Sequence Similarity

Andrew Torda, wintersemester 2010 / 2011, 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

Plan

- Last week
  - different ways to model systems

  1. Problem specific (grundlages Modell)
  2. generally applicable (statistical, machine learning..)

- today
  - two radically different ways to estimate protein similarity

Similarity of sequences

- Problem

  A**C**DEA**C**DE..

  A**D**DEA**Q**DE..

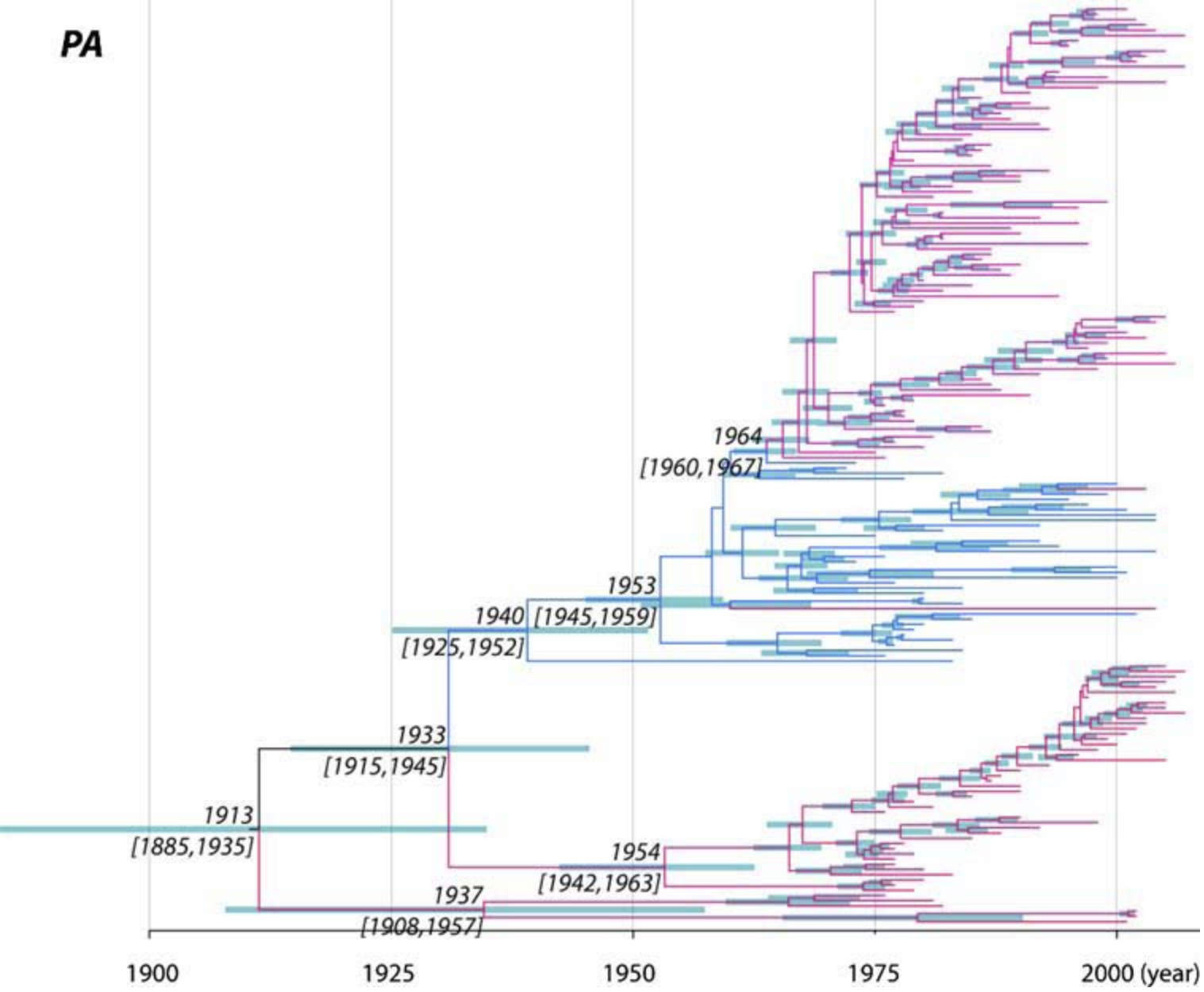
- how similar ?

