A Brief Introduction to Phylogenetic Analysis

Li Zhe

2007-04-16

Before Start

No model, no inference!

- Mathematical model is an idealization of the real-world phenomenon and never a complete accurate representation.
- We need a probability model to relate what we observe (data) to what we want to know (hypothesis or parameters).

Outline

- Quick Review of Probability and Statistics
- Distance Measure between Sequences
- Phylogenetic Trees

Probability – a Quick Review

- Sample space
- Events
- Probability measure
- Conditional probability
- Independent events
- Random variable

Survey of Statistics

- Population, sample, statistic
- Inference
  - Estimate
    - Maximum likelihood estimate
  - Hypothesis test

Outline

- Quick Review of Probability
- Evolutionary Distance between Sequences
  - What is Evolutionary Distance?
  - Approximate Methods
  - Maximum likelihood Methods
- Phylogenetic Trees

Evolutionary Distance

- An ideal evolutionary distance measure should be proportional to the divergence "time" between sequences

- An intuitive thought – count the difference between sequences – *p-distance*

$$p = \frac{\#different\_site}{\#total\_site}$$

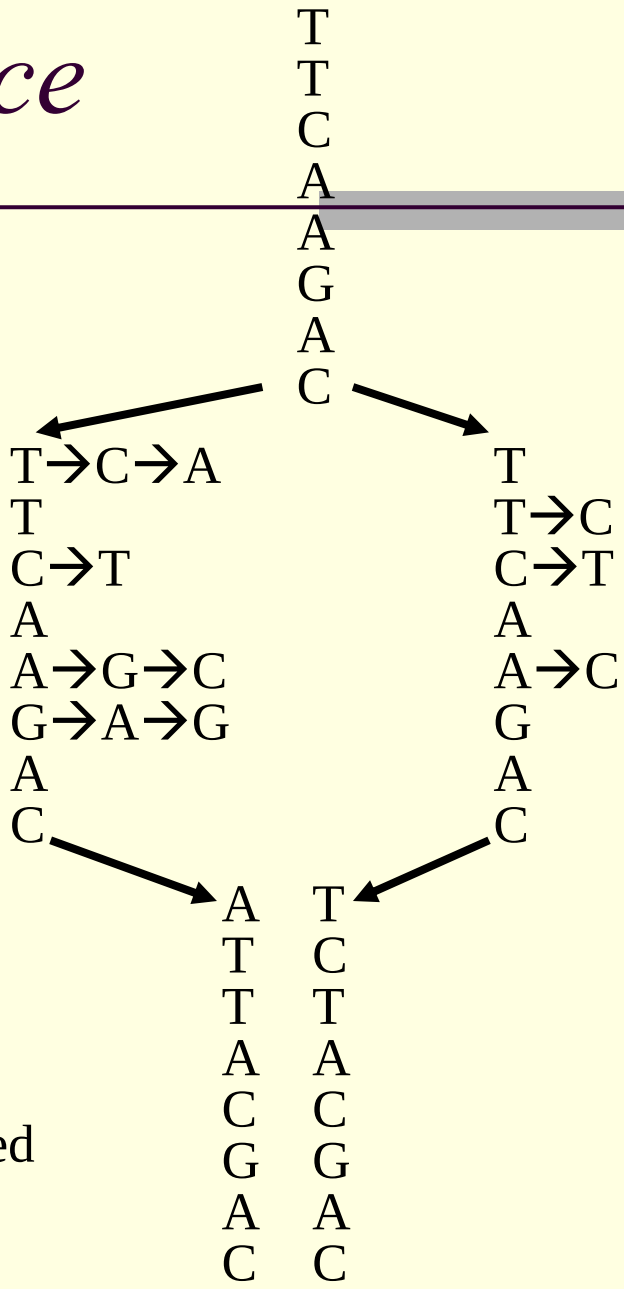
Only suitable for very closely related sequence…

Problem of *p-distance*

Multiple hits at the same site

multiple substitution
single substitution
parallel substation

convergent substitution
back substation

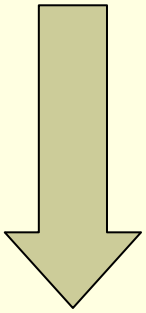


2 differences observed at 8 sites

10 substitutions have actually occurred

K-values

p-distances



Multiple substitution at the same site

K-values

Average substitution number among sites

K_s (d_S) – Synonymous substitution rate

K_a (d_N) – Nonsynonymous substitution rate

Estimate K-values

- Approximate method
  - Involve ad hoc treatments that cannot be justified rigorous
    - Nei and Gojobori 1986 (NG)
    - Yang and Nielsen 2000 (YN00)
- Maximum likelihood method
  - Estimate *K-values* based on explicit model of codon substitution
    - Goldman and Yang 1994
    - Muse and Gaut 1994

