Inferring phylogeny

Introduction • Distance methods • Parsimony method

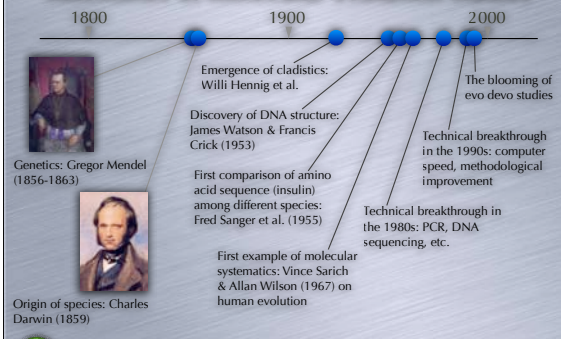
Jer-Ming Hu
胡哲明

Inst. Ecology & Evolutionary Biology
台大生態學與演化生物學研究所

Today's topics

- Overview of phylogenetic inferences
- Methodology
  - Methods using distance data
    - UPGMA, Neighbor-joining, minimum evolution
  - Methods using discrete data
    - Maximum parsimony
    - Maximum likelihood
    - Bayesian inference

Milestones of molecular evolution studies



Nature (1968) 217: 624

Contributions to molecular evolution

The patterns of evolution
Cladistics and phylogenetics:
Willi Hennig (1913-1976)

The mechanisms
The neutral theory of molecular evolution:
Motoo Kimura (木村資生, 1924-1994)

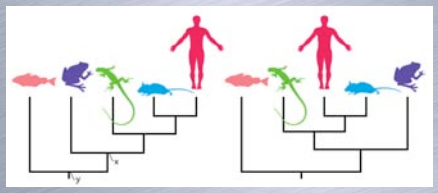
The models and methodological concerns
The uses of parsimony, maximum likelihood methods in phylogenetics:
Joseph Felsenstein
Genome Science and of Biology, University of Washington, Seattle

Reconstructing phylogeny

- The concepts of trees
- The data used to construct phylogenetic trees
  - Morphological data
  - Molecular data
- Homology of characters in data matrices
  - Sequence alignment

The basics of evolutionary trees

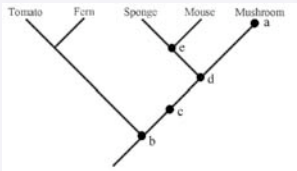
- The tree shapes



Are these two trees different?

Baum et al. (2005) *Science* 310:979

Most recent common ancestor (MRCA)



Which is the MRCA of a mushroom and a sponge? d

Which is the MRCA of a mouse and a fern? b

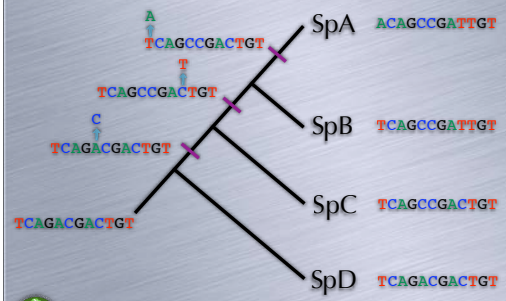
Types of data

- Character and character types
  - Quantitative and qualitative characters
  - Binary and multistate characters
- Assumptions about character evolution
  - Substitution models
  - Step matrix

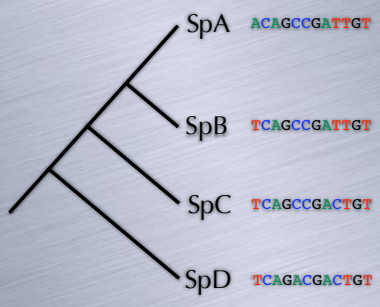
Distance method

- **UPGMA** (Unweighted Pair Group Method using Arithmetic averages)
- **Transformed distance method**
- **Neighbor-Joining method**
- **Minimum evolution**

Evolution of DNA sequences



In distance method, sequences were grouped based on their similarity, we need to measure the differences among sequences



Measuring genetic distance

- How different are the two sequences we have?
  - The ways of measuring nucleotide substitution
  - Simplest way
    - Hamming distance = n/N x100%

SpA ACAGCCGA-TGT
SpB TCAGCCGA-TGT
SpC TCAGCCGACTGT
SpD TCAGACGACTGT

1/12 = 8.3%

2/12 = 16.6%

Measuring differences between sequences

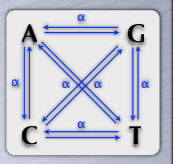
- The common ways
  - Direct counting
  - The probability of substitution at certain nucleotide site change from *i* to *j*.
- Nucleotide substitution models
  - **Jukes & Cantor**(1969)'s one parameter model
  - **Kimura**(1980)'s two parameter model
  - **F81** (Felsenstein, 1981)
  - **HKY85** (Hasegawa et al., 1985)
  - **Generalized time reversible model** (GTR)