  A**C**DQRSTSRQDCAEA**C**DE..

  A**D**DQRSTSRQDCAEA**Q**DE..

- size counts - longer sequences are more similar
  - probabilistically - more chances to mutate

- a measure of (di)similarity – evolutionary distance
- units
  - years, generations (t)
- who cares ?



- avian flu tree

- Chen R. and Holmes, E.C., J Mol Evol (2010) 70,98–105

Naïve Estimate

- Mutations are rare (bad for you)
- 1 mutation takes t, 2 mutations takes $2t$
- difference / distance
  - time t
- rate of mutation λ
- sequence length n_{res}
- number mutations n_{mut}
- $n_{mut} = t\,\lambda\,n_{res}$ so $\quad t = \dfrac{n_{mut}}{\lambda n_{res}}$
- not very good

naïve estimate

- Zero mutations
  - A→A or A→C →A ?
- What if I see A→C
  - was it A→G →C ?
- how to account for this… Jukes-Cantor model (or others)
- the answer.. $t \propto -\ln\left(1 - \frac{4}{3}\frac{n_{mut}}{n_{res}}\right)$ → similarity measure
- where does this come from ? Overview..

Jukes-Cantor distance

- for DNA, 4 base types
- Assumptions /definitions
  - probability of a specific mutation in time Δt is α
    - A→C or G→C or …
  - probability of seeing type A at time t is $p_{A,t}$
  - probability that you were not A is $(1 - p_{A,t})$
- aim – differential form, $p_{A,t}$

- probability of change in $\Delta t = 3\alpha$
- probability of no change $1 - 3\alpha$

- probability of $? \rightarrow \mathsf{A}$ in Δt
  - $\alpha(1 - p_{\mathsf{A},t})$
- probability of $\mathsf{A} \rightarrow \mathsf{A}$
  - $p_{\mathsf{A},t}\ (1 - 3\alpha)$
- probability $p_{\mathsf{A},t+\Delta t}$ of seeing A at $t + \Delta t$
  - $\alpha(1 - p_{\mathsf{A},t}) + p_{\mathsf{A},t}\ (1 - 3\alpha)$
- change in probability Δp in $\Delta t = p_{\mathsf{A},t+\Delta} - p_{\mathsf{A},t}$

$$\frac{\Delta p_{\mathrm{A},t}}{\Delta t} = p_{\mathrm{A},t+1} - p_{\mathrm{A},t}$$
$$= p_{\mathrm{A},t}(1-3\alpha) + \alpha(1-p_{\mathrm{A},t}) - p_{\mathrm{A},t}$$
$$= -4\alpha p_{\mathrm{A},t} + \alpha$$

- write in continuous form

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

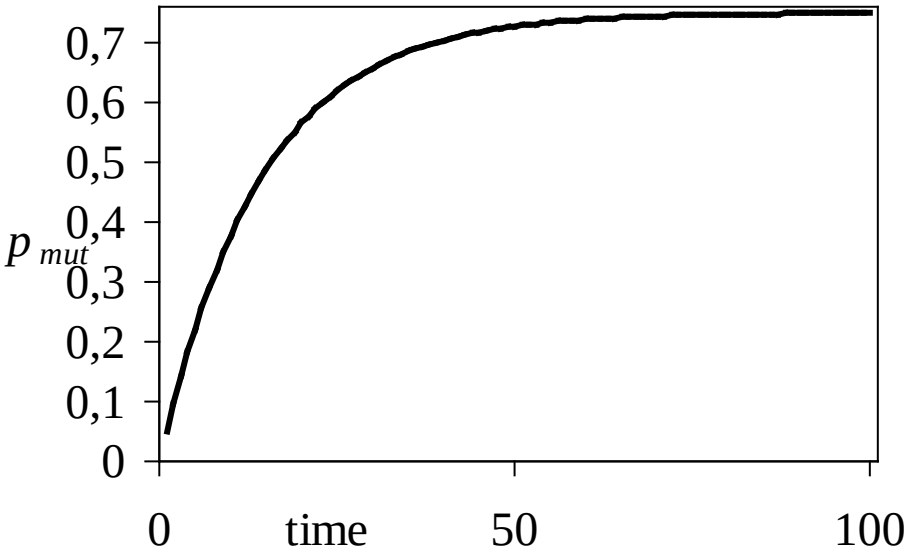
$$t = \int \left(\frac{1}{-4\alpha p_{\mathrm{A},t} + \alpha} \right) dp_{\mathrm{A},t}$$