Overview of Approximate Method

- Count the numbers of synonymous and nonsynonymous sites
  - Potential synonymous/nonsynonymous sites
- Count the number of synonymous and nonsynonymous differences between the two sequences
  - Evolutionary pathways for codons with two or three different sites
- Apply a correction for multiple substitutions at the same site
  - Based on nucleotide substitution model

Evolutionary Pathway

- More than one different sites between codons

- Every parsimony pathway (there is no back-substitution) is equally weighted

TTA(Phe) vs. GTA(Val)
2 pathways:
TTT(Phe) <-> GTT(Val) <-> GTA(Val)
TTT(Phe) <-> TTA(Leu) <-> GTA(Val)
0.5 synonymous substitutions and 1.5 nonsynonymous sustitutions

Nucleotide Substituion Models

Substitution model gives the matrix of substitution rate

Common assumptions: independent among sites, equal rates among sites.

JC69

$$\begin{bmatrix} . & \alpha & \alpha & \alpha \\ \alpha & . & \alpha & \alpha \\ \alpha & \alpha & . & \alpha \\ \alpha & \alpha & \alpha & . \end{bmatrix}$$

K80

$$\begin{bmatrix} . & \alpha & \beta & \beta \\ \alpha & . & \beta & \beta \\ \beta & \beta & . & \alpha \\ \beta & \beta & \alpha & . \end{bmatrix}$$

HKY85

$$\begin{bmatrix} . & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & . & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & . & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & . \end{bmatrix}$$

•Equal base freq.
•Transition rate equal transversion rate

Equal base freq.

Less assumption

Pros and Cons of Approximate method

- Pros
  - Quick computation
  - Reasonable results
- Cons
  - No explicit models, ad hoc approximations
    - implicit assumptions
    - difficult to extend
    - hard to evaluate
  - Not a good estimation in statistics (bias, consistency, efficiency)
  - Unrealistic results under some circumstances

Maximum Likelihood Method

- Based on explicit codon substitution model
  - Continuous Markov chain
  - Different base frequency
  - Different transition/transversion rates
  - Different rates on codon sites
- Use maximum likelihood (ML) method to fit the model according to data

Markov Model for Substitution

- Continuous stationary Markov chain

$$i \xrightarrow{P_{ij}(t)} j \qquad P_{ij}(t) = P\{X(t+s) = j \mid X(s) = i\}$$

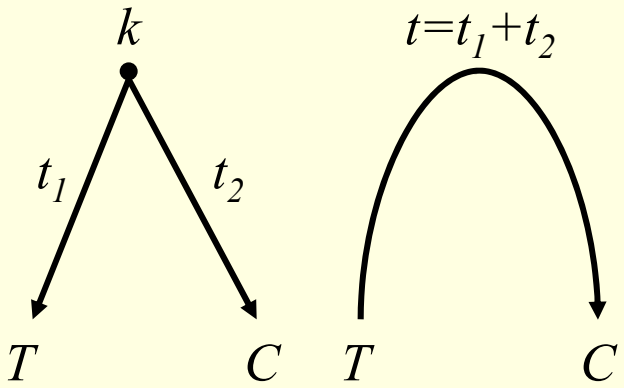
$$i, j \in \{\text{codons}\}$$

- Time reversibility

Equilibium assumption

$$i \underset{P_{ji}(t)}{\overset{P_{ij}(t)}{\rightleftarrows}} j \qquad \pi_i P_{ij}(t) = \pi_j P_{ji}(t)$$

$$i, j \in \{\text{codons}\}$$



Transition Probability

$$P(t) = \left[p_{ij}(t)\right]_{61\times 61} = e^{Qt}, \text{ where } Q = \left[q_{ij}\right]_{61\times 61}, \text{ and } \sum_{j=1}^{61} q_{ij} = 0$$

Q is the instantaneous substitution rates, i.e., the substitution rates in an infinitesimal of time *dt*.

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at more than one position,} \\ \pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion,} \\ \kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition,} \\ \omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition.} \end{cases}$$

κ – transition/transversion ratio

ω – nonsynonymous/synonymous rate ratio

Scale of Instantaneous Rate Matrix

Q is scaled that time *t* is effectively measured as expected numbers of nucleotide substitution per codon.

$$\sum_{i=1}^{61}\sum_{j\neq i}\pi_i q_{ij} = -\sum_{i=1}^{61}\pi_i q_{ii} = 1$$

Maximum Likelihood Estimate of K-values

The probability of observing a codon site with codons *i* and *j* in the two sequences is

$$f_{ij}(t) = \pi_i p_{ij}(t)$$

Assuming independence and same rates of codon sites, the log-likehood function is

$$\ell(t, \kappa, \omega) = \sum_{i,j} n_{ij} \log\{ f_{ij}(t)\}$$

Thus we can get ML estimates of parameters (*t*, κ , ω).

