

Sequence Similarity

Andrew Torda, wintersemester 2009 / 2010, AST, Angewandte ...

- What is the easiest information to find about a protein ?
 - sequence
 - history - amino acid sequencing
 - today - DNA / mRNA sequencing
 - consequence
 - lots of sequences
 - want to find similar proteins
- Mission
 - similarity of sequences – ways to estimate

Similarity of sequences

- Problem

ACDEACDE . .

ADDEAQDE . .

- how similar ?

ACDQRSTSRQDCAEACDE . .

ADDQRSTSRQDCAEAQDE . .

- size counts - longer sequences are more similar

- probabilistically - more chances to mutate

- a measure of (di)similarity – evolutionary distance

Too Simple Estimate

- difference / distance
 - time t
- rate of mutation λ
- few mutations
 - $A \rightarrow C$ but not $A \rightarrow C \rightarrow A$ (OK ?) if $p(\text{mutation})$ small
- sequence length n_{res}
- number mutations n_{mut}
- $n_{mut} = t \lambda n_{res}$ SO $t = \frac{n_{mut}}{\lambda n_{res}}$
- too simple

Jukes – Cantor distance

Simplification

- work with 4 base types (like DNA)

Rules and nomenclature

- probability of a specific mutation $A \rightarrow C$ or $G \rightarrow C$
 - in time Δt is α
 - set $\alpha = \lambda/4$
- probability of a change from type A at time t is $p_{A,t}$
- probability of seeing type A at time 1 is $p_{A,1}$
- initial probability at time 0 is $p_{A,0} = 1$

Jukes – Cantor distance

- probability of change in $\Delta t = 3\alpha$
- probability of no change $p_{A,1} = 1 - 3\alpha$
- probability of $A \rightarrow ? \rightarrow A$ in Δt
 - $\alpha(1 - p_{A,t})$

Fear not - slower
detailed explanation in
Übung

- what is the probability of seeing type A at a time $t+1$?
 - (no change) + ($A \rightarrow ? \rightarrow A$)
 - $p_{A,t+1} = p_{A,t} (1 - 3\alpha) + \alpha(1 - p_{A,t})$
- what change has occurred in time Δt ?

$$\begin{aligned}\frac{\Delta p_{A,t}}{\Delta t} &= p_{A,t+1} - p_{A,t} \\ &= p_{A,t} (1 - 3\alpha) + \alpha(1 - p_{A,t}) - p_{A,t} \\ &= 4\alpha p_{A,t} - 3\alpha\end{aligned}$$

Jukes – Cantor distance

- $\frac{dp_{A,t}}{dt} = -4\alpha p_{A,t} + \alpha$
- we want an estimate of t
- like any differential equation

$$\frac{dt}{dp_{A,t}} = \frac{1}{-4\alpha p_{A,t} + \alpha}$$

$$t = \int \left(\frac{1}{-4\alpha p_{A,t} + \alpha} \right) dp_{A,t}$$

- Übung – derivation of Jukes-Cantor rates...

Jukes – Cantor distance

- from
$$t = \int \left(\frac{1}{-4\alpha p_{A,t} + \alpha} \right) dp_{A,t}$$

- we get
$$p_{no_change} = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t} \qquad p_{change} = \frac{3}{4} - \frac{3}{4} e^{-4\alpha t}$$

- but this is for one site
- important what fraction of sites has changed ?
- estimate time

$$\frac{n_{mut}}{n_{res}}$$

$$t \propto -\ln\left(1 - \frac{4}{3} p_{change}\right)$$

$$t \propto -\ln\left(1 - \frac{4}{3} \frac{n_{mut}}{n_{res}}\right)$$

Simplifications made

- We have only worried about relative distances
 - no attempt to speak of years
- What is time ?
 - generations
 - years
- 4 bases for DNA (easy to change to 20 amino acids)

Comments on

- base composition equal at $t = 0$
- a residue can mutate to any other
- gaps / alignment quality
- uniform mutation rates

- some details on these issues...

