Introduction / Modelling

Andrew Torda, wintersemester 2010 / 2011, AST, Angewandte …

- who am I ?
- language .. Deutsch / English .. verhandelbar
- Zettel
  - **`www.bioinformatics.uni-hamburg.de/research/BM/torda/lehre.html`**
- + stine
- Übungen ebenfalls im web
- heute ? nicht 90 Minuten

Administration

People

- Andrew Torda 1. Stock / 105
  schade@zbh.uni-hamburg.de
  sekr (Annette Schade) 42838 7330

- Marco Matthies
- Thomas Margraf

Vorlesungen	Mi	16:30 – 18:00
Übungen	Mo	16:30 – 18:00

Homework / Übungen

Not too much

- enough from other courses

Übungen

- very short report (schriftlich)
- individuelle / eigene

Textbooks

- any biochemistry book (Stryer, Biochemistry as per chem dept)
  - expensive, not used too much
- Leach, Andrew, “Molecular Modelling” very good for future semesters
- Folien should be sufficient

Exams

- any facts that are mentioned in these lectures and Übungen
- schriftliche Klausur

Ausgleichung

- Informatikers – familiarity with proteins /chemistry /structure
- MLS/Chemiker/.. - some scripting, linux command line

	informatiker	MLS/Chemiker/..
Mi 27. Nov	organization	
Mo. 1. Nov	Protein struct 1	Linux command line / Scripting 1
Mo. 8. Nov	Protein struct 2	Linux command line / Scripting 2

Lecture Plans

1	18. Okt. 09	Models
2	30. Okt. 09	Similarity - protein sequences
3	6. Nov. 09	Cluster Analysis
4	13. Nov. 09	Secondary structure prediction
5	20. Nov. 09	Secondary structure prediction
6	27. Nov. 09	Protein domains
7	4. Dez. 09	Protein domains
8	11. Dez. 09	Protein function prediction
9	18. Dez. 09	Protein function prediction
10	8. Jan. 10	Protein function prediction
11	15. Jan. 10	Sequence design
12	22. Jan. 10	Sequence design
13	29. Jan. 10	Fold recognition
14	5. Feb. 10	Fold recognition

Themes - Applications

Theme of Semester

- given some information about a macromolecule (protein)
  - what can be calculated ? predicted ?
  - how much would you trust predictions ?
    - limitation, applicability, reliability

- typical information
  - a protein sequence (lots known)
  - a protein structure (less known)
  - a DNA sequence (think of genomes)

Today

- meaning of modelling
- similarity is not easy

Specific and general models

Dream

- Feed data to box and have it interpreted
  - given my protein, what is the structure ?
  - given my spectrum where is the centre of the peak ?

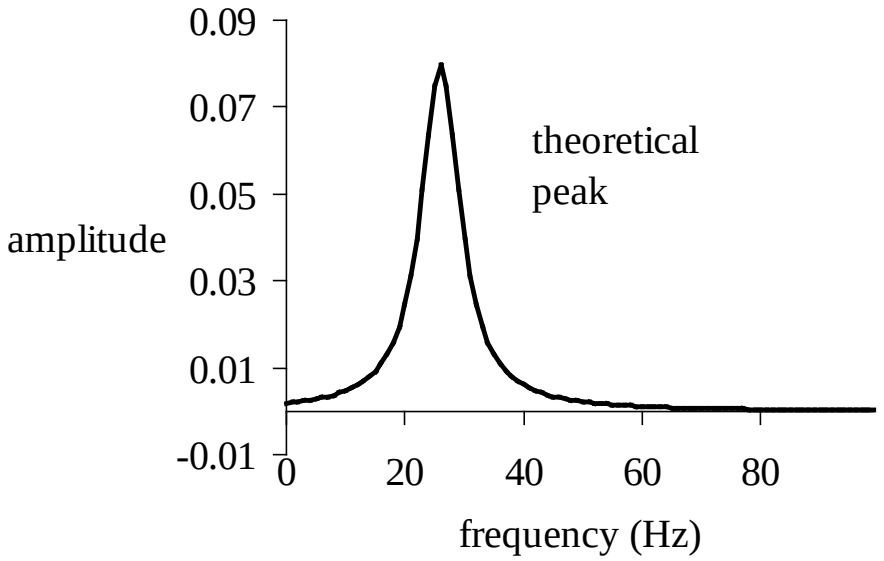
Model types

- Specific
  - you know the structure of your data, fit points to the observations
- General
  - look for some patterns in data – little understanding of the underlying theory