The probability of nucleotide substitution

SpC TCAGCCGACTGT
SpD TCAGACGACTGT

- In order to know the probability of the sequences to be different (***p***), we can calculate the probability of the sequences being the same ($I_{(t)}$) first.
- Let's start with Jukes-Cantor model

We can just use another estimator, ***K***, the number of substitutions per site since time of divergence between the two sequences



Therefore, $K = 2 \times 3\alpha t = 6\alpha t$

$\because \quad 8\alpha t = -\ln(1-\frac{4}{3}p) \qquad \therefore \quad K = -\frac{3}{4}\ln(1-\frac{4}{3}p)$

where p = the observed proportions of different nucleotides between the two sequences

Example of calculation

If there are 80 transitions and 20 transversions between two sequences (length=1000bp), the number of substitutions per site ***K*** can be estimated:

Under **J-C model**,

$$K = -\frac{3}{4}\ln(1-\frac{4}{3}p) = 0.10732 \qquad (p=0.1)$$

Under **Kimura 2P model**:

$$K = \frac{1}{2}\ln(\frac{1}{1-2P-Q}) + \frac{1}{4}\ln(\frac{1}{1-2Q}) \qquad (P=0.08,\ Q=0.02)$$

$$= 0.10943$$

where P = the proportions of transitions and Q = the proportions of transversions. p is the proportion of total substitution.

Overview of Distance method

1. Transform the data matrix into distance table

	sequences (sites)						
	1	2	3	4	5	6	7
1	T	T	A	T	T	A	A
2	A	A	T	T	T	A	A
3	A	A	A	A	A	T	A
4	A	A	A	A	A	A	T

	distances		
2	3		
3	5	4	
4	5	4	2
	1	2	3

sequences

2. Use this new data matrix to evaluate/ construct the tree

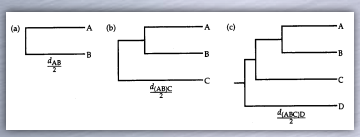
(Page & Holme 1998)

UPGMA

OTU	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

OTU	(AB)	C
C	$d_{(AB)C}$	
D	$d_{(AB)D}$	d_{CD}

where $d_{(AB)C} = (d_{AC} + d_{BC})/2$ and $d_{(AB)D} = (d_{AD} + d_{BD})/2$



(Graur & Li 2000)

Example

	dog	bear	raccoon	weasel	* seal	* sea lion	cat	monkey
dog	0	32	48	51	**50**	**48**	98	148
bear	32	0	26	34	**29**	**33**	84	136
raccoon	48	26	0	42	**44**	**44**	92	152
weasel	51	34	42	0	**44**	**38**	86	142
* seal	**50**	**29**	**44**	**44**	**0**	**24**	**89**	**142**
* sea lion	**48**	**33**	**44**	**38**	**24**	**0**	**90**	**142**
cat	98	84	92	86	**89**	**90**	0	148
monkey	148	136	152	142	**142**	**142**	148	0

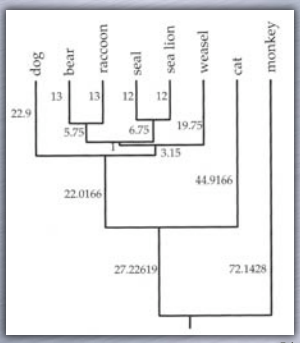
Example

	dog	bear	raccoon	weasel	SS	cat	monkey
dog	0	32	48	51	49	98	148
bear	32	0	26	34	31	84	136
raccoon	48	26	0	42	44	92	152
weasel	51	34	42	0	41	86	142
SS	49	31	44	41	0	89.5	142
cat	98	84	92	86	89.5	0	148
monkey	148	136	152	142	142	148	0

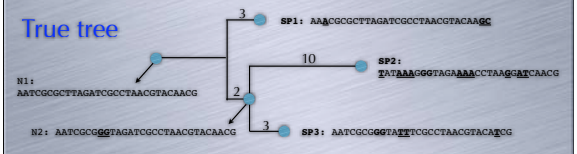
Example

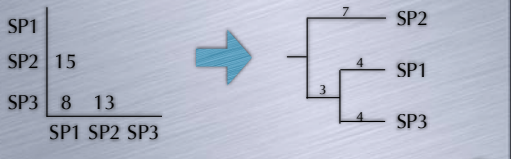
	dog	* BR	weasel	* SS	cat	monkey
dog	0	**40**	51	**49**	98	148
* BR	**40**	0	**38**	**37.5**	**88**	**144**
weasel	51	**38**	0	**41**	86	142
* SS	**49**	**37.5**	**41**	0	**89.5**	**142**
cat	98	**88**	86	**89.5**	0	148
monkey	148	**144**	142	**142**	148	0