Base Composition

Not a problem

- think back to slide on integration - constant c
- solved by assuming $p_{A,0}=1$ but could be any value

Different kinds of mutations

- We assumed
 - $p_{XY} = \alpha$ for all XY types
- Wrong:
 - DNA: $A \rightarrow G$ not as bad as $A \rightarrow C$ or $A \rightarrow T$
 - proteins: some changes easy ($D \rightarrow E$) some hard ($D \rightarrow W$)

Different kinds of mutations

- can be fixed with more parameters
 - simple case DNA
 - rate α for purine \rightarrow purine, β for purine \rightarrow pyrimidine
 - protein:
 - 19 different probabilities (for each amino acid type)

Gaps

- so far ignored
- more generally
 - we have assumed proteins / DNA can be aligned

Gaps and Alignments

- gaps ignored
- more generally - assumption that sequences can be aligned

ACDQRSTSRQDCAEACDE . .

ADDQRSTSRQDCAEAQDE . .

- but what about

ACDQRATSRQDQRSTSRQ . .

ADDQRSTSRQDCAEAQDE . .

- or

ACDQRATSRQDQRSTSRQ . .

ADDQRSTSRQDCAEAQDE . .

- the more distant the sequences, the less reliable the alignment

Uniform mutation rates

- Between organisms
 - fruit flies have short generations
 - bacteria have very short generations
 - within one class of organisms rates vary (DNA repair)
- Neglect of
 - duplication, transposition, major re-arrangements
- Different proteins mutate at different rates
 - essential – DNA copying
 - less essential
 - copied proteins (haemoglobins)
- Functional changes
 - similar proteins in different organisms – different functions
- Within one protein
 - some sites conserved, some mutate fast
- Complete neglect of natural selection

Similarity of sequences so far

- For very related sequences, not many back mutations
 - even simple mutation count (n_{mut}/n_{res}) OK
- Better to allow for back mutations
- Jukes-Cantor (and related) models
 - can include some statistical properties (base composition)
 - can be easily improved to account for other properties (different types of mutation occur with different frequencies)
 - hard to calibrate in real years, but may not matter
 - will be less reliable for less related species / proteins

Statistical approach to similarity

- Completely different philosophy
- Are proteins A and B related ?
 - how is A related to all proteins (100 000's) ?
 - how strong is the AB relation compared to A-everything ?
- What we need
 - BLAST / fasta (more in Dr Willhoeft's lectures)
 - idea of distributions
 - measure of significance

Significance

e-value (expectation value)

- I have a bucket with 10 numbered balls (1 .. 10)
- I pull a ball from the bucket (and replace it afterwards)
- how often will I guess the correct number ?
 - *e*-value = 0.1
- you guess the number and are correct 0.25 of the time
 - much more than expected
 - what is the probability (*p*-value) of seeing this by chance ?
 - example distribution.. binomial