- examples

Interpreting spectroscopic data

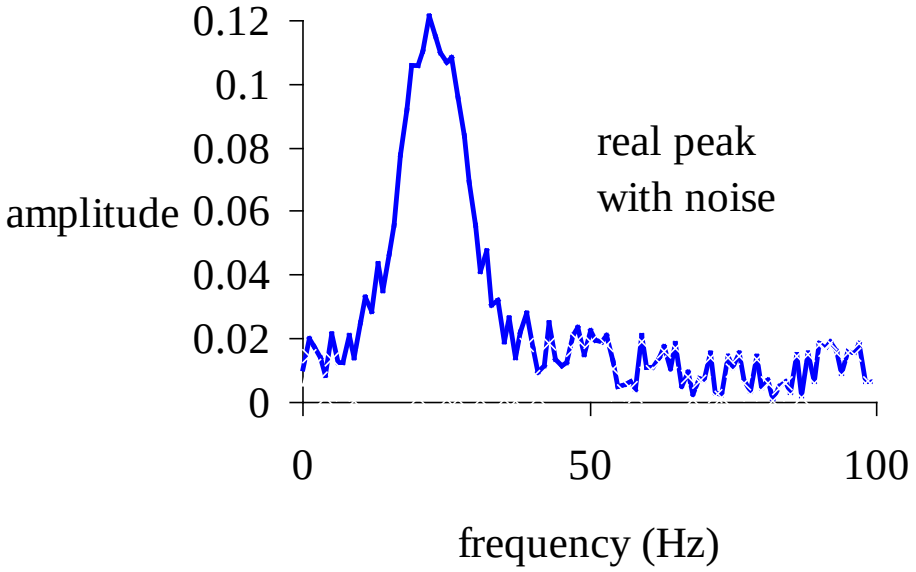
- just an example (no spectroscopy in this course)
- many kinds of peaks in spectroscopy look like



- my mission
- find centre (≈24) and height (≈0.08)
- but they have noise

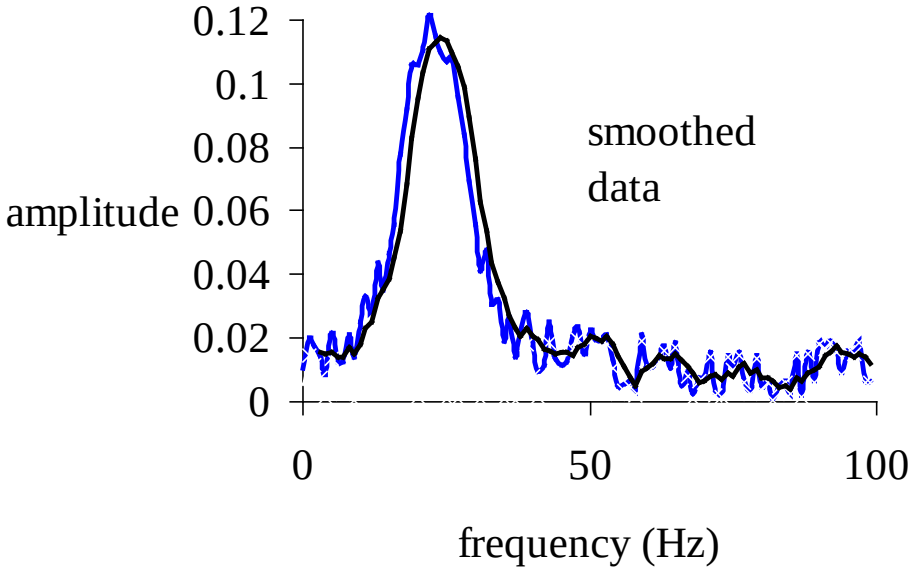
noisy data

- real world has noise
  - we still want centre, height



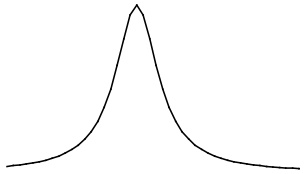
- try simple smoothing
  - no assumptions about data

