Protein Struktur

- Biologen und Chemiker dürfen mit Handys spielen (leise)
- go home, go to sleep
- wake up at slide 39

Andrew Torda, Wintersemester 2017/ 2018
Proteins - who cares?

Most important molecules in life? Ask the DNA / RNA people

- structural (keratin / hair)
- enzymes (catalysts)
- messengers (hormones)
- regulation (bind to other proteins, DNA, ..)
- industrial – biosensors to washing powder
- receptors
- transporters (O₂, sugars, fats)
- anti-freeze ...
Proteins are easy

- data (protein data bank, www.rcsb.org)
  - $\approx 10^5$ files
- literature on function, interactions, structure
- software
  - viewers, molecular dynamics simulators, docking, ..
- nomenclature and rules

Proteins are not friendly

- one cannot take a sequence and predict structure/function
- data formats are full of surprises
- data contains error and mistakes
Protein Rules, Physics, Folklore

Physics / Chemistry
• protein + water = set of interacting atoms
  • can be calculated (not really)

Rules (not quantified)
• proteins unfold if you heat them (exceptions ?)
• many charged amino acids.. they are soluble
• if they are more than 300 residues, they have more than one domain,
• proteins fold to a unique structure (could you prove this ?)
  • lowest free energy structure
Protein chemistry

Chemists / biochemists
- sleep, go home
- one tiny surprise at the end of the lectures

Short version
- proteins are sets of building blocks (amino acids, residues, Reste)
- 20 types of residue
- chains of length few to $10^3$ (100 or 200 typical)
- small ones ($< \approx 50$ residues) are peptides
- they fold up to nice stable structures – why?

Longer version..
The Plan

- polymers
- different kinds of sidechain
- structure due to backbone (secondary structure)
- properties of sidechains
- representation
### Sizes

1 Å = 10^{-10} m or 0.1 nm

<table>
<thead>
<tr>
<th>structure</th>
<th>size</th>
</tr>
</thead>
<tbody>
<tr>
<td>bond CH</td>
<td>1 Å</td>
</tr>
<tr>
<td>CC</td>
<td>1.5 Å</td>
</tr>
<tr>
<td>protein radius</td>
<td>10 - 10^{2} Å</td>
</tr>
<tr>
<td>α-helix spacing</td>
<td>5 ½ Å</td>
</tr>
<tr>
<td>$C_i^\alpha$ to $C_{i+1}^\alpha$</td>
<td>3.8 Å</td>
</tr>
</tbody>
</table>

myoglobin picture 2w6w
Proteins are polymers

simple polymers\[ A \rightarrow X \rightarrow B \]

many times gives
\[ A \rightarrow X \rightarrow X \rightarrow X \rightarrow X \rightarrow X \rightarrow X \rightarrow X \rightarrow X \rightarrow B \]

element

what kind of polymer would this give?

Do you know what R is?
Why are proteins interesting polymers?

Why do boring polymers not have well-defined structures?

Each part of polymer wants to interact with all other parts equally
- no structural preferences
- plastic bags, Haushaltsfolie
- no regular structures

Properties that make proteins different from plastics..
Giving proteins character 1

More complicated backbone with H-bond

donor

acceptor

• basis of standard regular structures in proteins (secondary structure)
• repeating polymer unit:

If this was all there was
• all proteins would be the same
amino acids (monomers) all look like:

\[
\begin{align*}
\text{NH}_3^+ & \quad \text{C} \quad \text{C} \quad \text{O} \\
\text{H} & \quad \text{R} & \text{C} & \text{O} \\
\text{R} & \quad \text{H} & \quad \text{C} & \quad \text{C} & \quad \text{H} & \quad \text{O} & \quad \text{OH} \\
\text{NH}_2 & \quad \text{C} & \quad \text{C} & \quad \text{O} \\
\text{R} & \quad \text{H} & \quad \text{C} & \quad \text{C} & \quad \text{O} & \quad \text{OH} \\
\end{align*}
\]

How can we construct specific structures?
• different kinds of "R" groups
Putting monomers together

A → X → B

• protein synthesis story (biochemistry lectures)
• peptides and proteins
  • < 30 or 40 residues = peptide
  • > 30 or 40 residues = protein
Backbone peptide bonds

How many backbone angles?