UPGMA tree

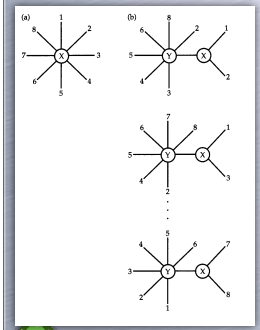


True tree



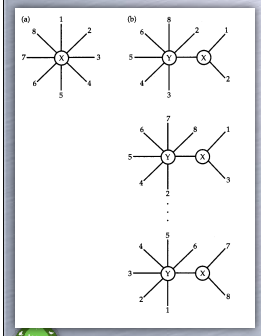


Neighbor-Joining (NJ) method



- Saitou & Nei (1987); Studier & Keppler (1988)
- NJ does not assume a clock and approximates the minimum evolution method

NJ procedure



1. For each tip, compute

$$u_i = \sum_{j:\, j \neq i}^{n} \frac{D_i}{(n-2)}$$

2. Choose the *i* and *j* for which $D_{ij}-u_i-u_j$ is smallest

3. Join *i* and *j*, compute the new node (*ij*) to every other nodes

$$D_{(ij),k} = \frac{(D_{ik} + D_{jk} - D_{ij})}{2}$$

4. Delete tips *i* and *j* and make new table with the new node (*ij*)

5. Continue till resolve the tree

Limitation on distance methods

- All nucleotide sites change independently.
- When evolutionary rates vary from site to site, than the data set needs to be corrected
- The substitution rate is constant over time and in different lineages.
- The base composition is at equilibrium.
- The conditional probabilities of nucleotide substitutions are the same for all sites and do not change over time.

Discrete methods

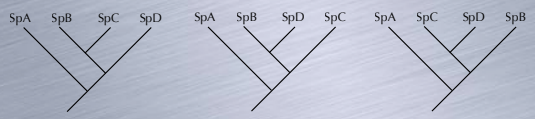
- Maximum parsimony (高度簡約原理)
- Maximum likelihood (最大似然性原理)
- Bayesian inference (貝葉氏導出式分析)
- Others

Parsimony methods

- The goal is to find the most parsimonious tree
- The criteria are to calculate the changes of character states, i.e. the *evolutionary steps*
- First, we have to know the way to evaluate a given tree

A simple data matrix w/ discrete characters

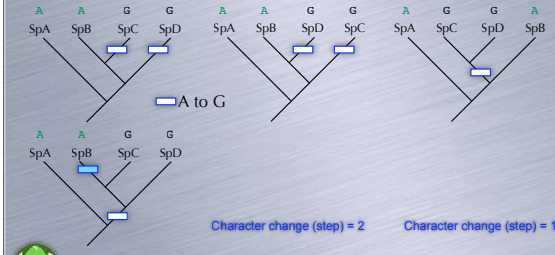
```
SpA: TCAGACGATTGTCAGACCATTG
SpB: TCAGTCGACTGTCAAACCATTG
SpC: TCGGTCAATTGTCAAACGATTG
SpD: TCGGTCAATTGTCAAACGATTG
```



A simple data matrix w/ discrete characters

For position no. 3

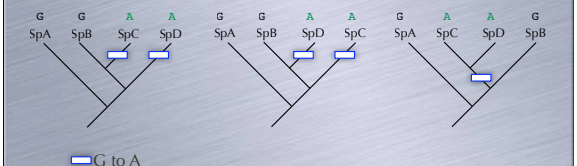
```
SpA: TCAGACGATTGTCAGACCATTG
SpB: TCAGTCGACTGTCAAACCATTG
SpC: TCGGTCAATTGTCAAACGATTG
SpD: TCGGTCAATTGTCAAACGATTG
```



A simple data matrix w/ discrete characters

For position no. 7

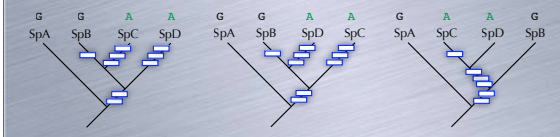
SpA: TCAGACGATTGTCAGACCATTG
SpB: TCAGTCGACTGTCAAACCATTG
SpC: TCGGTCAATTGTCAAACGATTG
SpD: TCGGTCAATTGTCAAACGATTG



Character change (step) = 2 | Character change (step) = 2 | Character change (step) = 1

A simple data matrix w/ discrete characters

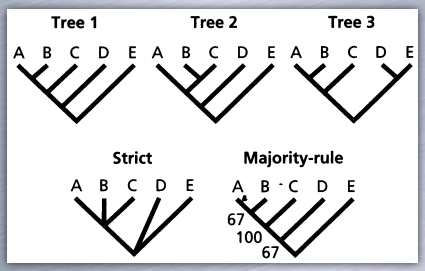
SpA: TCAGACGATTGTCAGACCATTG
SpB: TCAGTCGACTGTCAAACCATTG
SpC: TCGGTCAATTGTCAAACGATTG
SpD: TCGGTCAATTGTCAAACGATTG



Character change (step) = 9 | Character change (step) = 9 | Character change (step) = 6

Consensus trees

Finding the underlying information among trees

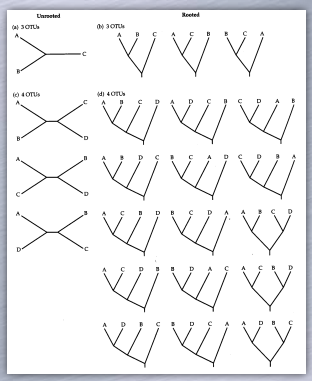


(Page & Holme 1998)

So, how many trees out there are we dealing with?