- claim
  - centre around 23
  - looks believable



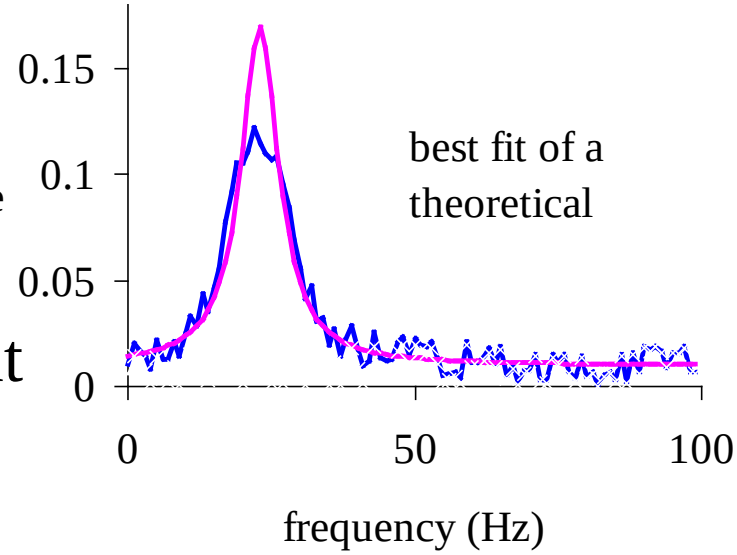
Using prior knowledge

- I expect peaks like $\frac{a^2}{(a^2 + x^2)}$

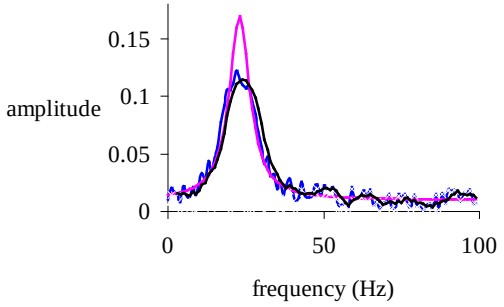


- A fit of a calculated peak…
  - something is clearly wrong
    - if peak has a certain width it must have an appropriate height

$$\frac{a^2}{(a^2 + x^2)}$$

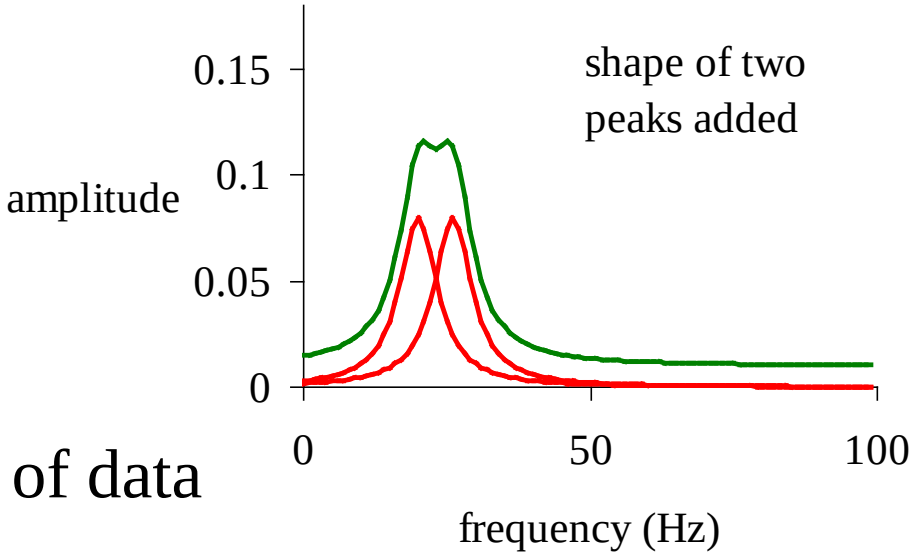


- What looked good is not the correct form

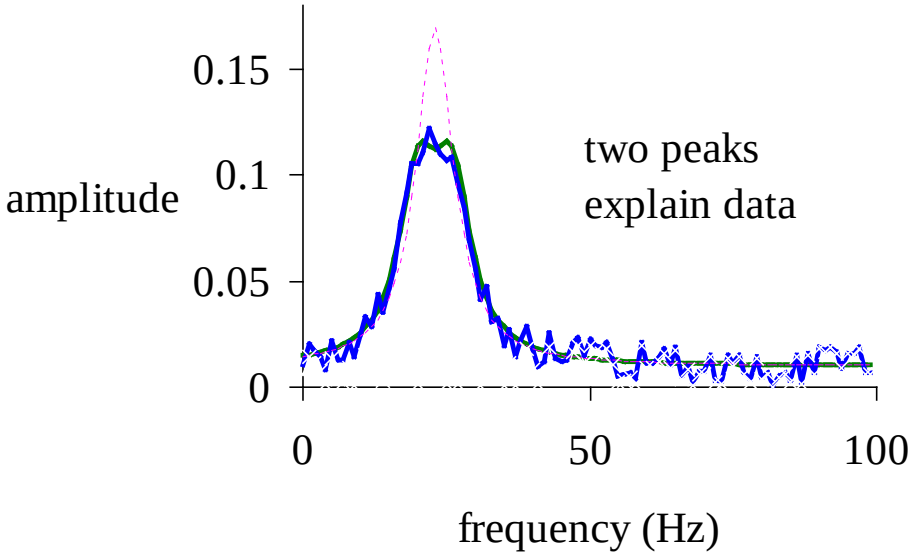


More appropriate fitting

- what if we used two peaks ?

