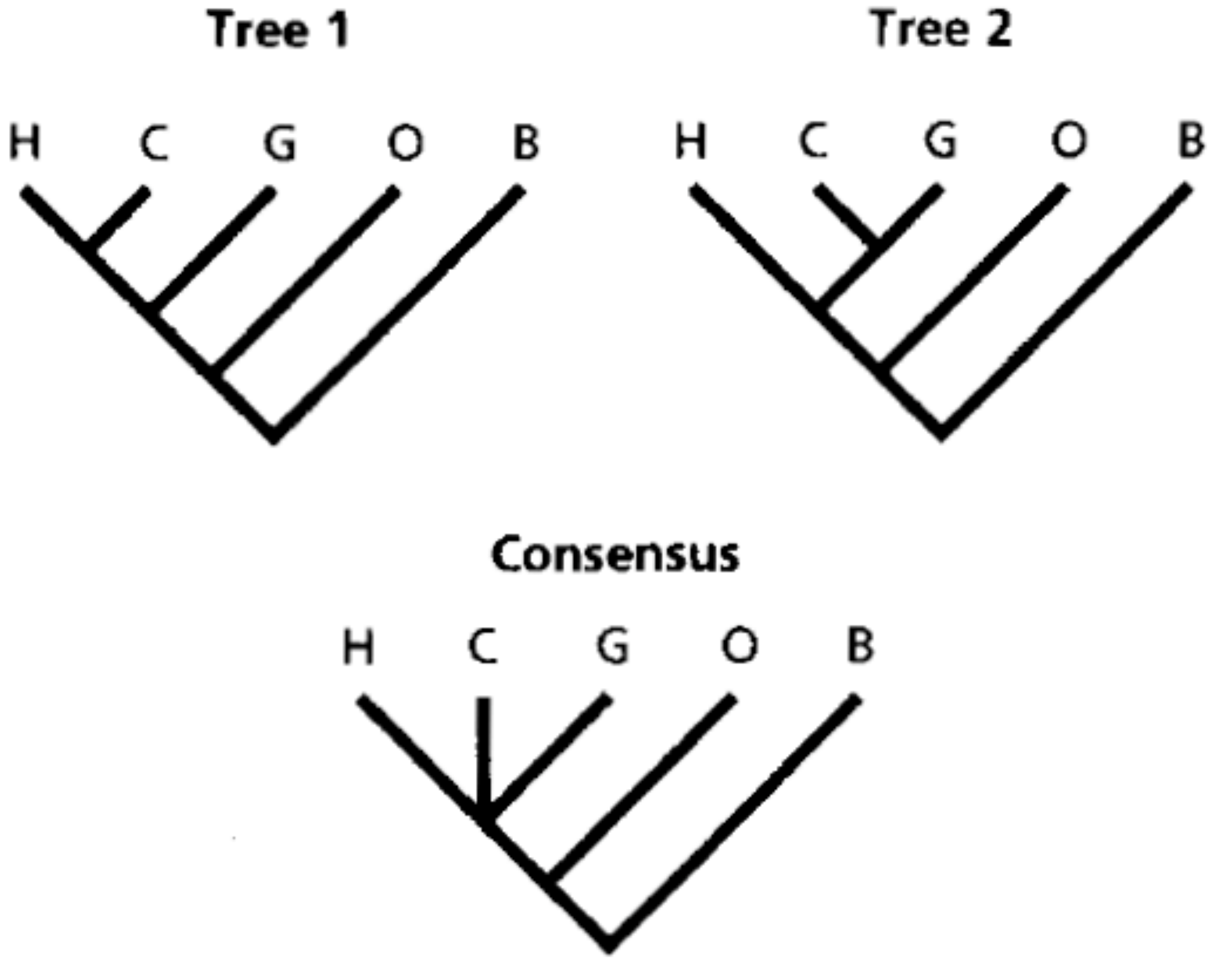
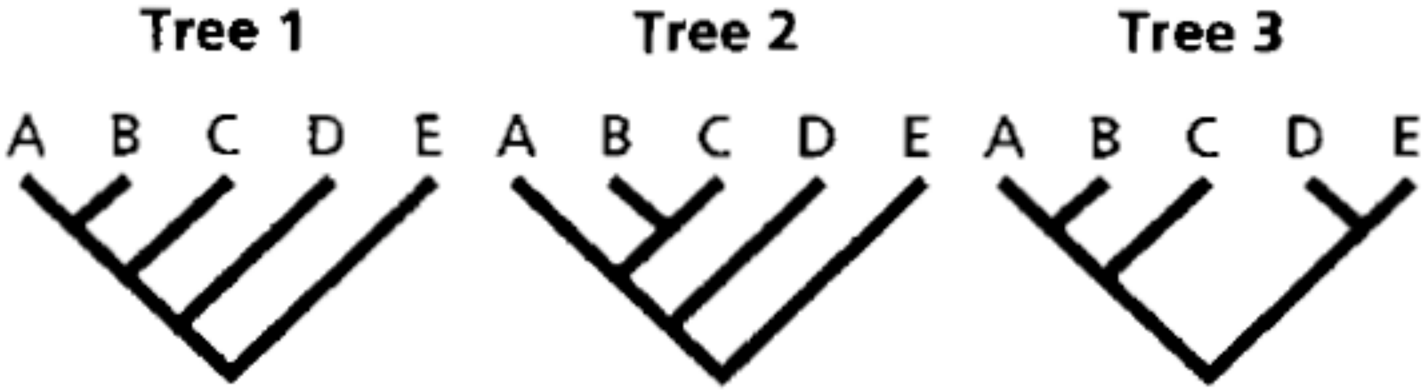
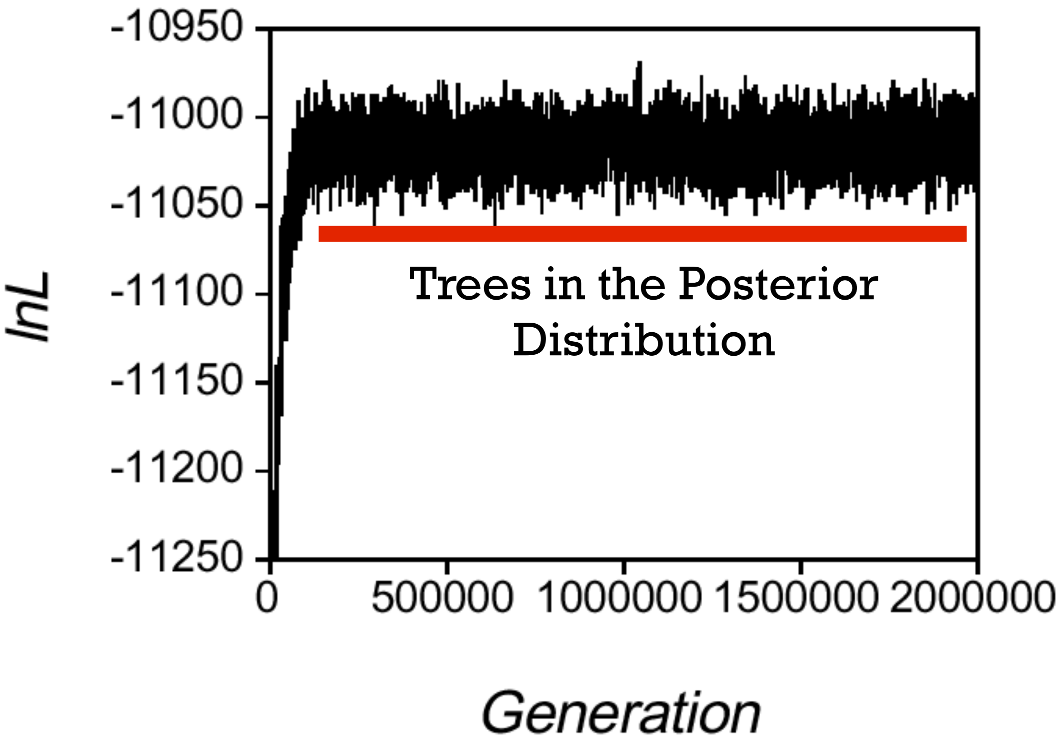