3 taxa

4 taxa



Tree searching

- The number of unrooted trees:

$$N_{unrooted} = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

- The number of rooted trees:

$$N_{rooted} = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

Taxon number	All possible unrooted tree number
3	1
4	3
5	15
6	105
7	945
8	10,395
9	135,135
10	2,027,025
11	34,459,425
12	654,729,075
13	13,749,310,575
14	316,234,143,225
15	7,905,853,580,625
16	213,458,046,676,875
17	6,190,283,353,629,375
18	191,898,783,962,510,625
19	6,332,659,870,762,850,625
20	221,643,095,476,699,771,875
21	8,200,794,532,637,891,559,375
22	319,830,986,772,877,770,815,625
23	13,113,070,457,687,988,603,440,625
24	563,862,029,680,583,509,947,946,875
25	25,373,791,335,626,257,947,657,609,375 **$>10^{29}$**
...	etc.

37

Note

- Say a computer can evaluate 10^6 trees per second.
- If we want to evaluate all of the trees for 25 taxa, we will need
  - $x = 10^{29}/10^6/60/60/24/365 = 3.17x10^{15}$年（三千兆年）

We got to have some ways to approximate the true tree.

38

Ways to solve time paradigm

- Skip unnecessary calculation
- Change the tree searching ways

39

The procedure can be simplified

```
SpA: TCAGACGATTGTCAGACCATTG
SpB: TCGGTCGACTGTCAGACCATTG
SpC: TCAGTCGATTGTCA-ACGATTG
SpD: TCAGTCGATTGTCA-ACGATTG
SpE: TCAGTCGATCGTCA-ACGATTG
```

Parsimonious uninformative

Parsimonious informative

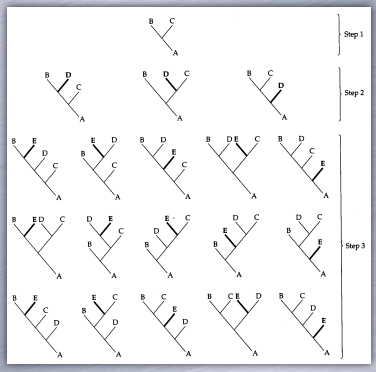
40

Searching for optimal trees

- Exhausted search
- Branch-and-bound method
- Heuristic approach
  - Stepwise/closest/random addition
  - Star decomposition
  - Branch swapping

41

Exhausted search



Page & Holme (1998) *Molecular evo...*

42

Distribution of tree length

```
Frequency distribution of lengths of 1000000 random trees:

            mean=6528.095481 sd=65.005662 g1=-0.322317 g2=0.159848
6114.00000 /-------------------------------------------------------------------
6146.30000 | (1)
6178.60000 | (3)
6210.90000 | (24)
6243.20000 | (84)
6275.50000 | (302)
6307.80000 | (1071)
6340.10000 |# (3493)
6372.40000 |### (9300)
6404.70000 |######## (22649)
6437.00000 |################# (48576)
6469.30000 |################################## (94873)
6501.60000 |################################################### (144323)
6533.90000 |################################################################### (187000)
6566.20000 |######################################################################## (199355)
6598.50000 |########################################################## (153445)
6630.80000 |################################ (88764)
6663.10000 |############# (36489)
6695.40000 |### (8865)
6727.70000 | (1266)
6760.00000 | (117)
            \-------------------------------------------------------------------
```

Data of bird ovomucoids

Tree-island profile:

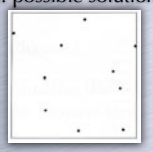
Island	Size	First tree	Last tree	Score	First replicate	Times hit
1	101	1471	1571	3890	3	1
2	69	4961	5029	3890	10	1
3	1	14236	14236	3890	30	1
4	5	14237	14241	3890	31	1
5	1236	1	1236	3891	1	2
6	96	8816	8911	3891	20	1
7	256	9524	9779	3891	23	2
8	960	14242	15201	3891	32	2
9	624	15202	15825	3891	33	1
10	3085	15870	18954	3891	37	1
11	1488	1911	3398	3892	5	1
12	626	3448	4073	3892	7	2
13	211	4074	4284	3892	8	1
14	96	5935	6030	3892	12	1
15	156	7211	7366	3892	14	2
16	236	7367	7602	3892	15	2
17	1053	7667	8719	3892	18	1
18	291	9780	10070	3892	24	1
19	576	10071	10646	3892	25	1
20	96	18987	19082	3892	39	1
21	676	4285	4960	3893	9	1
22	905	5030	5934	3893	11	1
23	96	8720	8815	3893	19	1
24	612	8912	9523	3893	21	1
25	32	18955	18986	3893	38	1
26	18	19083	19100	3893	40	1
27	339	1572	1910	3894	4	2
28	64	7603	7666	3894	17	1
29	3589	10647	14235	3894	26	1
30	234	1237	1470	3895	2	1
31	49	3399	3447	3895	6	1
32	44	15826	15869	3895	35	1
33	1180	6031	7210	3896	13	1