Jukes-Cantor – final estimate

- account for 4 types of base
- get out constant of integration
- forget the constants at the front

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

- final result…



Evolutionary distance

- Given two sequences
  - can work out the time (distance) between them
  - we do account for A→C →A

- a similarity measure based on a specific model !
  - start of story
- not a perfect model
  - not all mutations are equally likely – more later
  - what happens at $t \rightarrow \infty$?

- statistical approaches to similarity..

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 $p_{\mathtt{A},0} = 1$ but could be any value

Different kinds of mutations

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

  A**C**DQRSTSRQDCAEA**C**DE..

  A**D**DQRSTSRQDCAEA**Q**DE..

- 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

Uniform mutation rates

- Different proteins mutate at different rates
  - essential – DNA copying
  - less essential
  - copied proteins (haemoglobins)

- Within one protein
  - some sites conserved, some mutate fast
- Complete neglect of selection pressure
  - clear ? save for Übungzeit

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 Gonella'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 ?

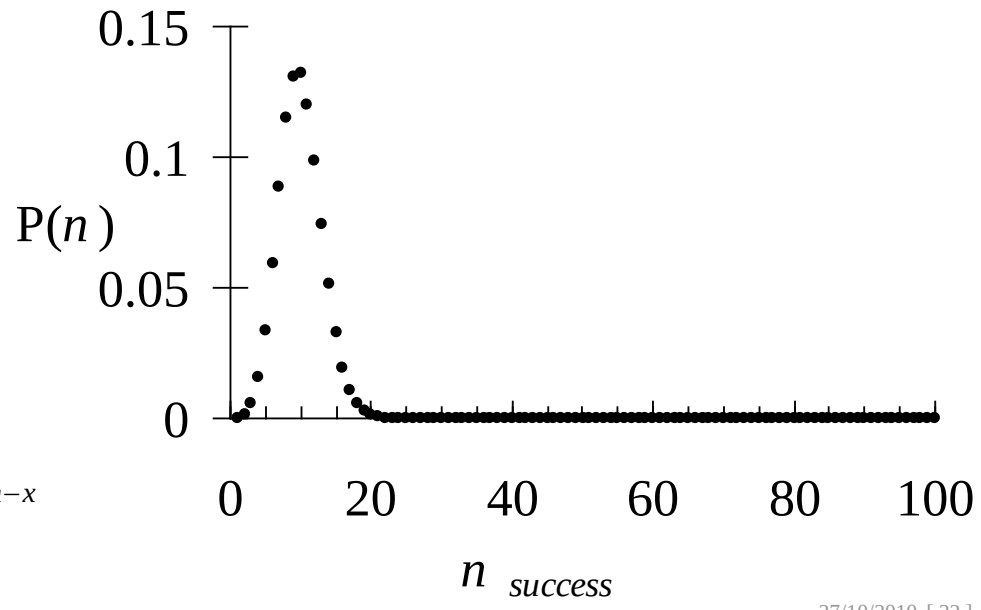
  - $P(0) = 0.9 \times 0.9 \times 0.9 \ldots = 2.7 \times 10^{-5}$

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

Binomial example

- probability that we make one correct guess
  - $P(1) = 0.1 \times 0.9 \times 0.9 \ldots +$
    $0.9 \times 0.1 \times 0.9 \ldots +$
    $0.9 \times 0.9 \times 0.1 \ldots + \ldots = 3.0 \times 10^{-4}$
  - $P(25) = 9.0 \times 10^{-6}$ 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

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

Connection to sequences

- Align two sequences
  - some bases / residues will be the same by chance
  - some will be meaningful