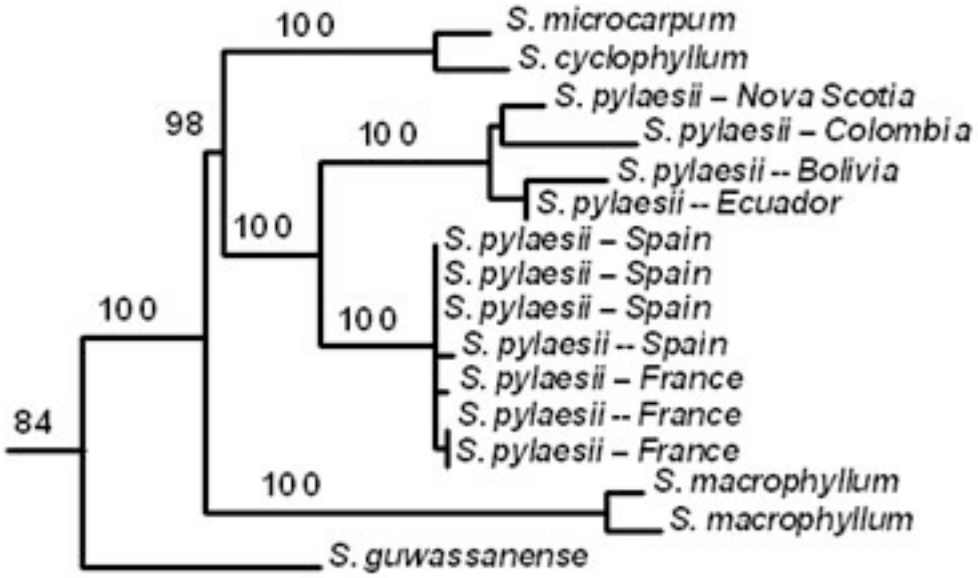
Figure 2.26 from Page and Holmes 1999

Majority Rule Consensus

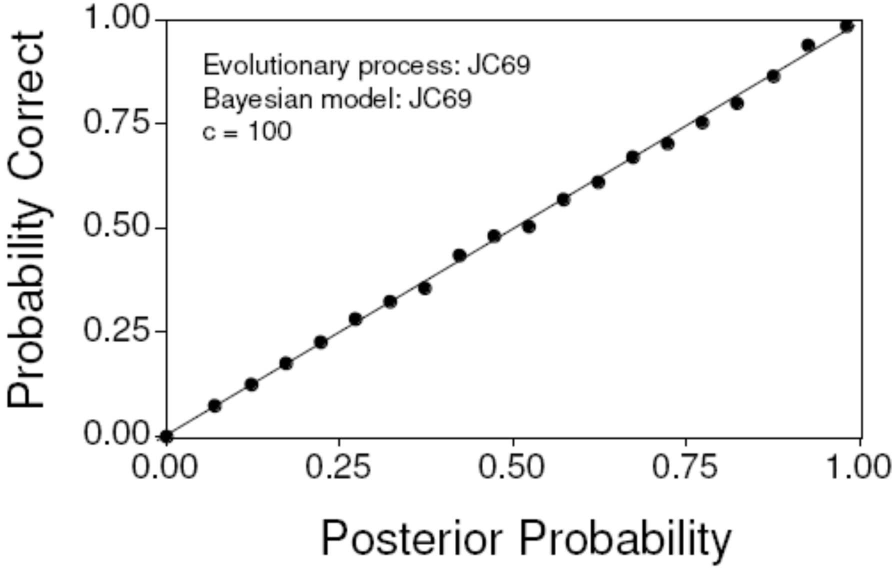


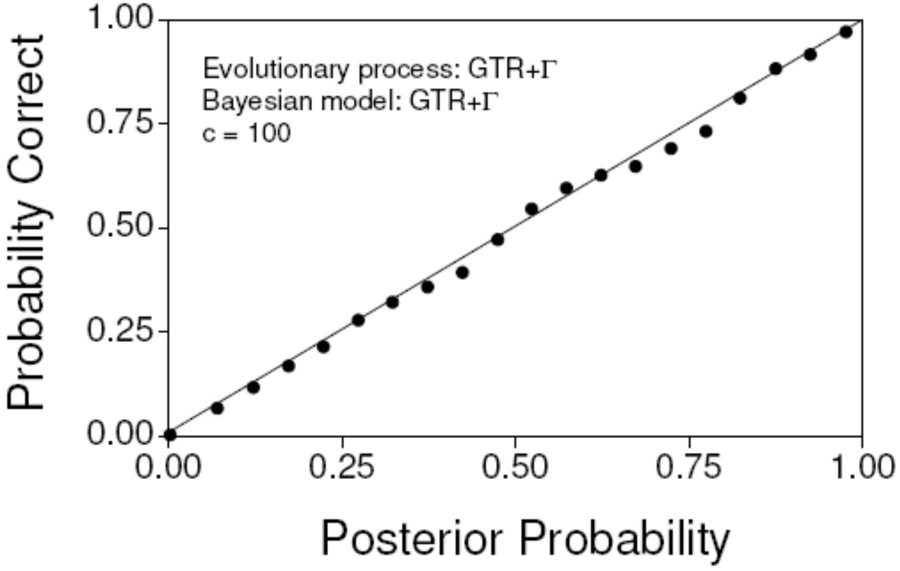
Bayesian Posterior Probabilties

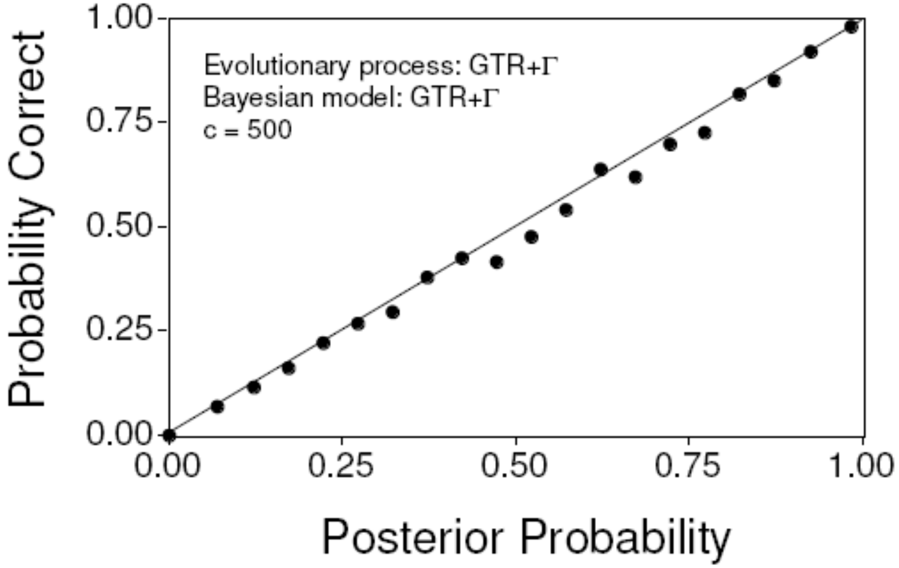


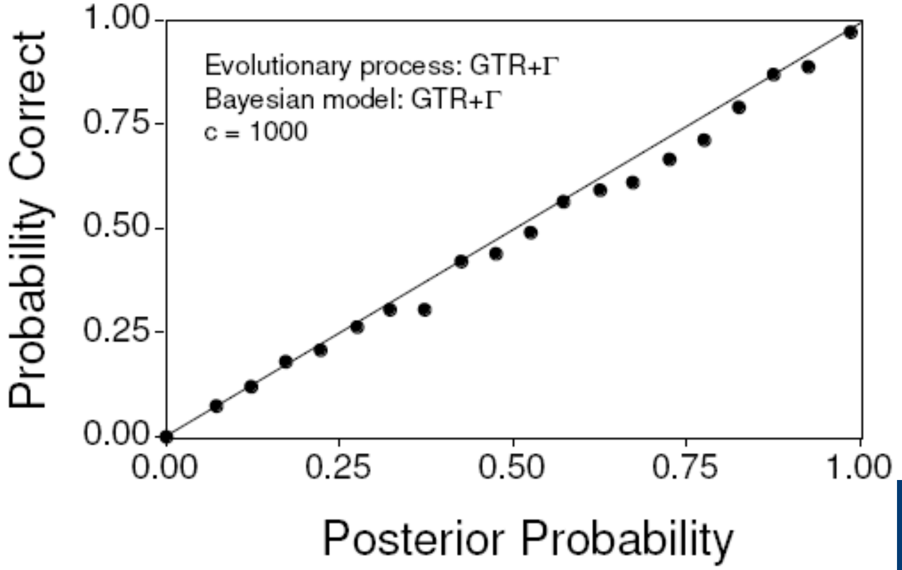


Posterior Probabilities



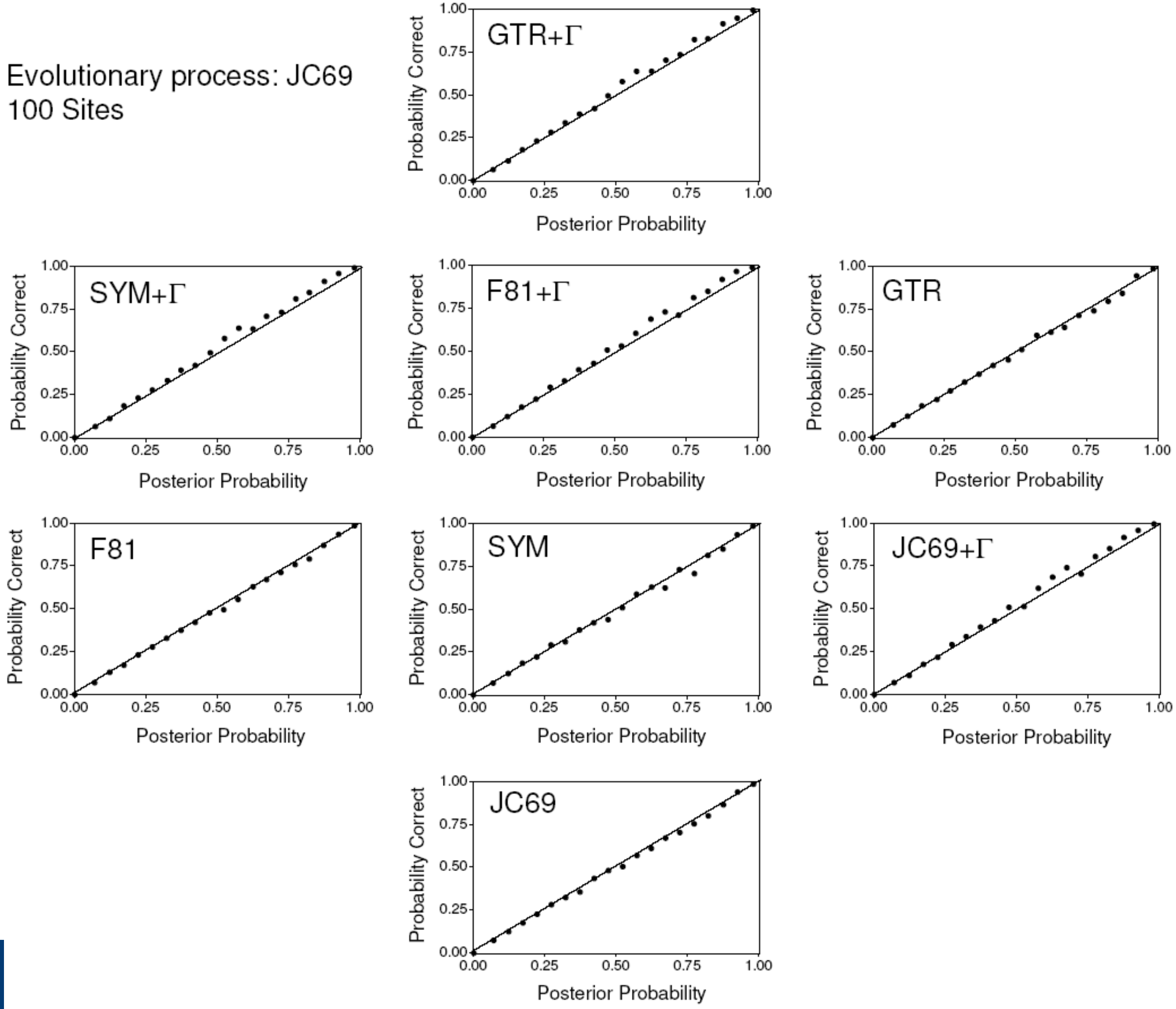