Results of heuristic search, with random addition from random starting points

Data of bird ovomucoids

Branch and bound searching

- Hendy & Penny (1982) first used this method for inferring phylogenies.
- An example - shortest Hamiltonian path (SHP)
  - For *n* cities, there will be *n*-1 cities to come next, so there will be *n*! possible solution

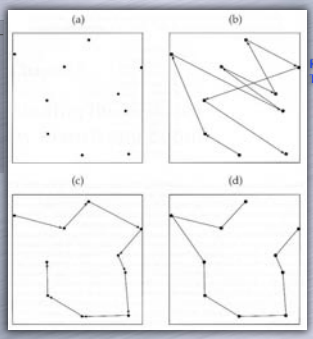


Hendy, M.D. & D. Penny (1982) *Mathem. Biosci.* 60: 133-142.

Felsenstein (2004) *Inferring phylogenies*

A shortest Hamiltonian path problem

Point	x	y
1	0.537	0.061
2	0.274	0.222
3	0.016	0.837
4	0.871	0.400
5	0.399	0.740
6	0.815	0.531
7	0.587	0.946
8	0.992	0.733
9	0.268	0.481
10	0.895	0.068



Random route:
Total length=5.4342

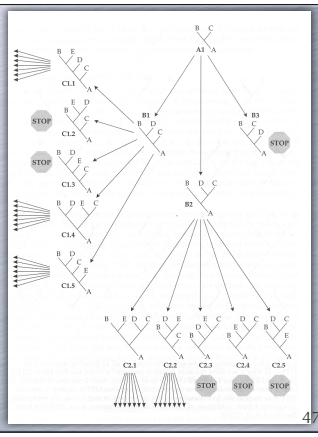
Adding by greedy manner:
Length=2.8027

Optimal solution:
Length=2.7812

Felsenstein (2004) *Inferring phylogenies*

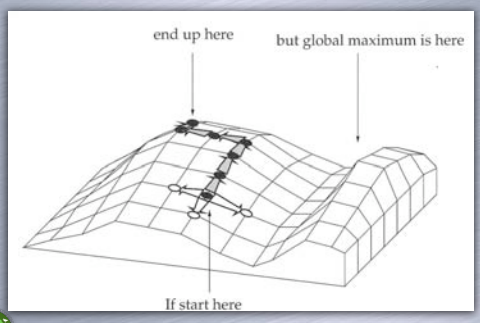
Branch-and-bound search

- This method can greatly improve the speed of search, but it still cannot escape from the constraints of the NP-hardness proof.



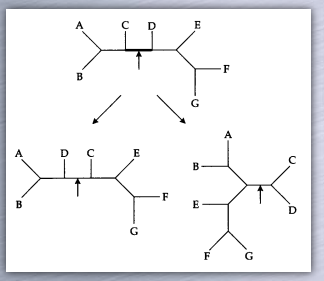
Page & Holme (1998) *Molecular evolution*

heuristic searching for best tree



Felsenstein (2004) *Inferring phylogenies*

Heuristic approach- Branch swapping: NNIs



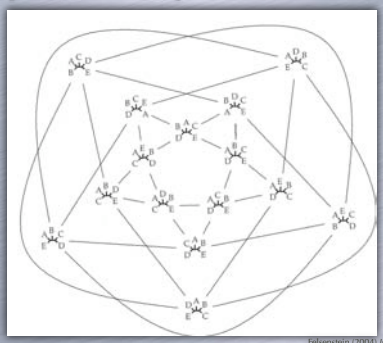
Nearest-neighbor interchanges (NNI)

Swofford et al. (1996) *Molecular Systematics*

How greedy should we be?

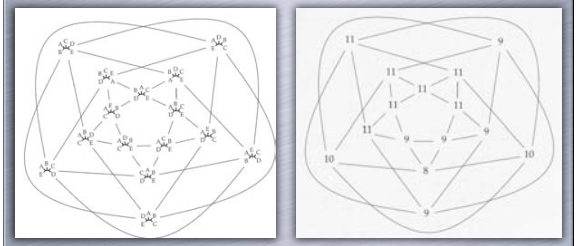
- In a tree with n tips, there will be n-3 interior branches.
- In all, 2(n-3) neighbors will be examined for each tree.
- Should we do NNI for each of the best trees, or stop at some point, or should we start from scratch more often?