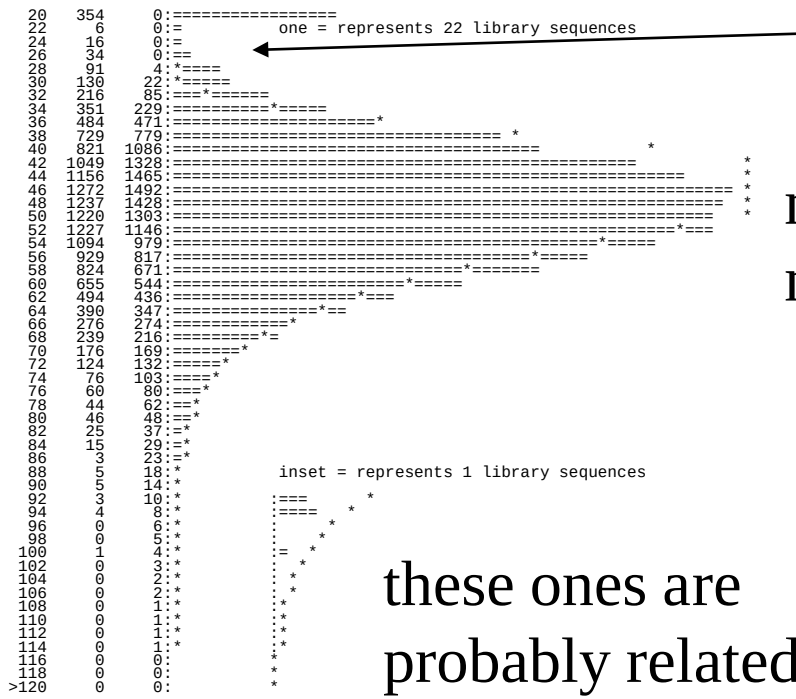
- pair 1 pair 2

ACGTACGT **ACGTACGT**

ACTTACGT **ACTTGCGT**

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



very few are radically different

many sequences match a bit

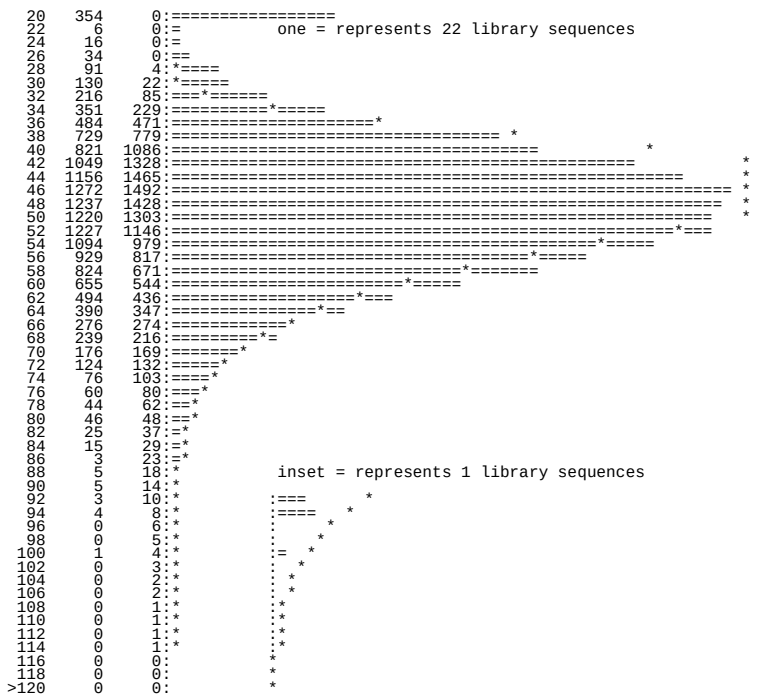
these ones are probably related

Distributions and sequences

- Can we put numbers on this ?
- model for the distribution
  - "extreme value distribution"
- Probability of score S≥ x

$$P(S \geq x) = 1 - \exp\left(-kMNe^{-\lambda x}\right)$$

- NM reflect sequence length



- one method
  - estimate λ and k for each sequence
- alternative
  - use a recipe and precalculate λ and k