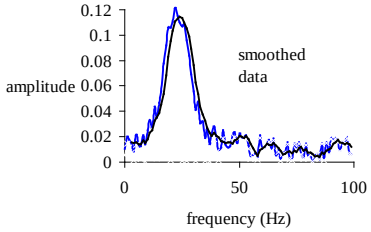


- peaks centred at 20 and 26
  - very different explanation of data

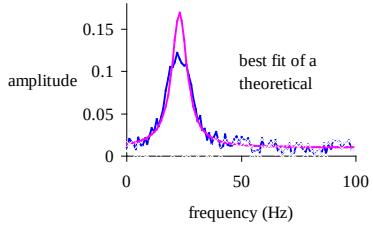


General vs appropriate modelling

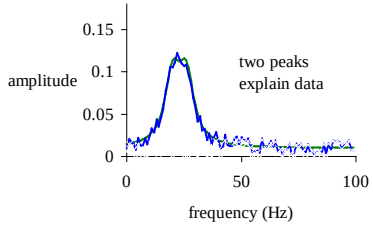
- general smoothing method suggested one peak
  - looks good
  - appears to explain observations
  - generally applicable



- testing with correct model suggested this is wrong



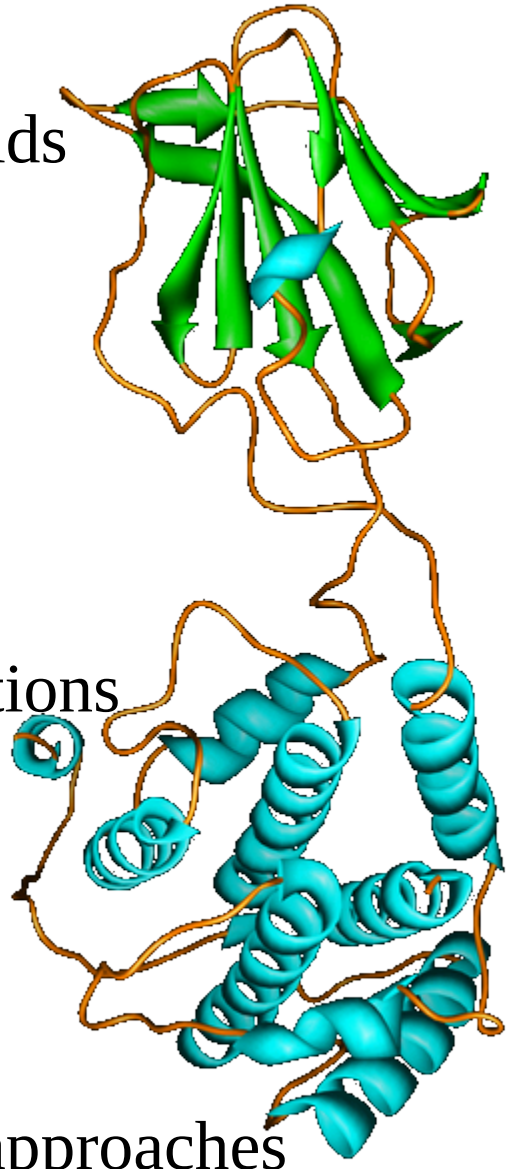
- fitting with best model (two peaks)
  - near perfect



- summary
  - if you know the underlying model, use it
  - always applicable ?
  - back to biological questions

General purpose modelling

- Proteins have "secondary structure
- It appears to reflect the sequence of amino acids
  - what is the rule ?
  - 20 amino acids, N positions,
    - 20^N sequences, patterns not clear