The space of all 15 possible unrooted trees



Felsenstein (2004) *Inferring phylogenies*

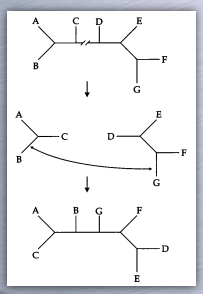
The space of all 15 possible unrooted trees



NNI is to moving in this graph

Felsenstein (2004) *Inferring phylogenies*

Heuristic approach- Branch swapping: TBR



Tree bisection and reconnection (TBR)

In a n_1+n_2 species in the tree, there will be ($2n_1$-3)($2n_2$-3) possible ways to reconnect the two trees.

Swofford et al. (1996) *Molecular Systematics*

Browsing tree space

- Rearrangement
- Random starting point strategy to avoid being trapped in "tree island" (Maddison 1991)
- Parsimony ratchet
  - Parsimony ratchet uses a re-weighted subset of characters as a starting point to browse the tree space, and tries to find as many tree islands as possible by using adjacent-tree searching method. (Nixon 1999)

Maddison, D.R. (1991) *Syst. Zool.* 40: 315-328.
Nixon, K.C. (1999) *Cladistics* 15: 407-414.

Parsimony ratchet

- Described by Nixon (1999), and available in various programs (NONA, WINCLADA, PAUPRat, etc.).
- Parsimony ratchet uses a re-weighted subset of characters as a starting point to browse the tree space, and tries to find as many tree islands as possible by using adjacent-tree searching method.

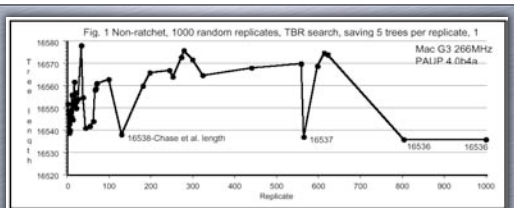
Nixon, K.C. (1999) *Cladistics* 15: 407-414.

55

Parsimony ratchet procedure

- Generate a starting "Wagner" tree
- Randomly select a subset of character (5-25%) and add 1 weight score
- Performing TBR branch swapping, keep 1 tree
- Reset the weighting to original
- Performing branch swapping, keep the best tree
- Return to step 2, repeat the iteration (step2-6) 50-200 times

56



Nixon, K.C. (1999) *Cladistics* 15: 407-414.

57

Notes on parsimony ratchet

- It seems to improve the effectivity of tree searching
- It can be used with any objective function based on character data: compatibility, distance matrix, and likelihoods.
- The strategy can be modified

58

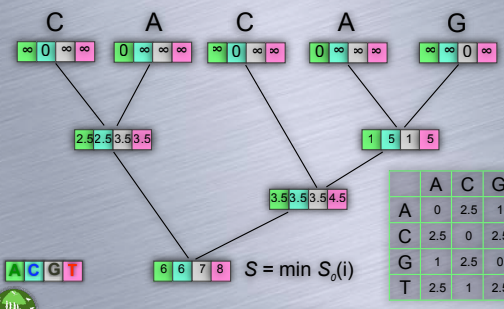
Weighted parsimony method

- Incorporating simple nucleotide models into MP analysis
  - Transition vs. transversion scores

	A	C	G	T
A	0	2.5	1	2.5
C	2.5	0	2.5	1
G	1	2.5	0	2.5
T	2.5	1	2.5	0

59

The Sankoff algorithm applied to the tree



	A	C	G	T
A	0	2.5	1	2.5
C	2.5	0	2.5	1
G	1	2.5	0	2.5
T	2.5	1	2.5	0

Modified from Felsenstein (2004) *Inferring phylogeny*

60

How do we know the tree we got is right?

The evaluation

Evaluating trees

- Character properties (CI, RI)
  - Examining how "clean" is the data on a given tree
- Confidence level of trees
  - Comparisons between obtained trees
    - Partition distance
    - Kishino-Hasegawa test
    - Distance test
    - Likelihood ratio test
  - Bootstrap/jackknife (internal support)
  - Bremer support (for parsimony)

62

Properties of characters

- **Consistency index** (CI) of each character is:

$$\text{character CI} = m_i/S_i$$

- Tree CI for all characters in a specific tree:

$$\text{tree CI} = \frac{\sum_{i=1}^{n} m_i w_i}{\sum_{i=1}^{n} w_i S_i}$$

m_i =所有可能演化樹中的i特徵的最少可能演化步驟 (minimum conceivable steps)
S_i =特定演化樹中的i特徵的演化步驟
w_i = i特徵的權重；n為特徵數

63

Properties of characters

- Retention index (RI) of each characters is:

$$\text{character RI} = \frac{M_i - s_i}{M_i - m_i}$$

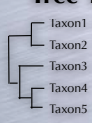
m_i =所有可能演化樹中的i特徵的最少可能演化步驟 (minimum conceivable steps)
M_i =所有可能演化樹中的i特徵的最多可能演化步驟 (maximum conceivable steps)
S_i =特定演化樹中的i特徵的演化步驟