Maximum Likelihood Estimate of K-values

The synonymous and nonsynonymous substitution rate per codon are

$$\rho_S^* = \sum_{i \neq j, aa_i = aa_j} \pi_i q_{ij}, \text{ and } \rho_N^* = 1 - \rho_S^*$$

With ω = 1 fixed, the numbers of synonymous and nonsynonymous sites per codon are

$$\rho_S^1, \text{ and } \rho_N^1$$

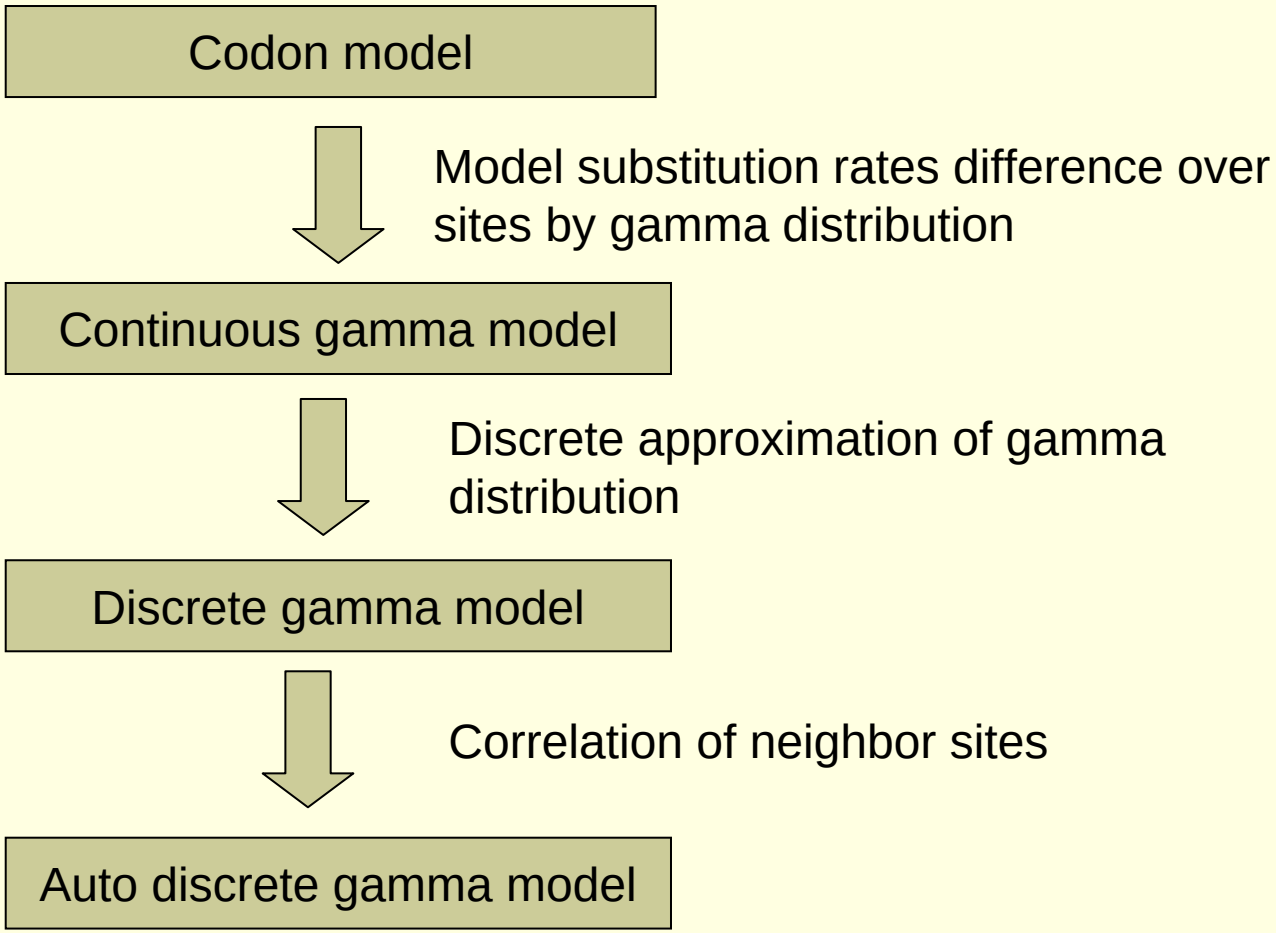
Then the K-values are

$$K_s = t\rho_S^* / (3\rho_S^1),$$

\ and

$$K_a = t\rho_N^* / (3\rho_N^1).$$

Extension of Codon Model



Pros and Cons of ML Method

- Pros
  - Explicit probability model
    - Easy to interpret
    - Easy to evaluate (likelihood ratio test)
    - Good statistic (consistency, efficiency)
    - Easy to extend
  - Can handle multiple sequences alignment with guided tree
- Cons
  - Need large data set (>=300codons)
  - Strong assumption on equilibrium
    - Erroneous results when equilibrium is not reached
  - Intensive computation
    - Much more slower than approximate method