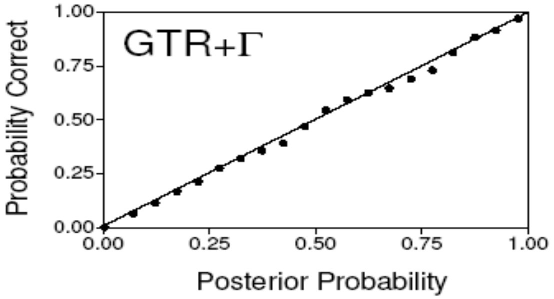
Posterior Probabilities

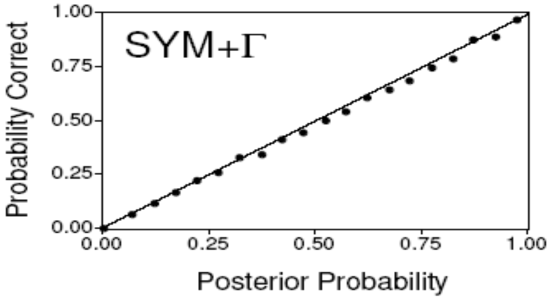
Evolutionary process: JC69
100 Sites

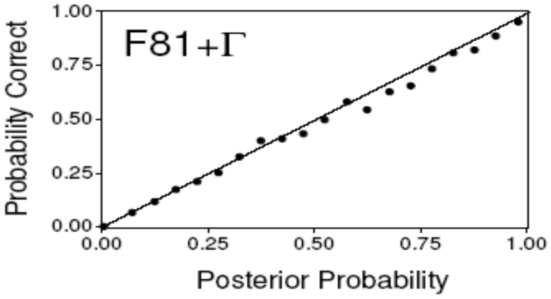


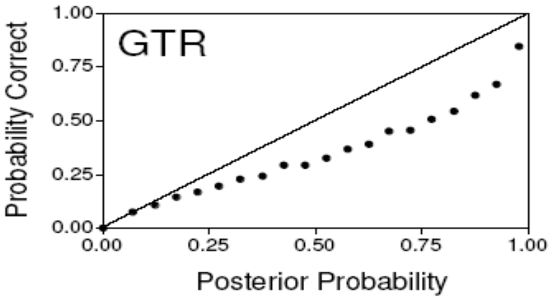
Posterior Probabilities

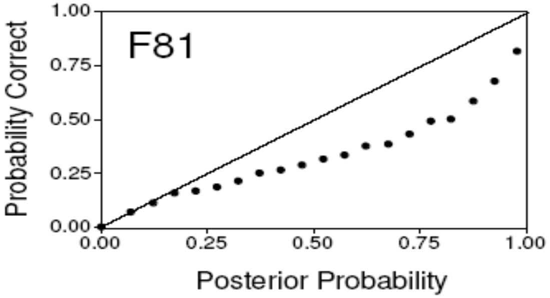
Evolutionary process: GTR+Γ
100 Sites

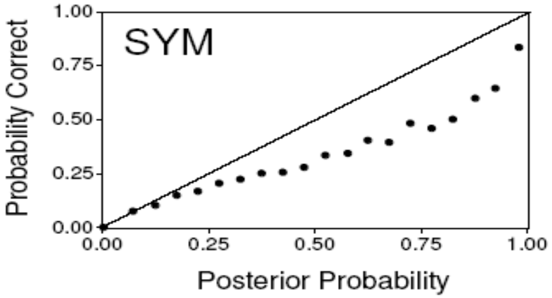


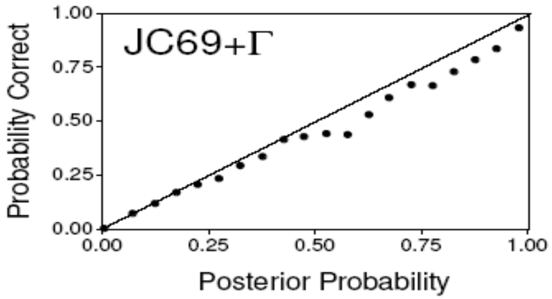


