

Bioinformatics

Task 2012

1, Compare the usage of the protein & nucleotide sequence alignments (1)

Nucleotide: harder to align, mutations happen on this level,
some methods have theoretical background only for nuc. seq.
→ nucleotide substitution models
some methods work only with nucleotide seq.
→ max. likelihood (mostly)

Protein: easier to align, some distance matrices developed for proteins, parsimony methods developed for proteins, in case of frame shifts, homology is not meaningful, losing info of same-sense mutations

2, Define the following terms: homolog, ortholog, paralog (3)

Homolog: two proteins are homologues if they share a common ancestor. There are two cases: orthologs & paralogs

Two proteins are orthologous if they are in different species, but they have the same function

Two proteins are paralogs if they are in the same organism and they have different functions

3, Define the distance in a cladogram (2)

The distance between two taxa ~~are~~ is the number of branching events between them.

...ology: two proteins are homologues if they share a common ancestor. There are two cases: orthologs & paralog. Two proteins are orthologous if they are in different species, but they have the same function. Two proteins are paralogous if they are in the same organism and they have different functions.

3, Define the distance in a cladogram (2)

The distance between two taxa ~~are~~ is the number of branching points between them.

4, What are the main steps of a typical phylogenetic analysis? Briefly describe each step! (5)

Data selection

→ sequences and/or other data?, gene tree or species tree?

Aligning sequences

→ use supporting info

→ good alignment is necessary but not sufficient precondition

Generating trees

→ Topology, branch lengths, consensus trees

→ ~~Boot~~ Bootstrapping tree scoring

Answer scientific questions

5, What is the homoplasy? Write 2 reasons that can result it (3)
 Similarity that is not homologous:

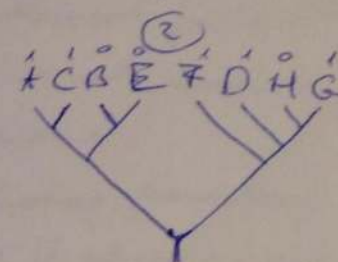
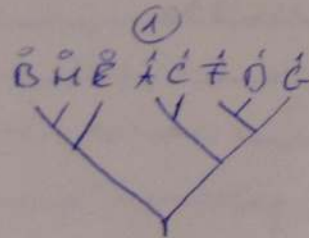
- independent evolution (analogy, convergence, parallelism)
 (e.g. fins of fishes & whales, wings of insects, birds & bats)
- (back mutation) ← Reversal

We can't deduce the relationship of taxa having homoplasy.

→ restore the original sequence & hence the original phenotype

6, We have a dataset & 2 trees about some animals, which tree represents better the evolution of the given attribute of the animals? (6)

A	B	C	D	E	F	G	H
1	0	1	1	0	1	1	0



→ The first tree, because it has the smaller sum of branch lengths and the tree with the smallest sum of branch lengths is the best tree!

7, Write 2 types of tree generation! Describe them briefly! (5)

Parsimony

- simple phylogenetic method close to numerical taxonomy
- works for any data type
- directly deals with the characters & their states

... easier to understand

→ The first tree, because it has the smaller sum of branch lengths and the tree with the smallest sum of branch lengths is the best tree!

7, Write 2 types of tree generation! Describe them briefly! (5)

Parsimony

- simple phylogenetic method close to numerical taxonomy
- works for any data type
- directly deals with the characters & their states
- most popular, easy to understand

Distance based methods

- clustering methods from statistics

UPGMA → clusters the two closest taxa, then recalculates the distance matrix

Neighbor joining → task 10

Model based phylogenetics

- nucleotide substitution models
- sophisticated statistics

Maximum likelihood - probability of seeing the observed data (D) given a model/theory (T) $Pr(D/T)$

Bayesian inference - probability that the model is correct given the observed data $Pr(T/D)$

8, What are the advantages & disadvantages of Parsimony? (4)

Advantages

- simple method, easy to understand
- doesn't depend on an explicit model of evolution
- gives both trees & associated hypotheses of character evolution
- reliable results if
 - data is well structured
 - homoplasy is either rare or widely distributed on the tree

Disadvantages

- misleading if homoplasy is common or concentrated in particular parts of the tree
 - phenophoric convergence
 - base composition biases
- long branch attraction
- underestimating branch lengths
- often justified on purely philosophical grounds

9, List minimum 4 diff. models for DNA sequence evolution! (4)

- Jukes - Cantor model (equal base frequencies ($\frac{1}{4}$), single overall mut.)
- K80 model (diff. mutation rates for transitions & transversions)
- TM92 (unique base frequencies, multiple mutation rates)

reading of homoplasy is common & concentrated in particular parts of the tree

→ homophilic convergence

→ base composition biases

- long branch attraction

- underestimating branch lengths

- often justified on purely philosophical grounds

9, List minimum 4 diff. models for DNA sequence evolution! (4)

Jukes - Cantor model (equal base frequencies ($\frac{1}{4}$), single overall mut.)