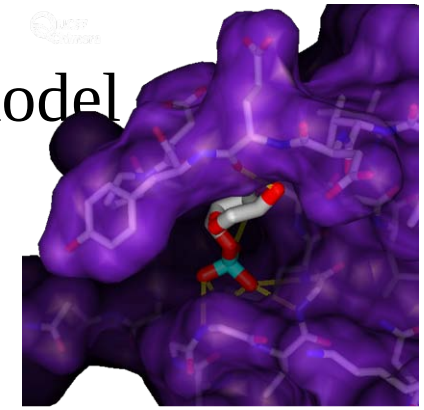
- what to do ?
  - correct model – think of all atomic interactions
    - see where atoms should be placed
      - not practical
  - or
  - forget physics
    - use dumb statistics / machine learning approaches

http://www.cgl.ucsf.edu/chimera/ImageGallery/, bluetongue virus capsid protein

Mixtures of specific and general

Will a ligand (Wirkstoff) bind to a protein ?

- with physics
  - model all atomic interactions, best physical model
  - calculate free energy (ΔG)
    - difference in solution / bound
- more generally
  - gather idea of important terms (H-bonds, overlap, ..)
  - try to find some function which often works
  - do not stick to real physics

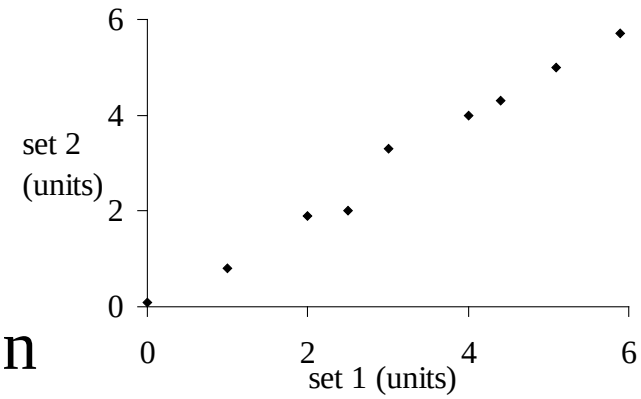


Will my drug dissolve in water or oil (lipid) ? (important)

- sounds like chemistry
  - usually approached by machine learning
    - number of atoms, types of atoms, …

http://www.cgl.ucsf.edu/chimera/ImageGallery/, bluetongue virus capsid protein

Similarity



- Important in all bioinformatics
  - I have a protein of unknown
    - structure / function / cell localisation
  - is it similar to one of known structure, function …
- Similarity seems obvious
  - two sets of numbers (above)
  - two protein sequences

    ACDEACDE rather similar - but quantified ?

    ADDEAQDE
    - how many positions differ ? how long are proteins ?
    - could the similarity be by chance ?