Method choice

- Approximate method (NG)
  - Equal (nearly) nucleotide frequency
  - Same rates among sites
  - Nearly neutral
  - Small data set
  - Equilibrium not reached
- ML method (Goldman & Yang)
  - Large data set
  - Equilibrium
  - Various rates among sites (*)
  - Selection pressure
  - Nested model test

Implementation – PAML

PAML: Phylogenetic Analysis by Maximum Likelihood

- baseml – ML analysis of nucleotide sequences
- codeml – implements the codon substitution model
- yn00 – implementation of approximate method (YN00)
- chi2 – conducts likelihood ratio test (between nested models)

Outline

- Quick Review of Probability and Statistics
- Evolutionary Distance between Sequences
- Phylogenetic Trees
  - What's Phylogenetic Trees?
  - Build Phylogenetic Trees by DistanceMethods
  - Validate Phylogenetic Trees by Re-sampling

Phylogenetic Trees

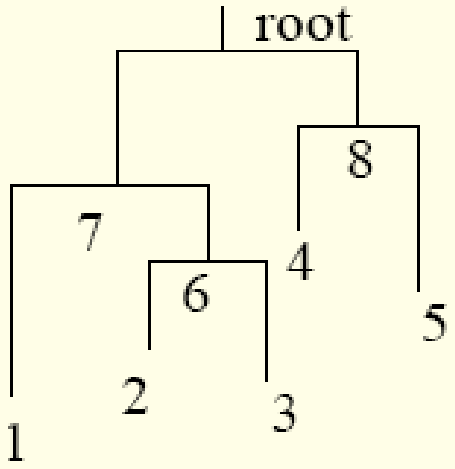
- *Phylogenetics* is the study of evolutionary relationships among organisms

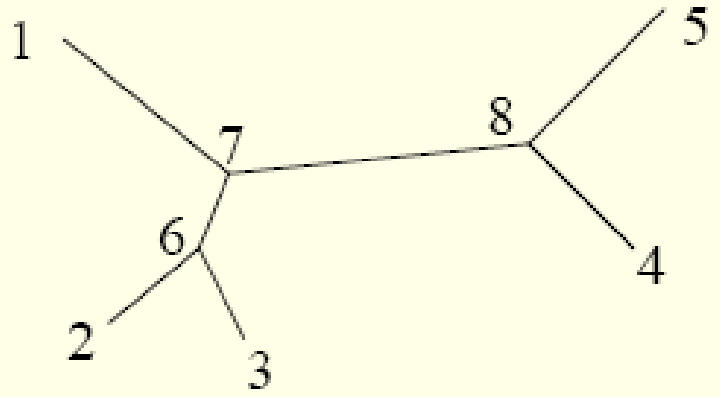
- A *phylogenetic tree* or *phylogeny* for a set of taxa (species, genes, …) is an evolutionary tree representing their relationships.
  - A tree is an acyclic graph: horizontal transfer is ignored
  - Edge weights *may* represent distance in evolution

Phylogenetic Trees

- Trees can be rooted or unrooted.
  - In the case of unrooted trees we can assume to have not enough data to determine the root of the tree
- The leaves of a phylogenetic tree usually represent the present day taxa, the internal nodes represent hypothesized ancestors.

Tree Topology





Why Phylogenetic Trees?

- Evolution of organisms (tree of species)
- Evolution of genes (tree of gene)
- Application:
  - Comparative Genomics
  - Gene function prediction

Models and Methods

- Maximum Parsimony methods
- Distance Matrix methods
- Maximum Likelihood methods

- Which is better?

Maximum Parsimony

- Variation is small
- All possible trees are evaluated
  - <=11 or 12 sequences concerned
  - Time-consuming
- Concensus tree for more than one MP trees

Distance Matrix methods

- Variation is intermediate
- Hierarchical inference
  - Rather faster then MP.
  - Large number of sequences
- The distance matrix can be derived from multiple alignment or evolution event or others like K-tuple method

Maximum Likelihood

- Variation could be some larger
- All possible trees are evaluated
  - <=11 or 12 sequences concerned
- Both topology and edge lengths are considered.
  - based on probability inference.

How many possible trees?