• 3 ($\phi$, $\psi$, $\omega$)

Peptide bond $\omega$ is planar

• partial double bond character (resonance forms)
• shorter than other C-N
• nearly always *trans*

Note: usually we do not draw H atoms
Backbone rotatable angles

Two rotatable angles $\phi, \psi$
some $\phi$ rotations

can we rotate freely?
• no... steric hindrance

• look at bottom – two unhappy O atoms
Ramachandran plot

can we rotate freely?
• no... steric hindrance

Ramachandran plot will reappear very often
Backbone H bonds

- oxygen is slightly negative
- NH bond is polar

H-bonds
- can be near or far in sequence
- fairly stable at room temperature
Secondary structure

Regular structures using information so far
• rotate phi ($\varphi$), psi ($\psi$) angles so as to
  • form H-bonds where possible
  • do not force side chains to hit each other (steric clash)

Two common structures
• $\alpha$-helix
• $\beta$-strand / sheet
\( \alpha \) helix

- each CO of residue \( i \) H-bonded to N of \( i+4 \)
- 3.6 residues per turn
- 2 H-bonds per residue
- side chains well separated
**β-sheet**

**β-strand**
- stretch out backbone and make NH and CO groups point out

**β-sheet**
- join these strands together with H-bonds
  (2 H-bonds/residue)

or parallel

anti-parallel
After \(\alpha\)-helix and \(\beta\)-sheet

Do helices and sheets explain everything? No
- there is flexibility in the angles (look at plot)
  - geometry is not perfectly defined
- there are local deviations and exceptions

Other common structures
- tighter helices
- some turns

Other structure
- coil, random, not named
What determines secondary structure?

So far
- secondary structure pattern of H-bonding
  - almost all residues have H-bond acceptor and donor
    - almost all could form $\alpha$-helix or $\beta$-sheet
Difference?
  - sequence of side-chains – overall folding
Why else are sidechains important
- chemistry of proteins (interactions, catalysis)
Fundamental dogma
- the sequence of sidechains determines the protein shape
side chain possibilities

- big / small
- charged +, charged -, polar
- hydrophobic (not water soluble), polar
- interactions between sites...

- a CS interaction is different to CC is different to RG ...

Andrew Torda 16.10.2017
Side chain properties

properties
• big / small
• neutral / polar / charged
• special (...)

eample
• phenylalanine side chain looks like benzene (benzin)
  • very insoluble
  • benzene would rather interact with benzene than water
• what if you have phe-phe-phe... poly-phe?
  • does not happen in nature (can be made)
  • would be insoluble
  • not like a real peptide
• phe is a constituent of real proteins – has a role

Andrew Torda 16.10.2017
Properties are not clear cut

You can be big / small, hydrophic / polar
• combinations are possible

Do not memorise this figure

Sidechain interactions

- ionic (if the sidechains have charge)
- hydrophobic (insoluble sidechains)
- H-bonds (some donors and acceptors)
- repulsive
Summary of amino acids (first dozen)

- Glycin (Gly)
- Threonin (Thr)
- Serin (Ser)
- Cystein (Cys)
- Tyrosin (Tyr)
- Asparagin (Asn)
- Glutamin (Gln)
- Arginin (Arg)
- Lysin (Lys)
- Histidin (His)
- Asparaginsäure (Asp)
- Glutaminsäure (Glu)
summary of amino acids (part 2)

Alanine (Ala)  \[ \text{H}_3\text{N}^+ \text{C} - \text{C} - \text{C} - \text{C} - \text{O}^- \]

Valine (Val)  \[ \text{H}_3\text{N}^+ \text{C} - \text{C} - \text{C} - \text{C} - \text{O}^- \]

Phenylalanine (Phe)  \[ \text{H}_3\text{N}^+ \text{C} - \text{C} - \text{C} - \text{C} - \text{O}^- \]

Methionine (Met)  \[ \text{H}_3\text{N}^+ \text{C} - \text{C} - \text{C} - \text{S} - \text{CH} - \text{C} - \text{C} - \text{O}^- \]