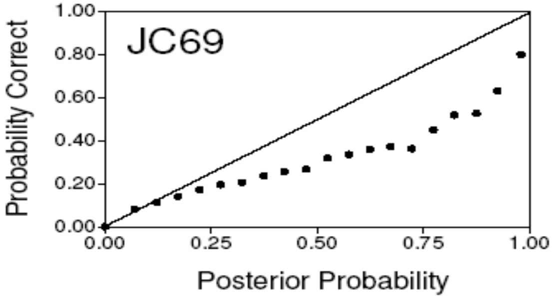






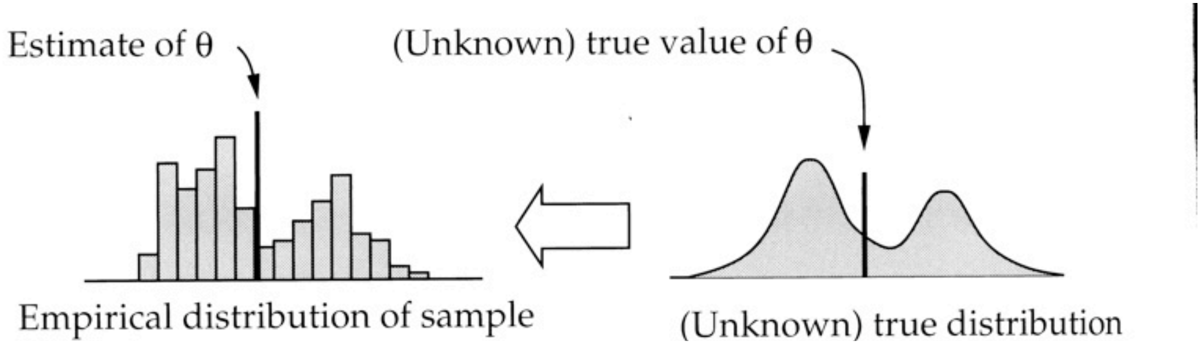






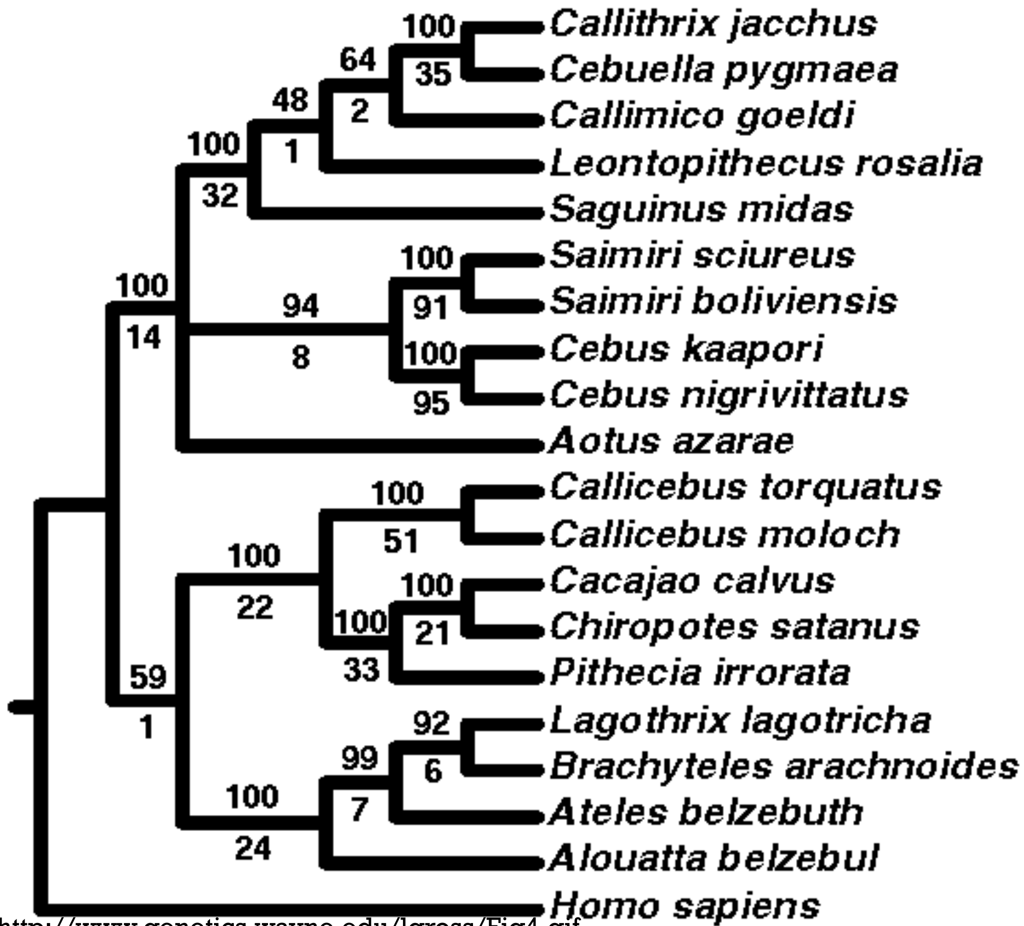
Bootstrapping

- Sampling with replacement



Bootstrapping a Phylogenetic Dataset

3. Ask which nodes are shared among analyses of bootstrapped datasets and construct consensus.

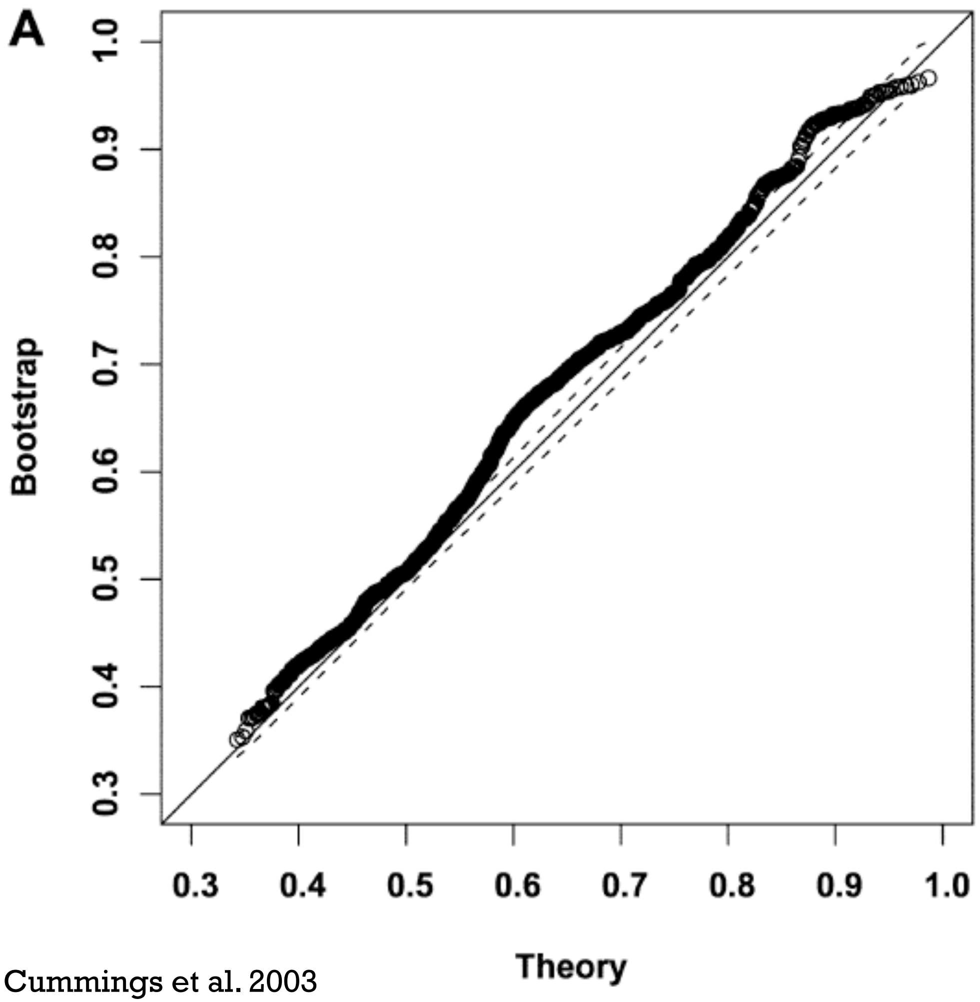