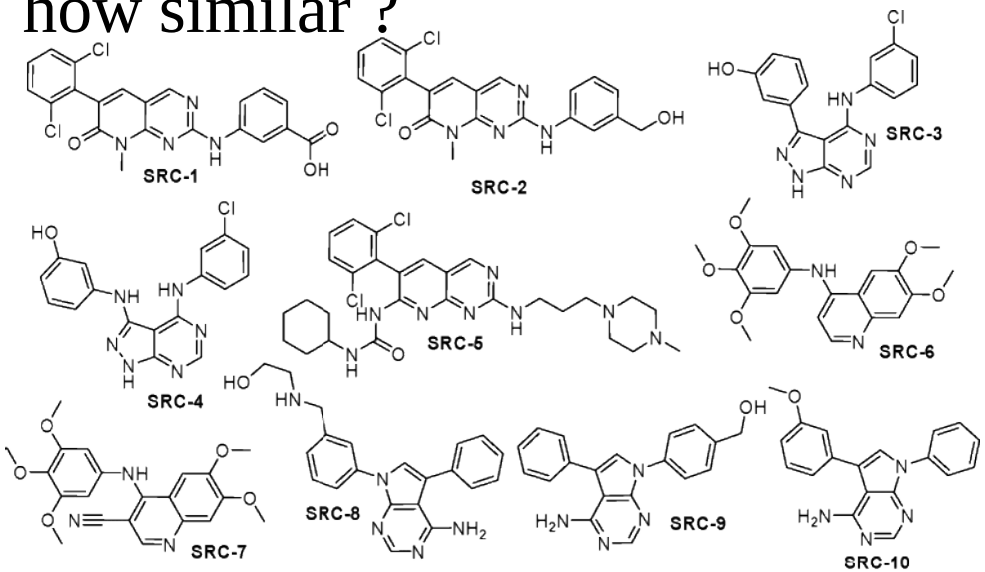
Similarity

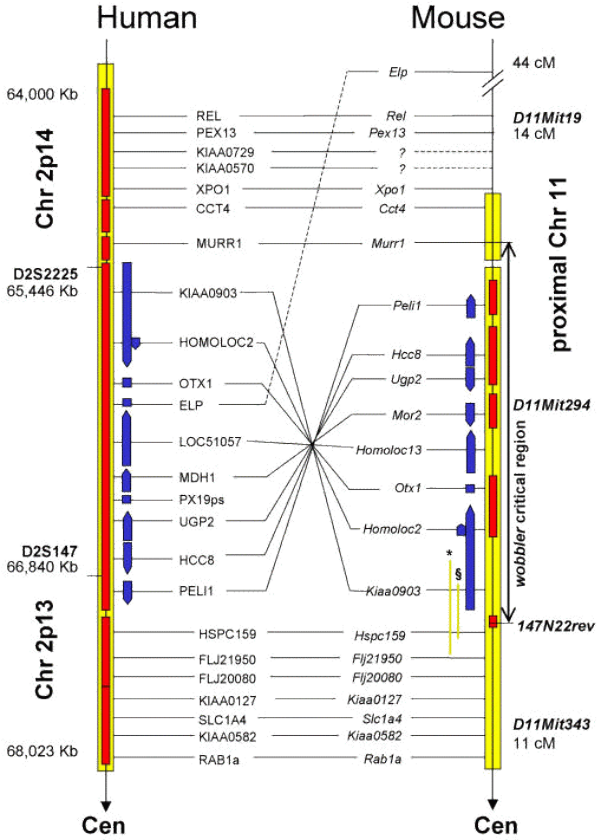
Two genomes similarity

- what are the descriptors ?
- how many genes are common ?
- is the order preserved ?

Potential drugs

- drug 1 binds, will drug 2 ?
- how similar ?





synteny plot: http://home.cc.umanitoba.ca/~umlawda/39.769/presentation/presentation.html, Fristensky, B.
ligands from, Wang, N., DeLisle, R. K. and Diller, D.J. (2005), J. Med. Chem., 48, 6980-6990

Detection and Quantification

- Models for prediction and interpretation
  - often not well justified

- Similarity in these applications
  - detection (finding / recognising)
  - quantification

- Each in the context of applications

- first protein structure …

Summary so far

A model can explain observations, make predictions or both

A model may be based

- on a belief of the underlying chemistry / physics
- purely mathematical, probabilistic

Similarity

- we have objects with some information (proteins, ligands, genomes, sequences, …)
- we want to find similar objects and hope they have the same properties
- simlarity has a different meaning in different areas

STOP

Sequence Similarity

- 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

  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

Too Simple Estimate

- difference / distance
  - time t
- rate of mutation λ
- few mutations
  - A→C but not A→C →A (OK ?) if p(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→C or G→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→ ? →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→ ? →A)
  - $p_{A,t+1} = p_{A,t}(1 - 3\alpha) + \alpha(1 - p_{A,t})$
- what change has occurred in time Δt ?

$$\frac{\Delta p_{A,t}}{\Delta t} = p_{A,t+1} - p_{A,t}$$

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

$$= 4\alpha p_{A,t} + \alpha$$

Jukes – Cantor distance

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

- we want an estimate of t
- like any differential equation

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

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

- Übung – derivation of Jukes-Cantor rates…

Jukes – Cantor distance

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

- we get $p_{no\_change} = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}$ $p_{change} = \frac{3}{4} - \frac{3}{4} e^{-4\alpha t}$

- but this is for one site
- important what fraction of sites has changed ? $\frac{n_{mut}}{n_{res}}$
- estimate time $t \propto -\ln\left(1 - \frac{4}{3} p_{change}\right)$

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

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 $p_{\mathtt{A},0}=1$ 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
- 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 selection pressure

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 ?