Leucine (Leu)  \[ \text{H}_3\text{N}^+ \text{C} - \text{C} - \text{C} - \text{C} - \text{O}^- \]

Proline (Pro)  \[ \text{H}_3\text{N}^+ \text{C} - \text{C} - \text{C} - \text{C} - \text{O}^- \]

Isoleucine (Ile)  \[ \text{H}_3\text{N}^+ \text{C} - \text{C} - \text{C} - \text{C} - \text{O}^- \]

Tryptophan (Trp)  \[ \text{H}_3\text{N}^+ \text{C} - \text{C} - \text{C} - \text{C} - \text{O}^- \]
Amino Acids by property

aromatic

tryptophan

phenylalanine

tyrosine
rather hydrophobic

leucine

\[ \text{N} - \text{C} - \text{C} - \text{C} - \text{N} \]

isoleucine

\[ \text{N} - \text{C} - \text{C} - \text{C} - \text{O} \]

cysteine

\[ \text{S} - \text{N} - \text{C} - \text{C} - \text{O} \]

methionine

\[ \text{S} - \text{N} - \text{C} - \text{C} - \text{O} \]

alanine

\[ \text{N} - \text{O} \]

proline

\[ \text{O} - \text{C} - \text{N} \]

glycine

\[ \text{N} - \text{O} \]

valine

\[ \text{N} - \text{C} - \text{C} - \text{O} \]
Polar

threonine

serine

glutamine

asparagine
charged

histidine

lysine

aspartate

arginine

 glutamate

• Muss ich alle Strukturen für die Klausur wissen?
Hydrophobicity – how serious?

Very serious, but simplified
- the lists above are
  - pH dependent
  - difficult to measure experimentally (some aspects)
- Is there a single definition for hydrophobicity?

Other properties - size

![Other properties - size](image)
Other properties – chemistry / geometry

Proline
- only one rotatable angle!
- peptide bond sometimes $cis$
- pro ramachandran plot
gly and cys

glycine
- no side chain
- can visit forbidden parts of phi-psi map

cysteine
- forms covalent links with other cys
Summary so far

- Proteins are heteropolymers.
- Backbone forms $\alpha$-helices and $\beta$-strands (and more)
  - Not sequence specific.
- Side-chains determine the
  - Pattern of secondary structure.
  - Overall protein shape.
- Special amino acids
  - Cys (forms disulfide bridges).
  - Gly (can visit "forbidden" regions of Ramachandran plot).
  - Pro (no H-bond donor).
- How many sequences can one have? $20^n$. 

Andrew Torda
16.10.2017
Nomenclature

Some rules are unavoidable

Alanine       Ala     A
Cysteine      Cys     C
Aspartic acid Asp     D
Glutamic acid Glu     E
Phenylalanine Phe     F
Glycine       Gly     G
Histidine     His     H
Isoleucine    Ile     I
Lysine        Lys     K
Leucine       Leu     L
Methionine    Met     M
Asparagine    Asn     N
Proline       Pro     P
Glutamine     Gln     Q
Arginine      Arg     R
Serine        Ser     S
Threonine     Thr     T
Valine        Val     V
Tryptophan    Trp     W
Tyrosine      Tyr     Y

Always write from N to C terminal (convention)
More definitions

- **primary structure**
  - sequence of amino acids
    - ACDF (ala cys asp phe...)
- **secondary structure**
  - $\alpha$-helix, $\beta$-sheet (+ few more)
    - structure defined by local backbone
- **tertiary structure**
  - how these units fold together
  - coordinates of a protein
distributions of residue types

Surprise coming

• 20 amino acid types – are they all equally common?
• Are you made of $\frac{1}{20} = 5\%$ of ala, leu, cys, ...?
amino acid frequencies

- leu
- ala
- gly
- val
- glu
- ser
- ile
- lys
- arg
- asp
- thr
- pro
- asn
- gln
- phe
- tyr
- met
- his
- cys
- trp

(swissprot (2014))
What would Darwin say?

