Introduction

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

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

Administration

People

- Andrew Torda 42838 7331 1. Stock / 105
  schade@zbh.uni-hamburg.de
  sekr (Annette Schade) 7330

- Gundolf Schenk
- Marco Matthies
- Thomas Margraf
- Stefan Bienert (more in RNA)

Vorlesungen	Freittag	13:15 – 14:45
Übungen	Montag	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

Protein Structure - the problem - sociological

- Easy ? boring ?
- Essential
- How many people have done biology ? chemistry ?

- Mein Vorschlag
  - Ich nutze die Übungszeit für Strukturgrundlagen
  - Donnerstag morgens

okt 26	basic proteins 1	people who have not done protein structure
nov 2	basic proteins 2	
nov 9	Jukes-Cantor model derivation	everybody

1, 23. 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

Broad themes

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)

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 ?

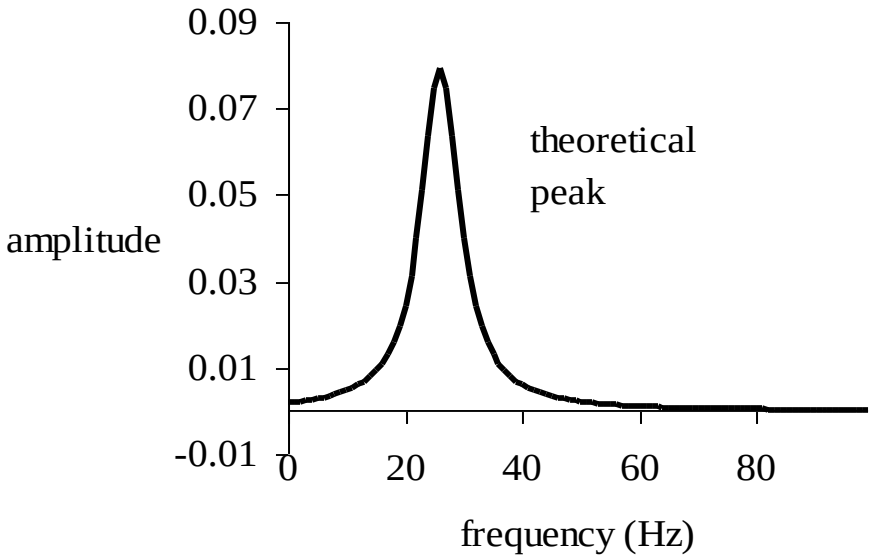
Model types

- 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

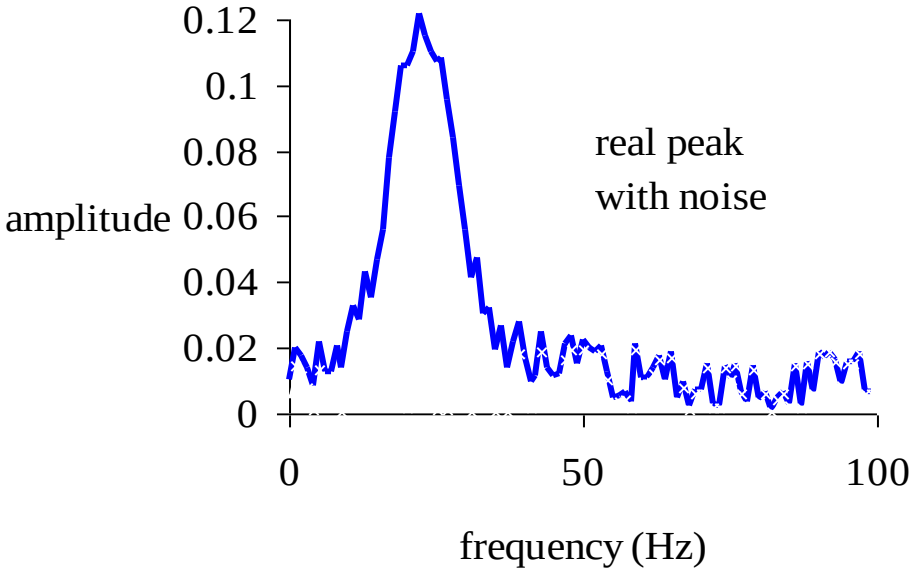
- just an example (no spectroscopy in this course)
- many kinds of peaks in spectroscopy look like