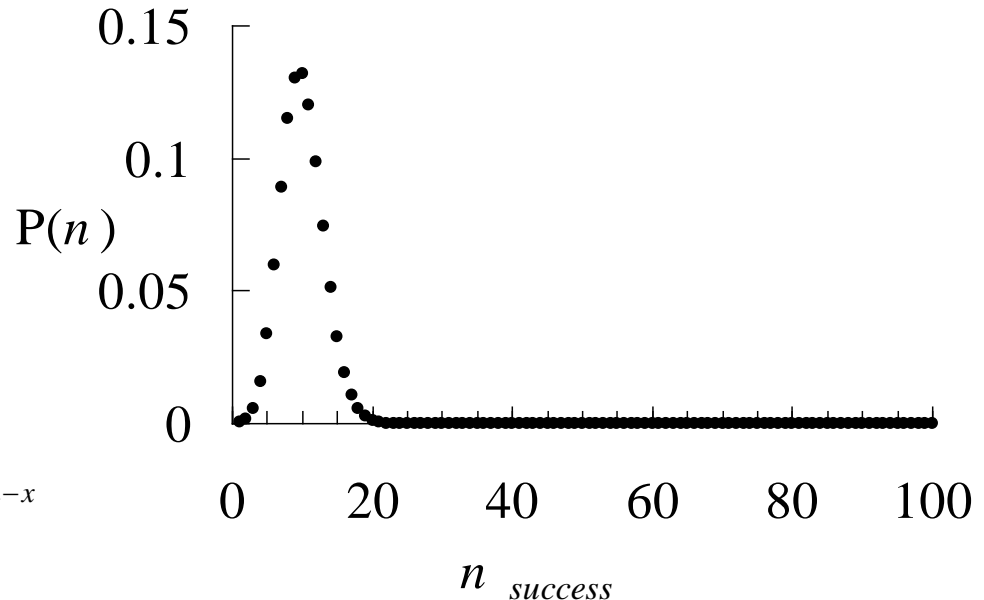
Binomial example

- we have 100 attempts ($n=100$)
- probability $p = 0.1$ of success on any attempt
- what is the probability that we are always wrong ?
 - $P(0) = 0.9 \times 0.9 \times 0.9 \dots = 2.7 \times 10^{-5}$
- probability that we make one correct guess
 - $P(1) = 0.1 \times 0.9 \times 0.9 \dots +$
 $0.9 \times 0.1 \times 0.9 \dots +$
 $0.9 \times 0.9 \times 0.1 \dots + \dots = 3.0 \times 10^{-4}$
 - $P(25) = 9.0 \times 10^{-6}$ my original question
 - $P(10) = 0.13$ what you would guess

Binomial example

- probability that we make one correct guess
 - $P(1) = 0.1 \times 0.9 \times 0.9 \dots +$
 $0.9 \times 0.1 \times 0.9 \dots +$
 $0.9 \times 0.9 \times 0.1 \dots + \dots = 3.0 \times 10^{-4}$
 - $P(25) = 9.0 \times 10^{-6}$ my original question
 - $P(10) = 0.13$ what you would guess

- this formula not for exams



formally

x number of success

n number trials

p probability per trial

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Distributions and sequences

- If I align two proteins, sometimes they will be similar (by chance)
- Take a protein and align to a large database
 - there will be a distribution of scores

```

20 354 0:=====
22 16 0:=
24 16 0:=
26 34 0:==
28 91 4:*====
30 130 22:*=====
32 216 85:====*=====
34 351 229:====*=====
36 484 471:====*=====
38 729 779:====*=====
40 821 1086:====*=====
42 1049 1328:====*=====
44 1156 1465:====*=====
46 1272 1492:====*=====
48 1237 1428:====*=====
50 1220 1303:====*=====
52 1227 1146:====*=====
54 1094 979:====*=====
56 929 817:====*=====
58 824 671:====*=====
60 655 544:====*=====
62 494 436:====*=====
64 390 347:====*=====
66 276 274:====*=====
68 239 216:====*=====
70 176 169:====*=====
72 124 132:====*=====
74 76 103:====*=====
76 60 80:====*=====
78 44 62:====*=====
80 46 48:====*=====
82 25 37:====*=====
84 15 29:====*=====
86 3 23:====*=====
88 5 18:*
90 5 14:*
92 3 10:*
94 4 8:*
96 0 6:*
98 0 5:*
100 1 4:*
102 0 3:*
104 0 2:*
106 0 2:*
108 0 1:*
110 0 1:*
112 0 1:*
114 0 1:*
116 0 0:
118 0 0:
>120 0 0:
    
```

one = represents 22 library sequences

inset = represents 1 library sequences

very few are radically different

many sequences match a bit

these ones are probably related

Two Distance Measures

One question

- what is the similarity of two sequences ?

Two answers

1. Given two sequences

* estimate evolutionary t

2. Given two sequences (one is in database)

* estimate whether they are really related

- when are they used ?