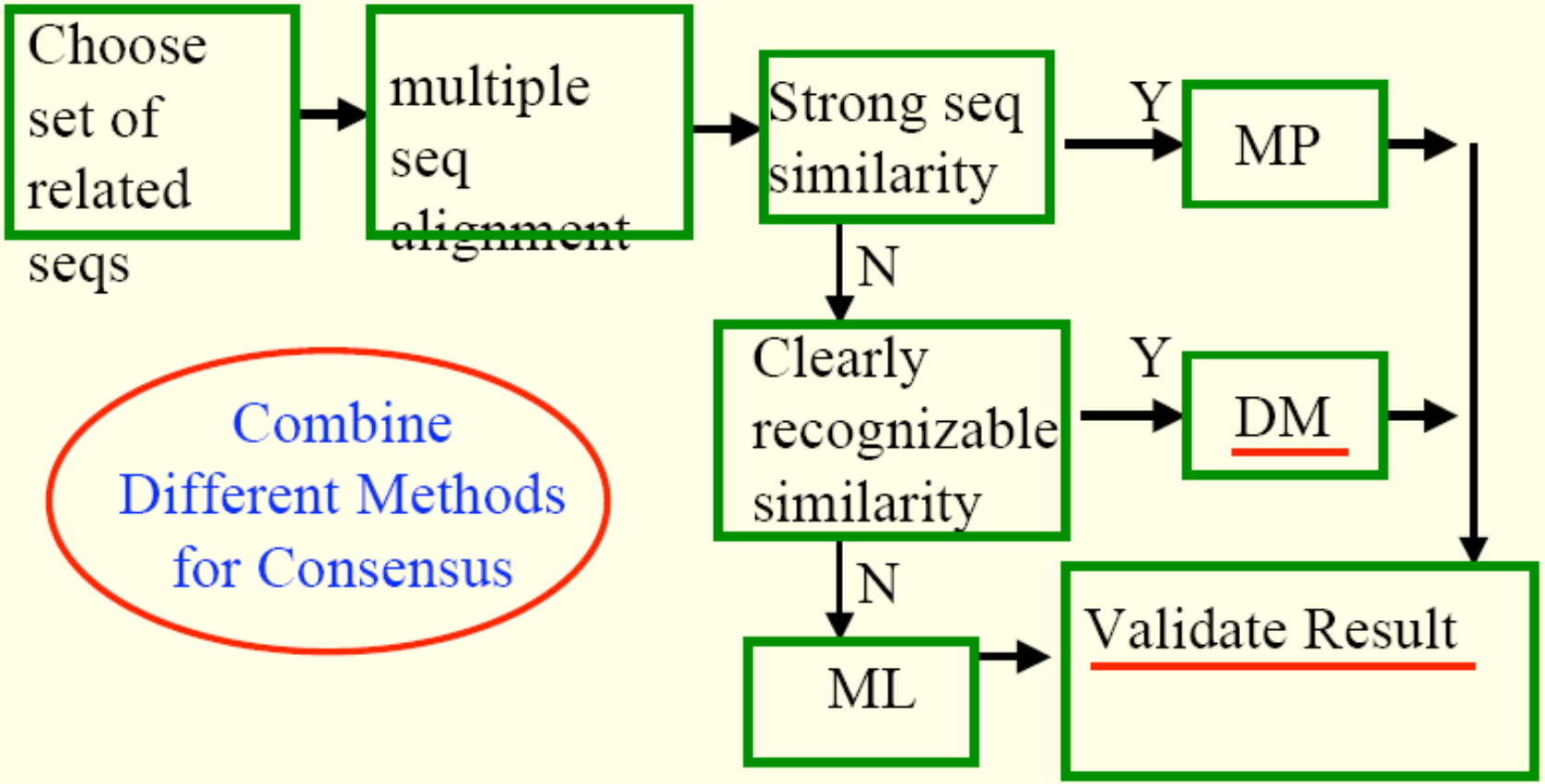
Rooted tree $\frac{(2m-3)!}{2^{m-2}\cdot(m-2)!}$ **m=10: 34,459,425**

Unrooted tree $\frac{(2m-5)!}{2^{m-3}\cdot(m-3)!}$ **m=10: 2,027,025**

A Quick Summary

	MP	DM	ML
Variation	+	++	+++
Computation Complex	++	+	+++
Edge Length Estimation	N	N	Y
Flexibility	+	+++	++

A General Protocol



Distance Methods

- Neighbors – the closest taxa
- Rather fast
- More reliable than MP when branch lengths vary (Jin and Nei, 1990; Swofford et al. 1996)
- Additive: the lengths be additive

Neighbors Joining

- Proposed by Saitou and Nei in 1987
  - Pearson et al. enhance NJ in 1999 (Not a single tree predicted)
- Pairing sequences based on the effect of the pairing on the sum of the branch lengths of the tree
- Starting from a star-like tree

Similarity to Distance

- Convert alignment scores to distances:

$$D = -\log S_{eff} = -\log\{(S_{obs} - S_{rand})/(S_{\max} - S_{rand})\}$$

S_{obs} is observed pairwise alignment score

$S_{\max}$ is the maximum score, the average of the score of aligning either sequence to itself.