64

Example: calculating CI & RI

Taxon1	0 0 0 0
Taxon2	0 0 0 1
Taxon3	1 1 0 1
Taxon4	1 1 1 1
Taxon5	1 1 0 1

Tree 1



Character(1) CI=1/1=1
Character(1) RI=(2-1)/(2-1)=1

$$TreeCI = \frac{1+1+1+1}{1+1+1+1} = 1$$

Tree 2

Character CI=1/2=0.5
Character RI=(2-2)/(2-1)=0

$$TreeCI = \frac{1+1+1+1}{2+2+1+1} = 0.67$$

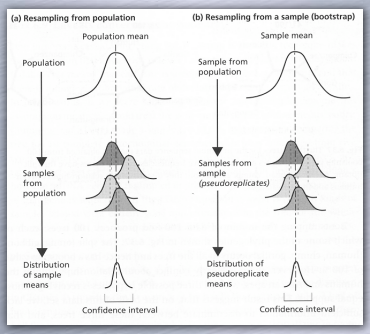
65

Evaluating trees

- Character properties (CI, RI)
- Confidence level of trees
  - Comparisons between obtained trees
  - Bootstrap/jackknife (internal support)
  - Bremer support (for parsimony)

66

Bootstrap method



Page & Holme (1998) *Molecular evolution*

Bootstrap method

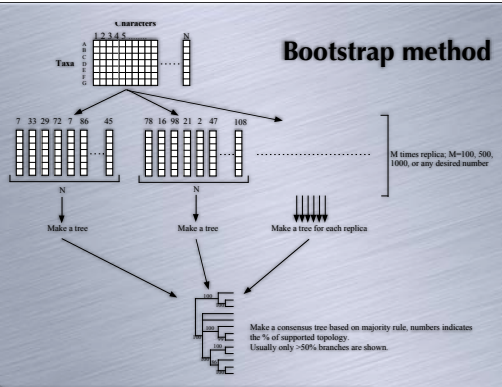


Diagram explaining the bootstrapping procedure used in phylogenetic analysis.

Jackknife method

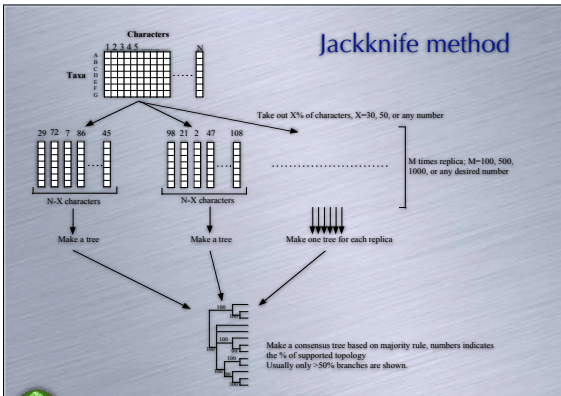
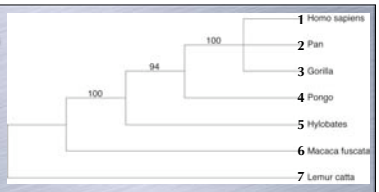


Diagram explaining the Jackknife procedure used in phylogenetic analysis.

Primate mtDNA phylogeny

MP criteria
Bootstrap 1000 replicates



1234567	Freq	%
.....**	1000.00	100.0%
...****	999.00	99.9%
....***	945.00	94.5%
.**....	497.33	49.7%
..*****	484.83	48.5%

Gorilla grouped with Homo/Pan or other primates

Bootstrapping

- Pseudo-resampling.
- An approximate measure of repeatability and accuracy of data.
- Will not correct the inconsistency caused by reconstruction method.
  - The tree constructed in every replicate still depend on the methods you choose, therefore subject to associated problems

Decay/support index

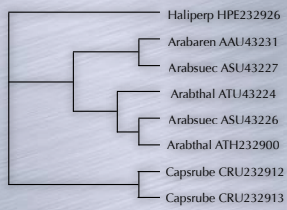
- Also called 'Bremer support' (Bremer 1988, 1994; Donoghue et al. 1992)
- An index of support calculating the difference in tree lengths between the shortest trees that contain versus lack a specific group.
- Constraint trees can be generated by AutoDecay.