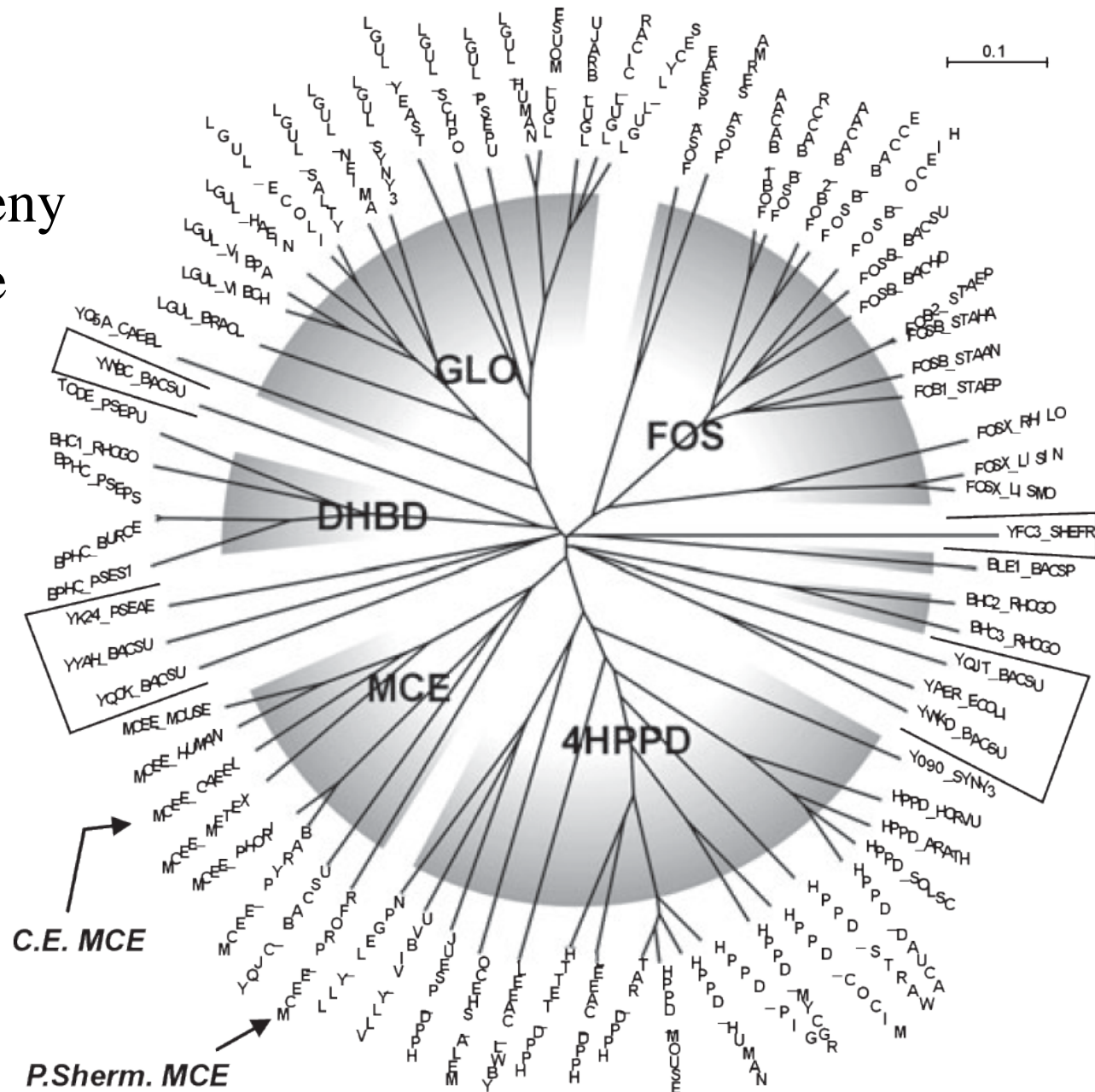
Two Distance Measures

Common uses

- Collection of sequences and want a phylogenetic tree ..
 - each sequence has mutated from another
 - use a measure like Jukes-Cantor
- One sequence
 - which are possibly related sequences ?
 - rank the similarities

An example phylogeny

- metabolic enzyme from a set of parasites



Two Distance Measures

- Collection of sequences and want a phylogenetic tree ..
 - each sequence has mutated from another
 - use a measure like Jukes-Cantor
- One sequence
 - which are possibly related sequences ?
 - rank the similarities
- model types ...

Model types

Connection to first lecture

- statistical approach
 - very little biology – sequences are objects + distribution
- Jukes - Cantor
 - problem-specific model (mutations, probabilities...)
- next topic – using these similarities - clustering