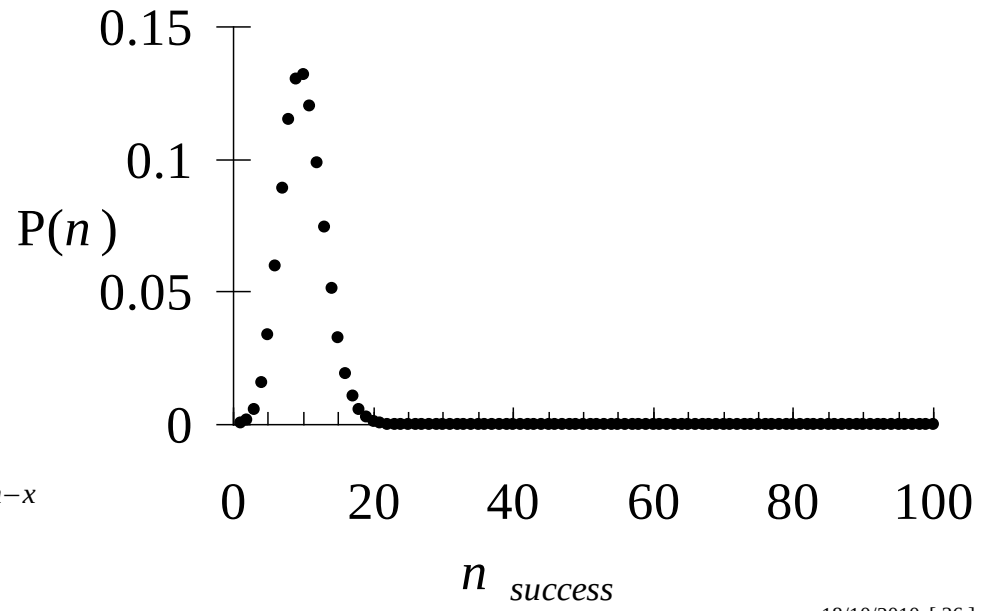
  - $P(0) = 0.9 \times 0.9 \times 0.9 \ldots = 2.7 \times 10^{-5}$