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

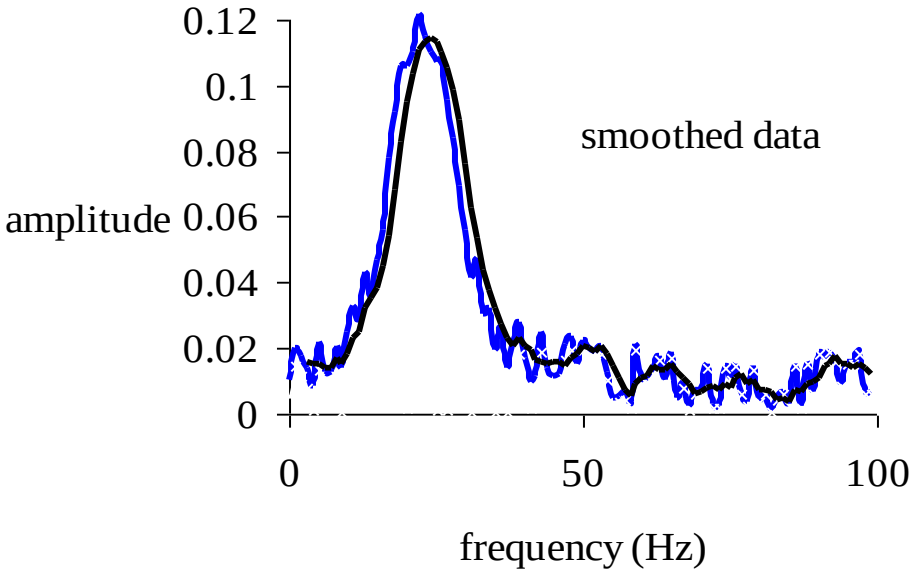
noisy data

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



- try simple smoothing
  - no assumptions about data

- claim
  - centre around 23
  - looks believable

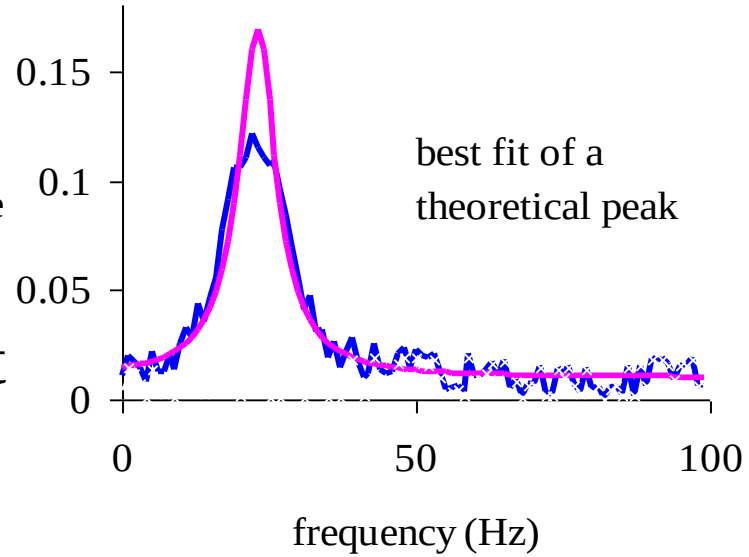


Using prior knowledge

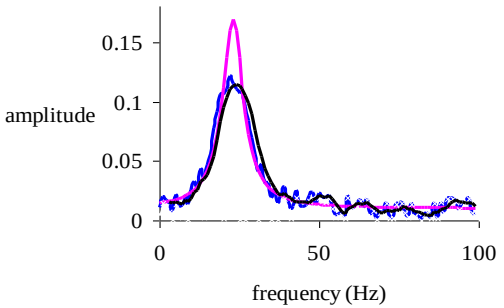
- I expect peaks like $\frac{a^2}{(a^2 + x^2)}$