http://www.genetics.wayne.edu/lgross/Fig4.gif

Interpreting Bootstrap Values

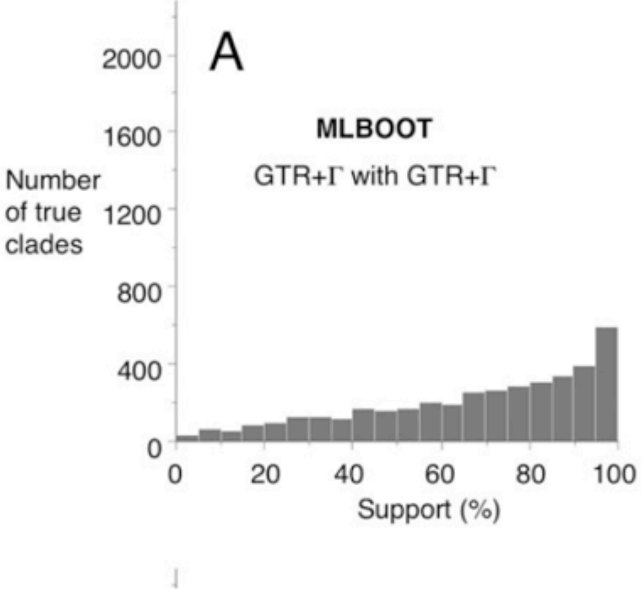
- Value on each node = percent of bootstrapped datasets that support that node.
- What is a “significant” bootstrap?
  - Most studies indicate that bootstrap values are conservative.
  - Depending on who you talk to, values ranging from 70-100% are considered reliable (Hillis and Bull 1993)