S_{rand} is the expected score for aligning two random sequences of the same length and residue composition, which can be calculated by random shuffling of the two sequences or by an approximate calculation given in Feng & Doolittle[1996]

Neighbour Joining Algorithm

- For each node i the distance from the rest of the tree is estimated by

$$r_i = \frac{1}{N-2}\sum_{k \neq i} d_{i,k}$$

- Choose the nodes *i* and *j* that for which

$D_{ij} = d_{ij} - r_i - r_j$ is smallest

join i and j (ij is new node)

- Compute branch length from i and j to ij

$$d_{i,(ij)} = \frac{1}{2}d_{i,j} + \frac{1}{2}(r_i - r_j), d_{j,(ij)} = \frac{1}{2}d_{i,j} + \frac{1}{2}(r_j - r_i)$$

- Compute the distances between the new cluster and each other cluster:

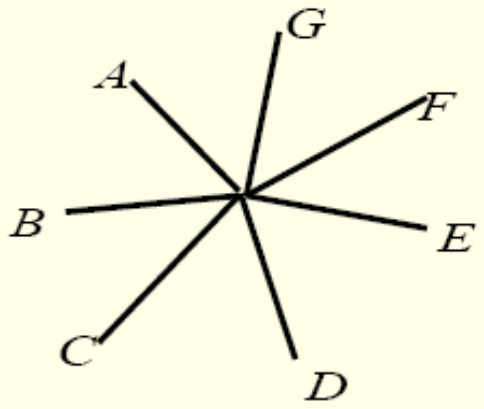
$$d_{(ij),k} = \frac{d_{i,k} + d_{j,k} - d_{i,j}}{2}$$

Neighbour joining algorithm(1)

	A	B	C	D	E	F	G	r_i
A		63	94	111	67	23	107	88.4
B	63		79	96	16	58	92	80.8
C	94	79		47	83	89	43	87
D	111	96	47		100	106	20	96
E	67	16	83	100		62	96	84.4
F	23	58	89	106	62		102	88
G	107	92	43	20	96	102		92

No molecular clock assumption

Start from the star-like tree
Calculate r_i



Neighbour joining algorithm(2)

	A	B	C	D	E	F	G	r_i
A		-106.2	-81.4	-73.4	-105.8	-153.4	-69.4	88.4
B	63		-88.8	-80.8	-149.2	-110.8	-80.8	80.8
C	94	79		-136	-84.4	-86	-136	87
D	111	96	47		-80.4	-78	-168	96
E	67	16	83	100		-110.4	-80.4	84.4
F	23	58	89	106	62		-78	88
G	107	92	43	20	96	102		92

Calculate D_{ij} , D and G are the closest

Calculate the branch lengths of D and G

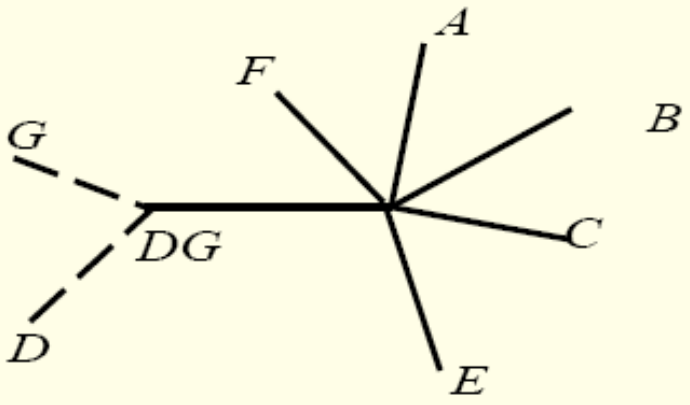
$d = 12$

$g = 8$

Neighbour joining algorithm(3)

	A	B	C	E	F	DG	r_i
A		63	94	67	23	94	85.25
B	63		79	16	58	84	75
C	94	79		83	89	35	95
E	67	16	83		62	88	79
F	23	58	89	62		94	81.5
DG	94	84	35	88	94		91.25

Join D and G, calculate the distances r_i from DG to other nodes



Neighbour joining algorithm(4)

	A	B	C	E	F	DG	r_i
A		-97.25	-86.25	-97.25	-143.75	-82.5	85.25
B	63		-91	-138	-98.5	-82.25	75
C	94	79		-91	-87.5	-151.25	95
E	67	16	83		-98.5	-82.25	79
F	23	58	89	62		-78.75	81.5
DG	94	84	35	88	94		91.25

Calculate D_{ij} , C and DG are the closest

Calculate the branch lengths of C and DG

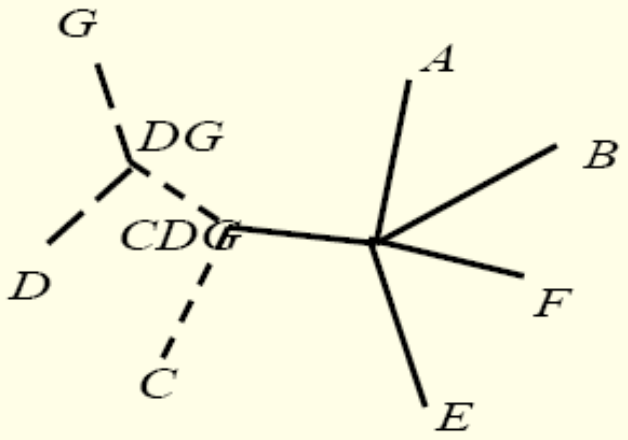
$c = 19.375$

$dg = 15.625$

Neighbour joining algorithm(5)