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

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

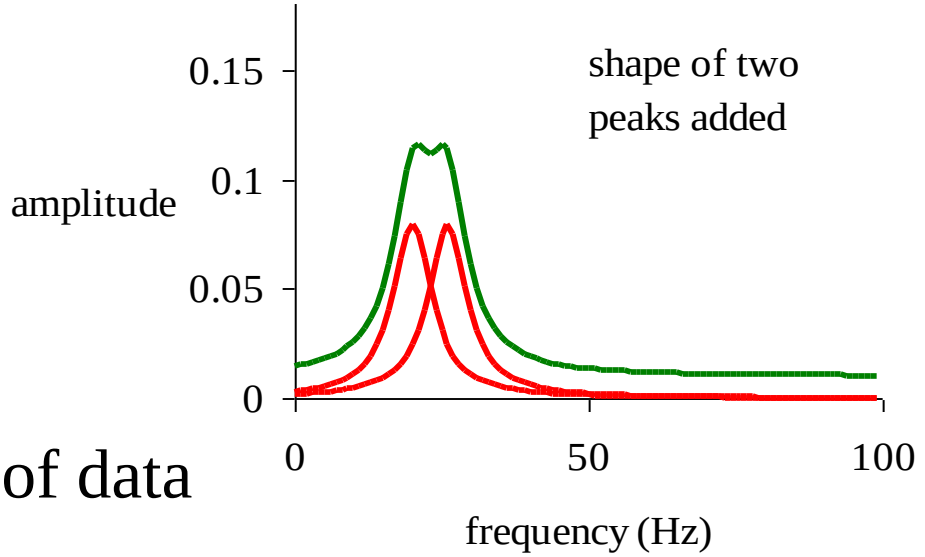


- What looked good is not the correct form

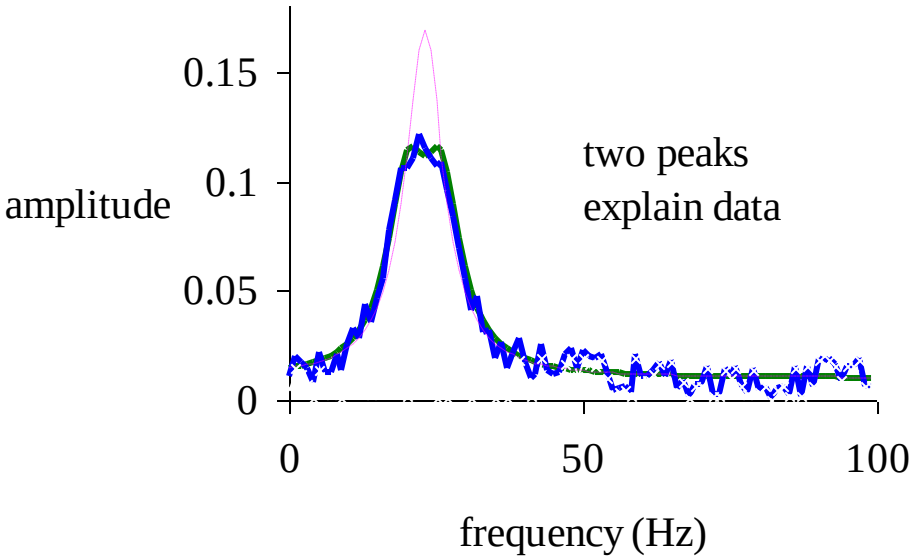


More appropriate fitting

- what if we used two peaks ?

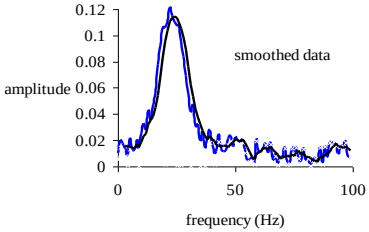


- peaks centred at 20 and 26
  - very different explanation of data

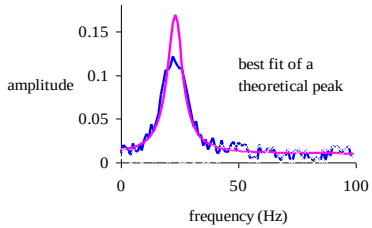


General vs appropriate modelling

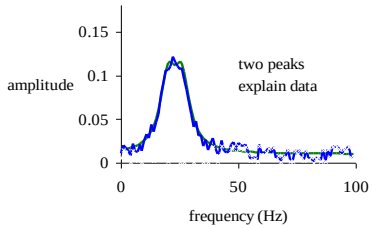
- general smoothing method suggested one peak
  - looks good
  - appears to explain observations
  - generally applicable