Method of calculation

- **1. Obtain most parsimonious tree(s) (MP)**
  - Generate a strict consensus tree
- **2. Obtain all trees one step longer (MP + 1)**
  - Generate a strict consensus tree
  - If branch is not supported, Decay index = 1
- **3. Obtain all trees one step longer (MP + 2)**
  - Generate a strict consensus tree
  - If branch is not supported, Decay index = 2
- **4. Obtain all trees one step longer (MP + 3)**
  - Generate a strict consensus tree
  - If branch is not supported, Decay index = 3

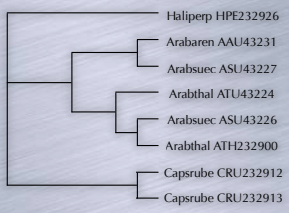
Bremer Support (decay index)

- the number of extra steps it takes to collapse a group
  - example: 1 MP tree, 110 steps



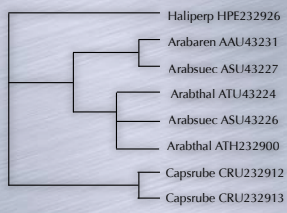
Bremer Support (decay index)

- 3 trees <= 111 steps



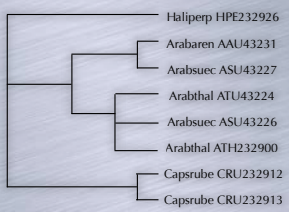
Bremer Support (decay index)

- 3 trees <= 111 steps



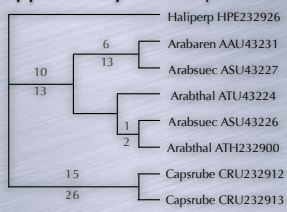
Bremer Support (decay index)

- 5 trees <= 117 steps



Bremer Support (decay index)

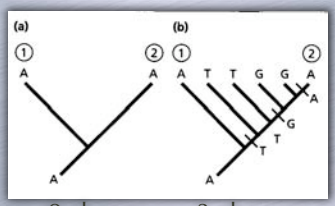
- Can not be larger than the branch length
- No direct connection to branch length otherwise
- Quantification of support in a parsimony framework



Methodological concerns

- Sampling problem
- Performance of phylogenetic methods under computer simulation

Effects of sampling



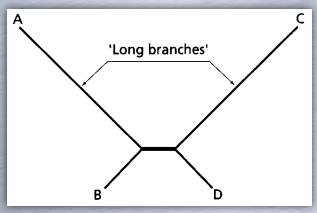
0 change 3 changes

Makes No.2 a fast evolving taxon

(Page & Holmes, 1998)

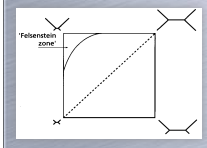
Long branch attraction

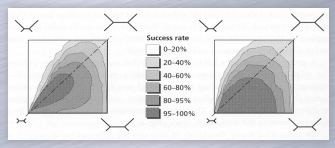
Joe Felsenstein (1978, *Syst. Zool.* 27: 401-410)



Page & Holme (1998) *Molecular evolution*

Simulation analysis

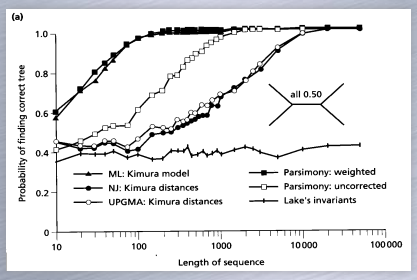




UPGMA Parsimony

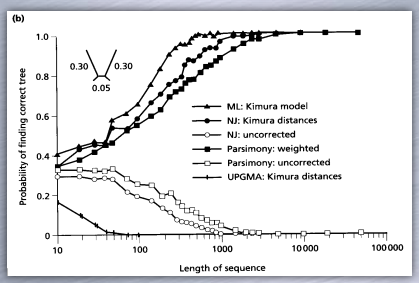
Page & Holme (1998) *Molecular evolution*

Equal rate of evolution

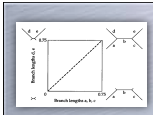


Page & Holme (1998) *Molecular evolution*

Unequal rate of evolution



Page & Holme (1998) *Molecular evolution*



Parsimony

Weighted parsimony

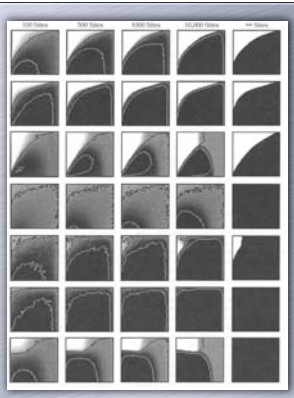
UPGMA

Lake's invariants

Maximum likelihood (Jukes-Cantor)

Maximum likelihood (Kimura)

Weighted least squares



Swofford et al. (1996) *Molecular Systematics*

References of phylogenetics

- Graur, D. and W.-H. Li. 2000. **Fundamentals of Molecular Evolution**. 2nd ed., Sinauer Assoc., Sunderland, MA, USA.
- Hall, B. G. 2004. **Phylogenetic trees made easy: a how-to manual, 2nd ed.** Sinauer Assoc., Sunderland, MA.
- Hillis, D. M., C. Moritz, and B. K. Mable (eds) 1996. **Molecular systematics**. Sinauer Assoc., Sunderland, MA.
- Page, R. D. M., and E. C. Holmes. 1998. **Molecular evolution - A phylogenetic approach**. Blackwell Science Ltd, Oxford, the United Kingdom.
- Yang, Z. H. 2006. **Computational molecular evolution**. Oxford University Press.