Bayesian Posterior Probabilities Versus Bootstraps



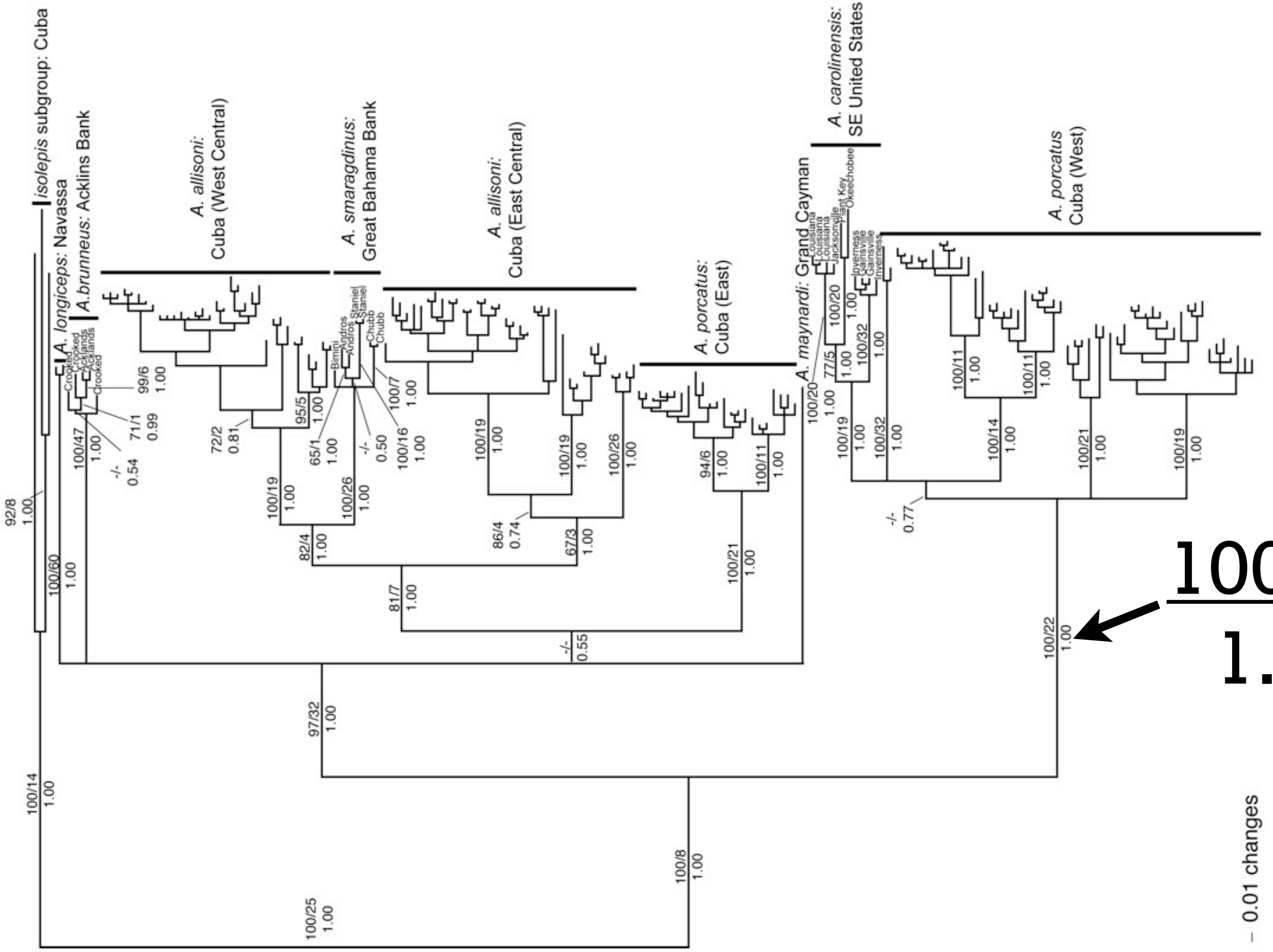
Cummings et al. 2003

Bayesian Posterior Probabilities Versus Bootstraps



Erixon et al. 2003

Multiple Measures



100/22

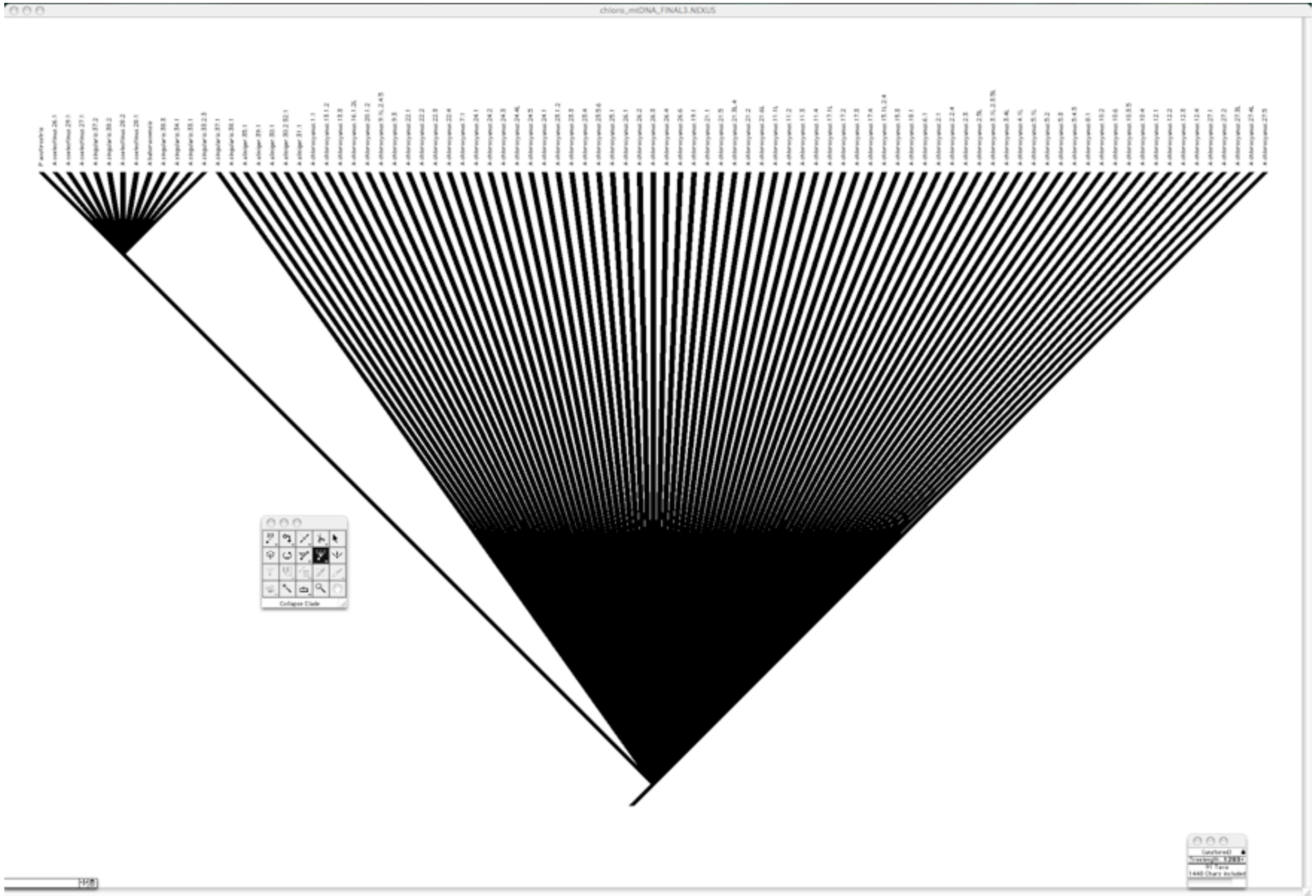
1.00

The Decay Index

- Also referred to as ‘Bremer support.’
- How much longer would a tree have to be to *not* include a particular node?

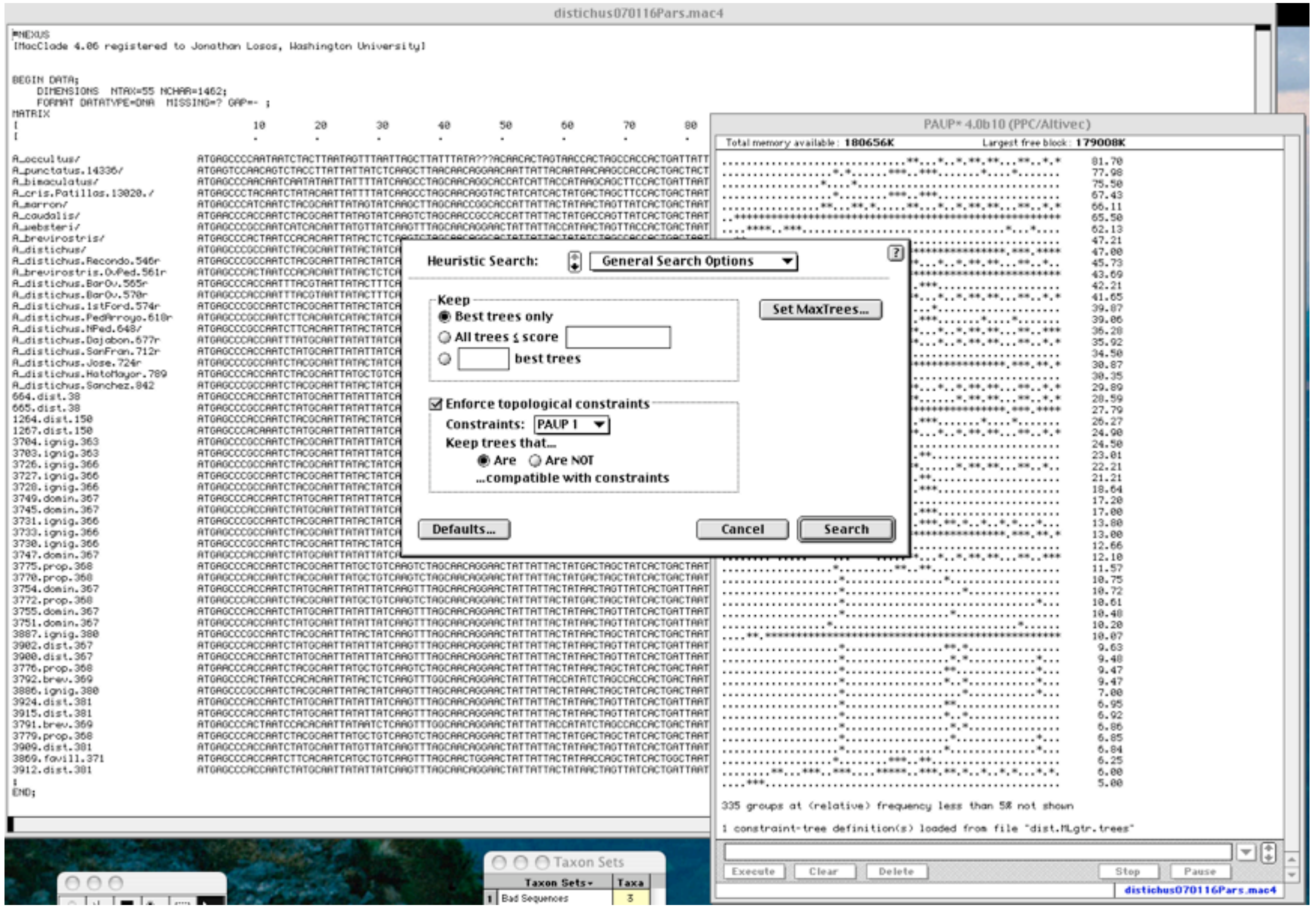
Calculating Decay Indices

1. Generate tree whose only structure consists of the node of interest.



Method for Calculating Decay Indices

2. Obtain the best tree lacking a particular node.



```
#NEXUS
[MacClade 4.06 registered to Jonathan Losos, Washington University]

BEGIN DATA;
    DIMENSIONS  NTAX=55 NCHAR=1462;
    FORMAT DATATYPE=DNA  MISSING=? GAP=- ;
MATRIX
```

PAUP* 4.0b10 (PPC/Altivec)

Total memory available: **180656K**   Largest free block: **179008K**

Heuristic Search: General Search Options

Keep
- ● Best trees only
- ○ All trees ≤ score
- ○ best trees

☑ Enforce topological constraints

Constraints: PAUP 1

Keep trees that...