- testing with correct model suggested this is wrong



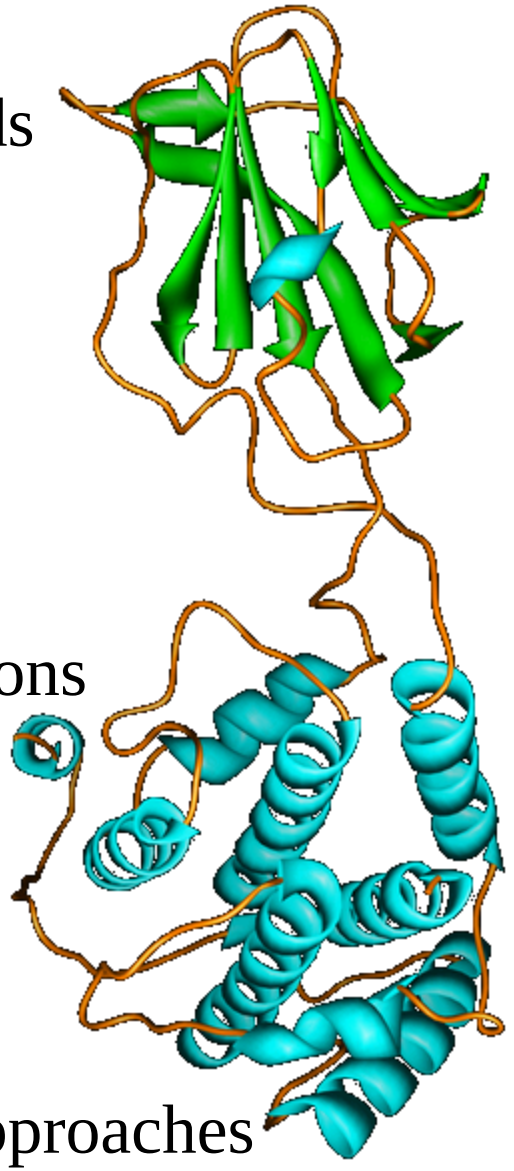
- fitting with best model (two peaks)
  - near perfect



- 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



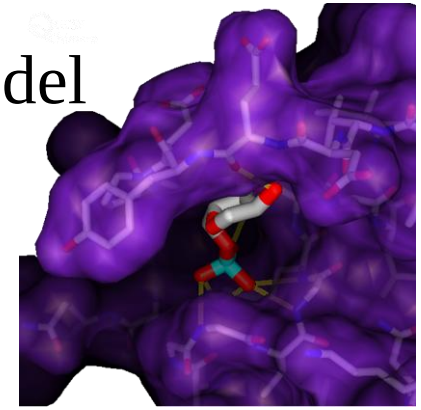
- what to do ?
  - correct model – think of all atomic interactions
    - see where atoms should be placed
      - not practical
  - or
  - forget physics
    - use dumb statistics / machine learning approaches

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

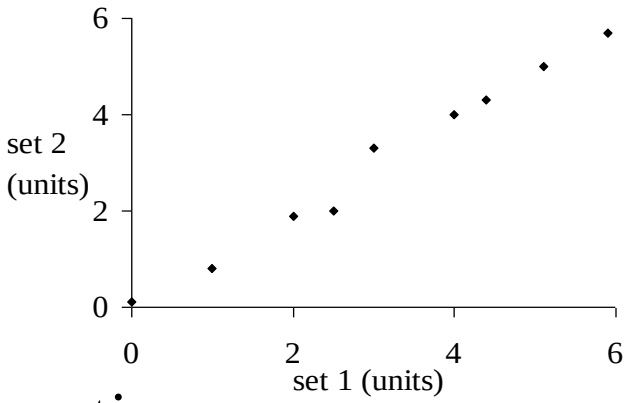


Will my drug dissolve in water or oil (lipid) ? (important)

- 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 ?