- 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



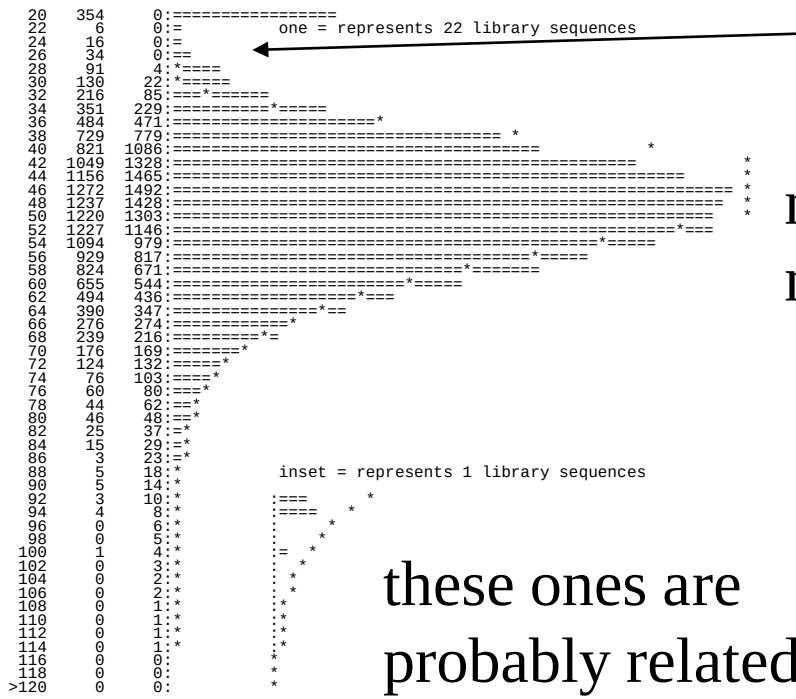
formally

x number of success
n number trials
p probability per trial

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

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

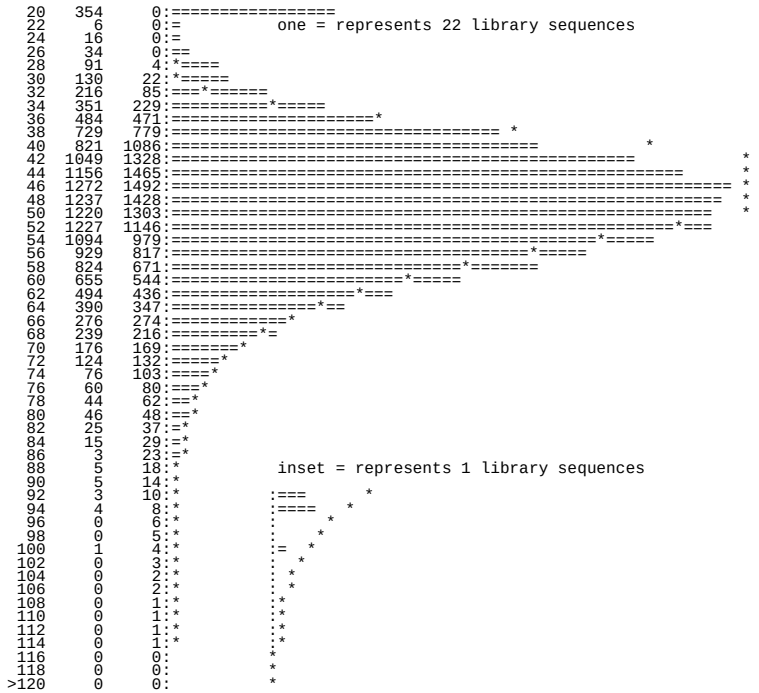
these ones are probably related

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

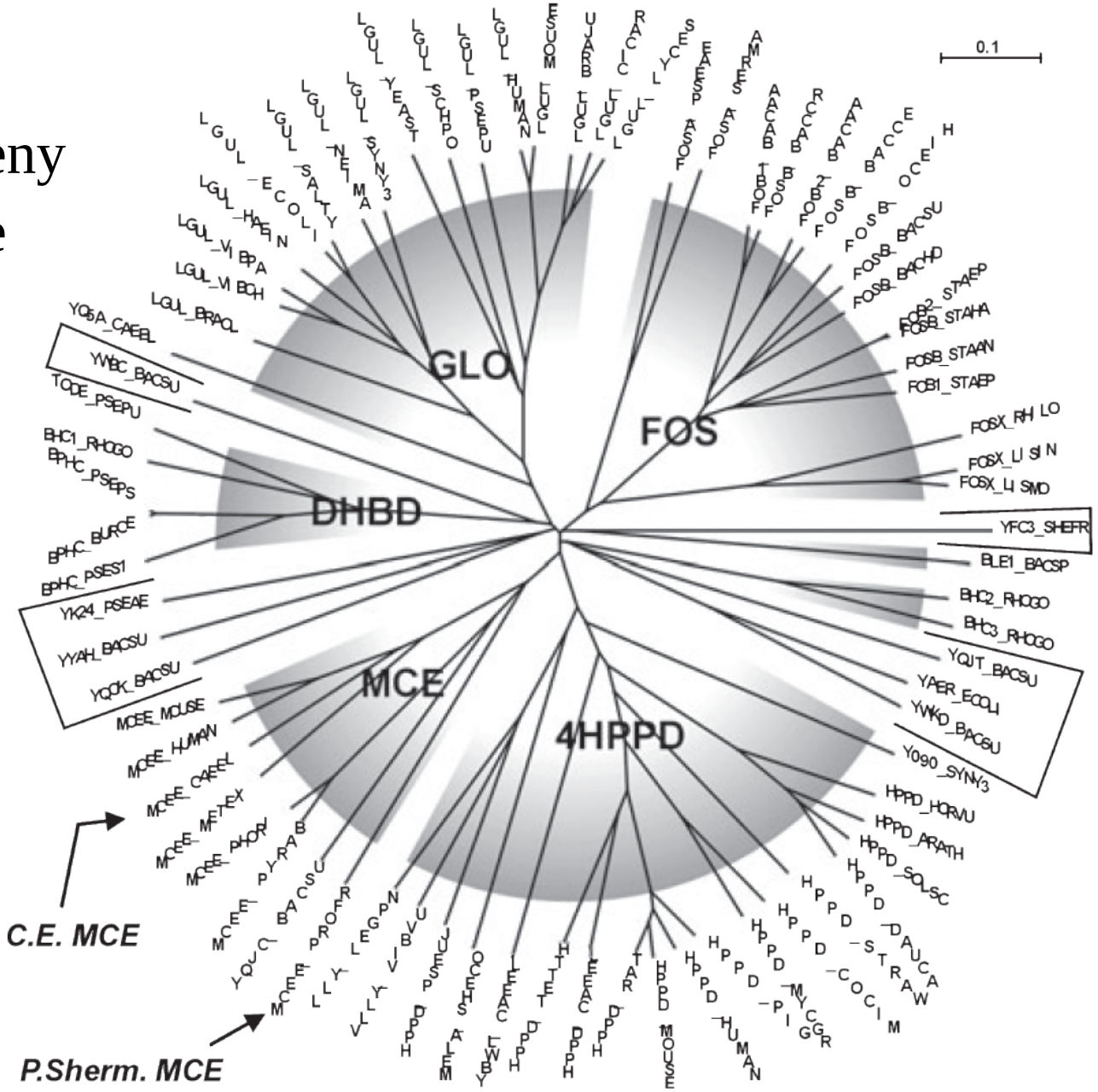
Common uses

- 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



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