● Are ○ Are NOT

...compatible with constraints

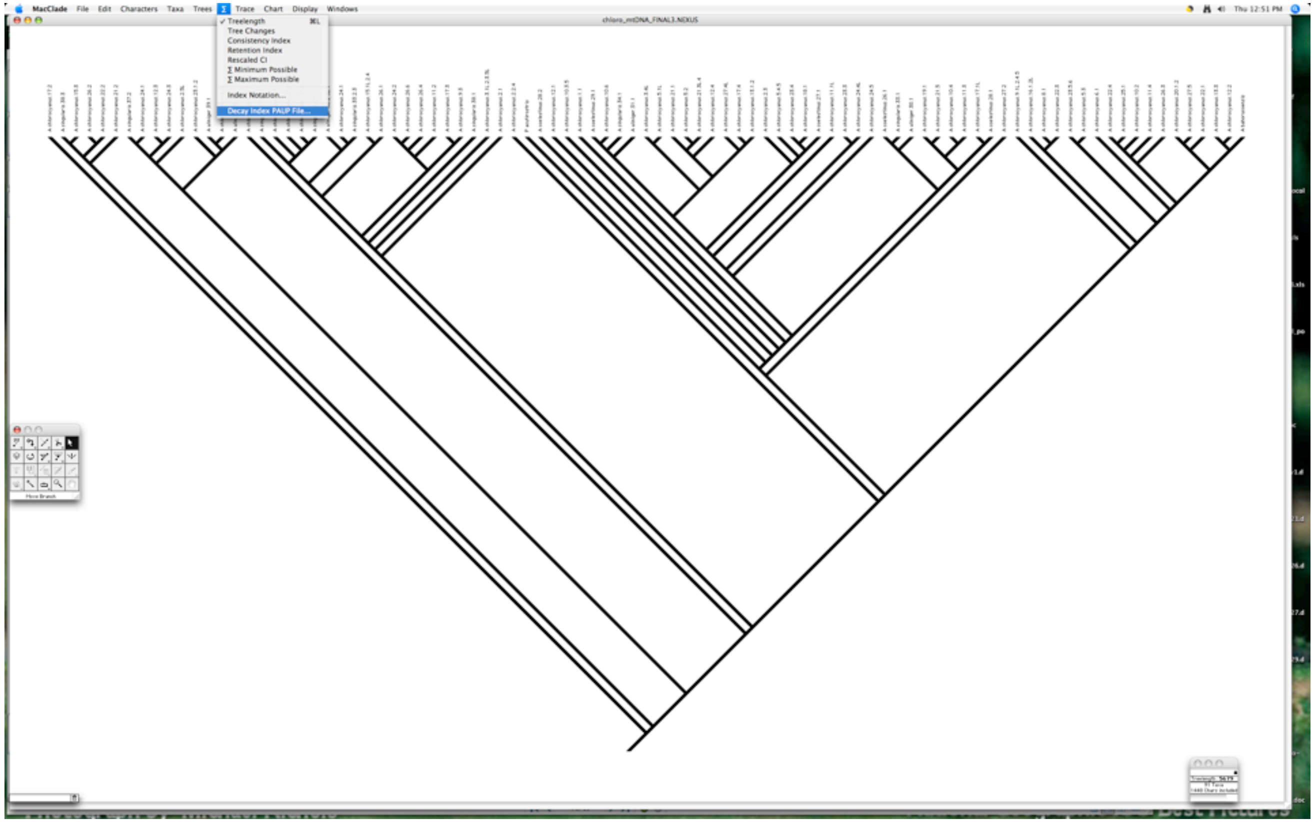
```
335 groups at (relative) frequency less than 5% not shown

1 constraint-tree definition(s) loaded from file "dist.MLgtr.trees"
```

Calculating Decay Indices

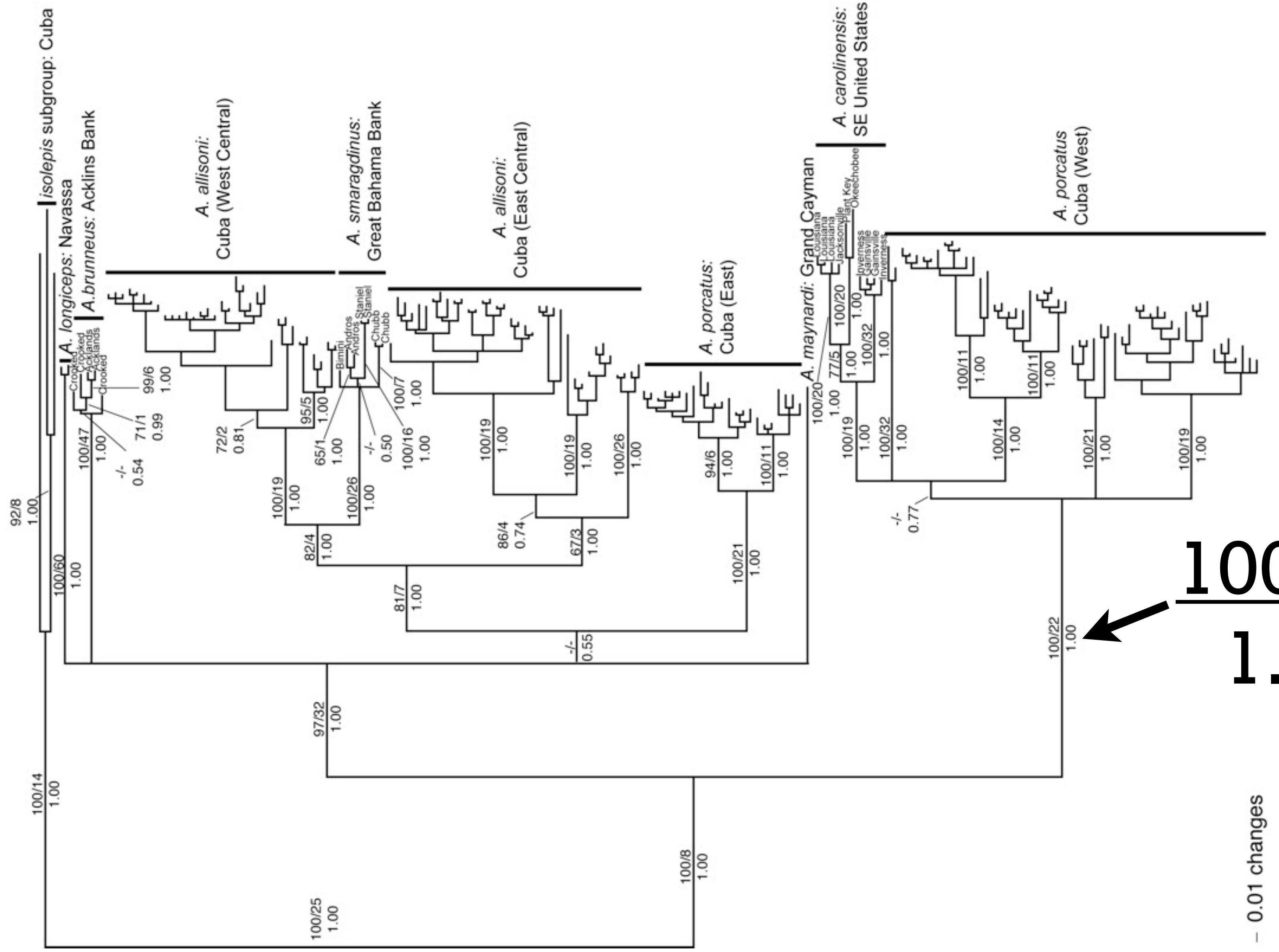
1. Generate tree whose only structure consists of a particular node of interest.
2. Use constraint function in PAUP to obtain the best tree lacking a particular node.
3. Calculate decay index
    - *Decay index* = score for most parsimonious tree without constraint minus score for most parsimonious tree lacking the node of interest.

Automation of Decay Index Calculation in MacClade

- Often calculated for every node on a tree.
- Use of MacClade to automate generation of constraint trees.



Well-supported Tree



100/22

1.00