K80 model (diff. mutation rates for transitions & transversions)

TN93 (unique base frequencies, multiple mutation rates)

HKY (unique b.f., often used for ML)

TC2 (GC content)

GTR (unique b.f., all mut. rates specified)

10, Describe briefly the Neighbour-joining algorithm! (4)

- constructs tree step-wise

→ starts with a star-shaped tree, adds two taxa (in each step) → calculates the sum of the branch length

→ keeps the shortest tree

- produces unrooted tree

- quick, good guess of true phylogeny

- many very similar implementations

11, Name 2 model based phylogeny algorithms, & define them with one sentence! (3)

Maximum likelihood $\rightarrow$ given the data, is the model/theory acceptable? $Pr(D/T)$

Bayesian inference $\rightarrow$ given the theory, is the data good enough? $Pr(T/D)$

12, What is the basic idea of Maximum likelihood method? (2)

$\rightarrow$ given the data, is the model acceptable? $Pr(D/T)$

13, Describe briefly the algorithm of the Bayesian phylogen. -! (4)

$\rightarrow$ given a model, is the data good enough? $Pr(T/D)$

- it uses the Monte Carlo Markov Chain algorithm

$\rightarrow$ efficient numerical solution

$\rightarrow$ applies subjective probability (controversial)

$\rightarrow$ for many/long nucleotide seq.

14, Why are multiple alignments useful? (3)

They reveal subtle differences

Using them we can find e.g. groups of proteins, & this helps in identifying the function of proteins

15, How can we describe multiple alignment using regular expression? (3)

16, Calculate the entropy of the following multiple align. (4)

$$- \sum_{x=A,T,G,C} P_x \log P_x$$

1. column: $P_A = 1$ $P_T = P_G = P_C = 0$

2. column: $P_A = P_C = 0$ $P_T = 1/2$ $P_G = 1/2$

3. column: $P_A = 1/4$ $P_T = P_G = P_C = 1/4$

4. column: $P_C = P_G = 1/2$ $P_T = P_A = 0$

$$\text{entropy} = - \left(1 \cdot \log 1 + \frac{1}{2} \log \frac{1}{2} + \frac{1}{2} \log \frac{1}{2} + 4 \cdot \left(\frac{1}{4} \log \frac{1}{4} \right) + 2 \left(\frac{1}{2} \log \frac{1}{2} \right) \right)$$

17, What is protein profile? (3)

- a matrix, that contains frequencies of the amino acids in a multiple alignment. The sum of the columns is always 1 & one means total conservation. It reflects how much is an amino acid conserved, it contains gaps too.

e.g.

G	H	T	A	...
G	H	T	P	...
G	P	T	G	...
T	P	T	G	...

MA

⇒

G	3/4		1/2	
H		1/2		
T	1/4		1	
P		1/2		1/4
A				1/4

18, What is the difference ~~that~~ betw. retrieval & similarity search in databases? (4)

- retrieval means an exact match
- similarity search gives mostly similar entries based on a numerical similarity measure

... alignment. The sum of the columns is always 1 & one means total conservation. It reflects how much is an amino acid conserved. It contains gaps too.

e.g.

G	H	T	A	...
G	H	T	P	...
G	P	T	G	...
T	P	T	G	...
M1				



G	3/4			1/2
H	1/2			
T	1/4		1	
P		1/2		1/4
A				1/4

18, What is the difference ~~that~~ betw. retrieval & similarity search in databases? (4)

- retrieval means an exact match
- similarity search gives mostly similar entries based on a numerical similarity measure

19, Describe the main steps of BLAST algorithm! (5)

- ~~It takes~~ - stores database in a hash table, with n-mer words & occurrences
- records words in a query, includes similar words, based on BLOSUM similarity
- selects close entries that share a given no of common words with the query
- from here it selects those where the shared words densely cluster in certain regions (seeds) → High scoring pairs (HSP)
- elongates HSP's & splices them together by some criterion
- calculates significance of the spliced regions using the BLAST formula
- ranks the top scoring entries by significance & presents the scores & the pairwise alignments for the top list

20, What is binary classification? What is a confusion matrix? (4)

- classification is an algorithmic procedure, which categorises an input data into classes. Binary: there are ~~are~~ two classes, e.g. true/false
- a binary decision can be described by a confusion matrix

properties: accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$

sensitivity = $\frac{TP}{TP+FN}$

specificity = $\frac{TN}{TN+FP}$

		true class	
		T	F
predicted class	T	TP true positive	FP false positive
	F	FN false negative	TN true negative

21, Why is it difficult to write classifiers for proteins groups?

In what sense can protein groups be diff from each other?

- protein groups have small between group distances & large within group distances. The members of a group are very similar, but the diff. groups are weakly connected
- protein families share domains btw each other, domain similarities are weak.
- a prot. group can contain 1 - > 30 000 members, & the average sequence length varies too.
- prot. groups can be diff. in sense of function, 3D structure, conserved sequences

22, What databases use a, annotated clusters b, non-annotated

- can protein groups be diff from each other?
- protein groups have small between group distances & large within group distances. The members of a group are very similar, but the diff groups are weakly connected
- protein families share domains btw each other, domain similarities are weak.
- a prot. group can contain 1 - > 30,000 members, & the average sequence length varies too.
- prot. groups can be diff. in sense of function, 3D structure, conserved sequences

22, What databases use a, annotated clusters b, non-annotated or partly ann. clusters? (2)

a, COG, SBASE

b, Unif

23, Name 2 major application areas of NGS! (4)

- de novo genome assembly
- resequencing a known genome from diff individuals
- personalized medicine

24, What types of templates can be used for NGS & what are they good for?

fragment: < 1kbp fragments of DNA

mate-pair: circular DNA cleaved subsequently

25, What is color space in NGS? (4)

- each 'base pair' has a colour code, so if we have the starting nucleic acid, & a table with the assigned colours, we can get the sequence from a colour code

26, What part of NGS reads are least reliable in terms of base quality? (4)

- the ending parts