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

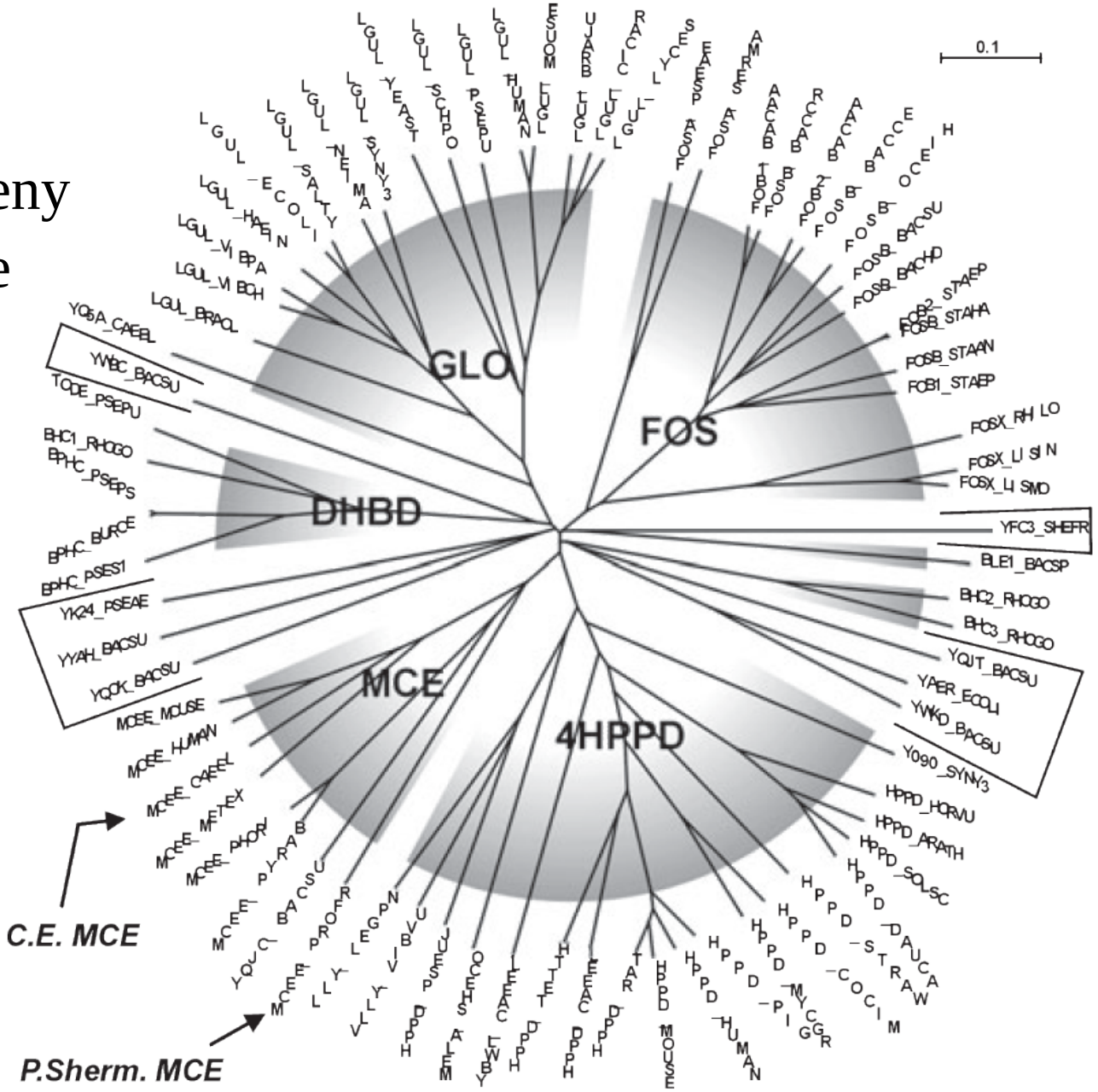
Common uses

- 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



Kühnl… Liebau, FEBS Journal 272 (2005) 1465–1477

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