	A	B	E	F	CDG	r_i
A		63	67	23	61	71.3
B	63		16	58	64	67
E	67	16		62	60	68.3
F	23	58	62		74	72.3
CDG	61	64	60	74		98.3

Join DG and C, calculate the distances r_i from CDG to other nodes



Neighbour joining algorithm(6)

	A	B	E	F	CDG	r_i
A		-75.3	-72.6	-120.6	-108.6	71.3
B	63		-119.3	-81.3	-101.3	67
E	67	16		-78.6	-90	68.3
F	23	58	62		-96.3	72.3
CDG	61	64	60	74		98.3

Calculate D_{ij} , A and F are the closest

Calculate the branch lengths of A and F

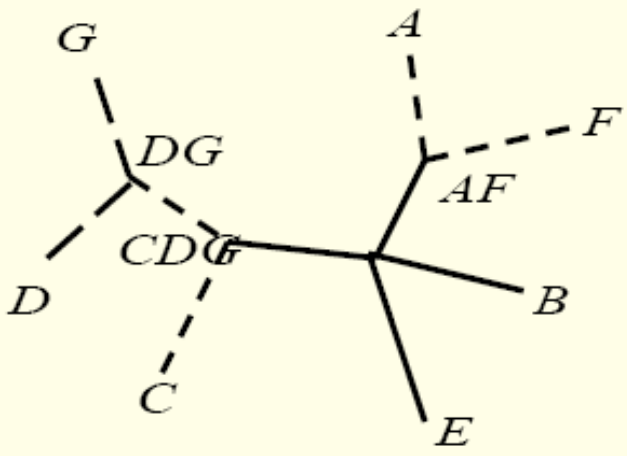
$$a = 11$$

$$f = 12$$

Neighbour joining algorithm(7)

	AF	B	E	CDG	r_i
AF		98	106	112	158
B	98		16	64	89
E	106	16		60	91
CDG	112	64	60		118

Join A and F, calculate the distances r_i from AF to other nodes



Neighbour joining algorithm(8)

	AF	B	E	CDG	r_i
AF		-149	-143	-164	158
B	98		-164	-143	89
E	106	16		-149	91
CDG	112	64	60		118

Calculate D_{ij} , B and E are the closest

Calculate the branch lengths of B and E

$b = 7$

$e = 9$

Neighbour joining algorithm(10)

	AF	BE	CDG	r_i
AF		-408	-408	300
BE	188		-408	296
CDG	112	108		220

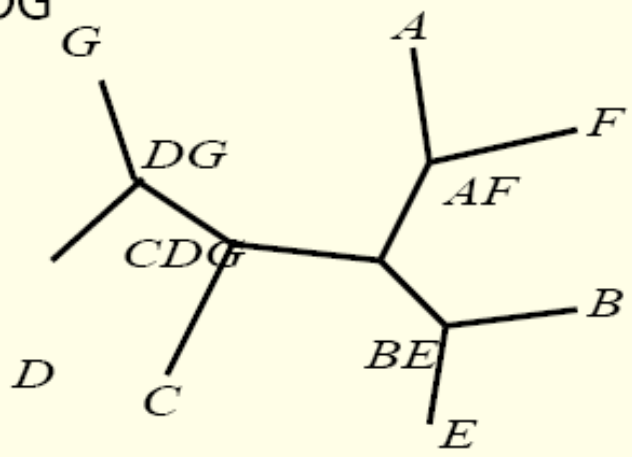
Calculate D_{ij} , BE and CDG are the closest