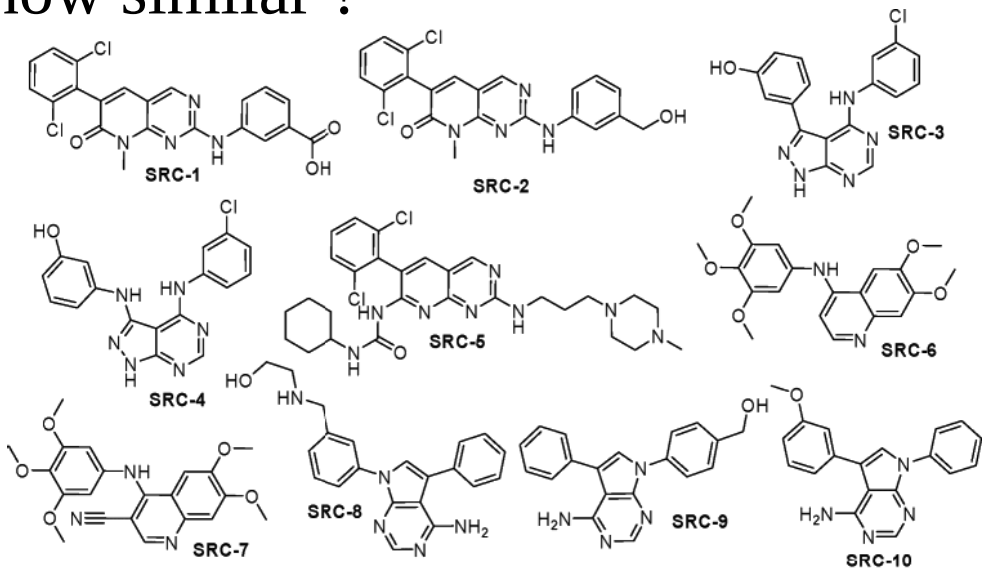
Similarity

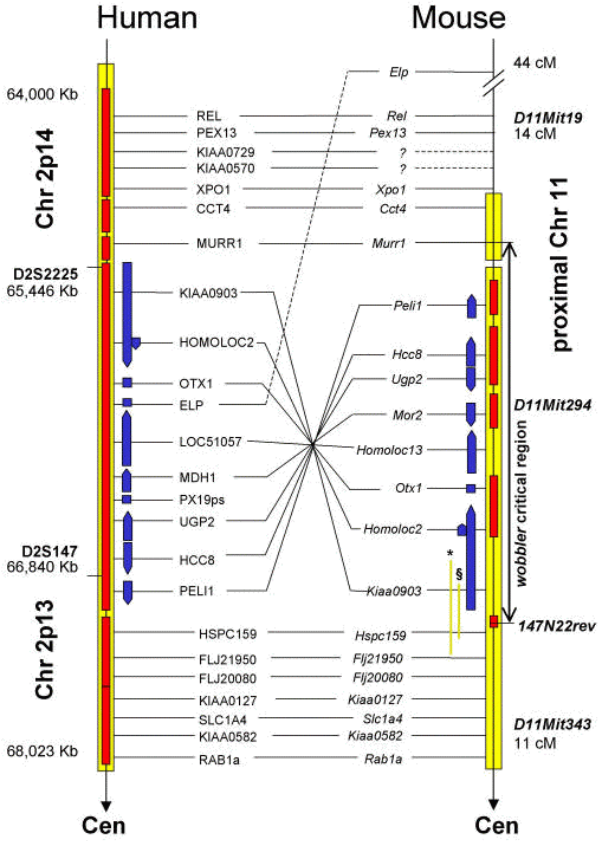
Two genomes similarity

- 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
- similarity has a different meaning in different areas

Montag

- Übungszeit: zwei Wochen für Grundlagen-Proteinstruktur benutzt