Calculate the branch lengths of BE and CDG

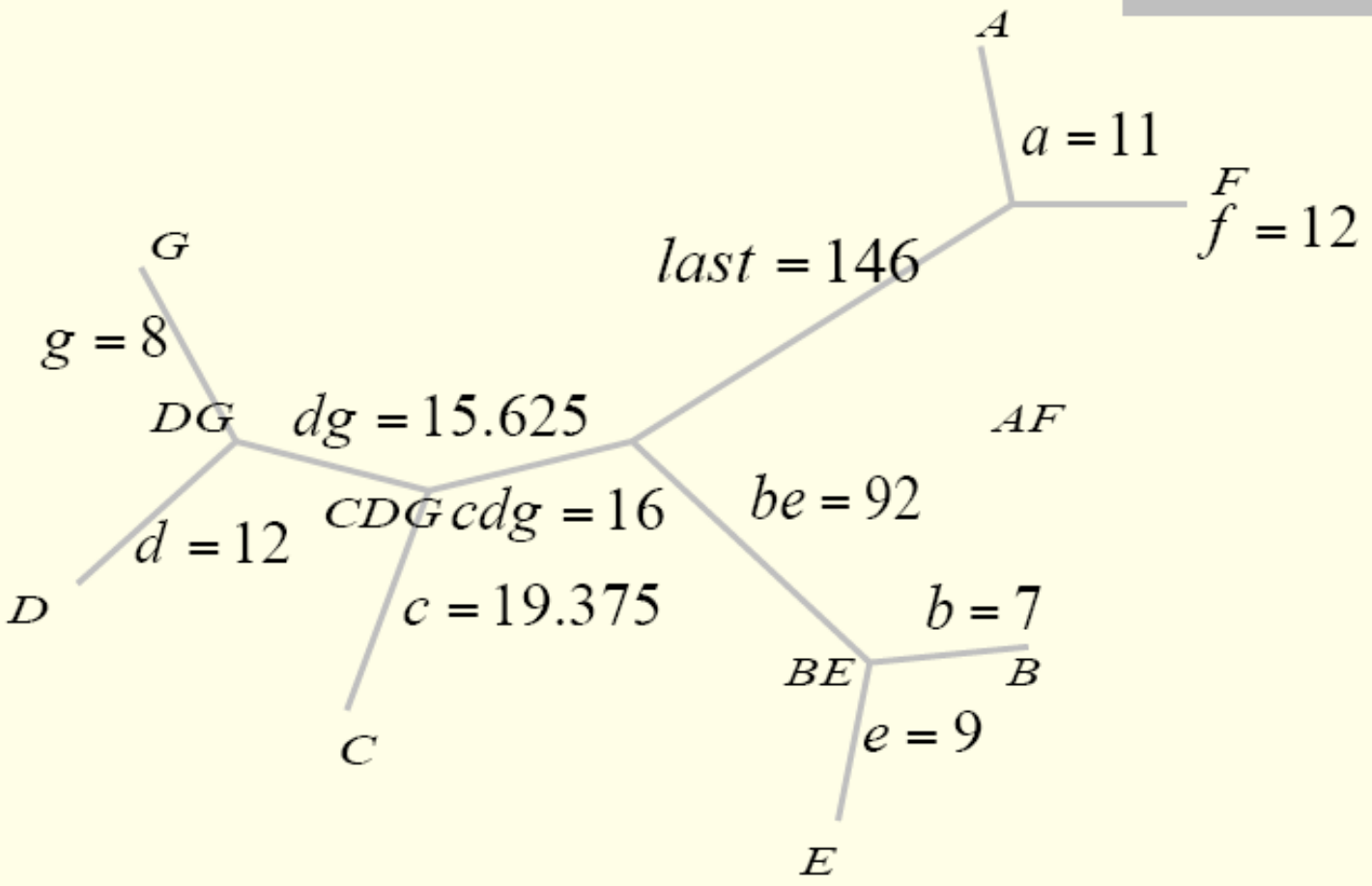
$be = 92$

$cdg = 16$

Join BE and CDG, calculate the distances from BECDG to the last node AF :146



Neighbour joining algorithm(11)



A Quick Summary

- NJ is fast and reliable for topology
  - But not edges length
- NJ do not necessarily assume molecular clock.
  - But it guarantees the assumption hold if required.
- Distances should hold Triangle Law.

Validate the Inference

- Phylogenetic trees are inferred based on Model
  - Hypothetical Inference
- How reliable are the result?
  - Reliability vs. Stability
  - Validate the result by Re-sampling.

Bootstrap(1)

- Given a dataset consisting of an alignment of sequences, an artificial dataset of the same size is generated
  - by picking columns from the alignment at random with replacement.

- One given column in the original dataset can therefore appear several times in the artificial dataset

Bootstrap(2)

- The tree building algorithm is then applied to this new dataset, and the whole selection and tree building procedure is repeated typically 100 times.
- The frequency with which a chosen phylogenetic feature appears is taken to be a measure of the confidence we can have in this feature.
- At last, a consensus tree is created

Validate the Tree

- To improve prediction of trees and assist with localization of the root, an outgroup could be set.

- An outgroup of the following criteria:
  - From species that are known to have separated from the others at an early evolutionary time
  - More distantly related with other sequences

More words on Outgroup

- More than one can be selected
- By independently information, such as fossil evidence
- Too distant an outgroup may lead to incorrect prediction

NJ @ PHYLIP

- Multiple alignment: clustalw, t-coffee, muscle
  - save the output in phylip format (*.phy)
- Bootstrap the sequence data: SEQBOOT
- Build Phylogenetic trees: NEIGHBOR
- Calc Consensus : CONSENSUS

Thanks for your patience!