Why?
- so much ala, leu
- so little trp, cys, his, met

A story
- Darwinist
- non-Darwinist

What would Darwin say?
- There is a chemical / biological reason
Think Darwinist

Empirical fact
• trp, cys, met are rare in proteins

Consequence
• too much trp is bad for you / expensive / dangerous

Possibilities
• metabolic cost issues
  • does it cost energy / nutrients to make trp ? cys with its sulfur ?
• protein structure – lots of chemical differences between amino acid types
  • if you put lots of trp / cys / met in a protein
    • does it not fold ? Does it become unstable ?
• if free trp toxic ?
Common amino acids

Leu and ala
• cheap to synthesise?
• do you get them as by-products from other biochemistry?
• what is their advantage in protein structure?
  • stability? rigidity? flexibility?

Forget Darwin – think neutral evolution
• what do we mean by Darwinism?
Very Darwinist

- protein sequences
  - mutation
    - new sequences
      - yes
        - protein OK?
        - yes
        - no
          - trash can
Think neutralist

- OK/not OK step (selection) less important
- What determines the sequences you see?
  - "mutation" step
- mutation step looks very simple
  - not really
- consider the meaning and biases

![Diagram]

- protein sequences
  - mutation
  - new sequences
    - protein OK?
      - yes
      - no
Codon bias

- look at the most rare amino acids...

- number of codons not quite everything

- some bases are more common than others

\[
p(\text{his}) = p(C) \cdot p(A) \cdot p(C) + p(C) \cdot p(A) \cdot p(C) \\
= 0.22 \cdot 0.30 \cdot 0.22 + 0.22 \cdot 0.30 \cdot 0.22 \\
\approx 0.03
\]

- does this predict the probability of all amino acids?
- if yes, there is no selection for amino acids

<table>
<thead>
<tr>
<th></th>
<th>U</th>
<th>A</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>22 %</td>
<td>30 %</td>
<td>22 %</td>
<td>26 %</td>
</tr>
</tbody>
</table>
How relevant is Darwinism?

One outlier (arg)

Little evidence of Darwinist selection in amino acid types

Logical consequence
- there are many sites in proteins where it does not matter which amino acid is used

Forget Darwinism and selection of amino acids?

No

- arg example
- lots of mutation data
  - for an enzyme
    - most mutations are a bit bad, some do not matter

- Do not be a pure Darwinist
- do not interpret everything you see in terms of fitness
Ultimately, our representation of a structure...

<table>
<thead>
<tr>
<th>x, y, z coordinates</th>
</tr>
</thead>
</table>

Drawing the structure?
• where are atoms? – therapeutic binding
• which residues could be involved in interactions?
What is the surface? Where could molecules fit?
Representations

Colour surface by hydrophobicity
Highlight / emphasise regular structures
Why does structure matter?

- what residues can I change and preserve function?
- what is the reaction mechanism of an enzyme?
- what small molecules would bind and block the enzyme?
- is this protein the same shape as some other of known function?

Where do structures come from?

- X-ray crystallography
- NMR
- + a bit of small angle X-ray scattering, electron diffraction, neutron diffraction...
resolution, precision, accuracy

Coordinates 27.370  13.558  -14.735
  • what do they mean?
Random errors
  • non-systematic / noise / uncertainty
  • should be scattered around correct point

X-ray crystallography has model for data
  • uncertainty (probability)
  • resolution (experimental)
    • < 1 Å (unusually good)
    • > 5 Å (bad, but examples..
     3LJ5  Full Length Bacteriophage P22 Portal Protein
     3M0C X-ray Crystal Structure of PCSK9 in Complex with the LDL receptor
X-ray crystallography

Non-systematic errors
- small problems: (O and N look the same)
- few huge problems
- newer structures are better

Proteins are not static
- overall motion
- local motion
NMR structures

Different philosophy to X-ray
- lots of little internal distances
- do not quite define structure

Generate 50 or $10^2$ solutions
- look at scatter of solutions

As with X-ray
- some parts are well defined
- some not

structure 1sm7
Summarise and stop

- roles of proteins
- heteropolymers – 20 types of amino acid / residue
- geometry – avoiding atomic clashes, forming H bonds
  - leads to regular secondary structure
- chemistry of amino acids very different to another
- unique structure for a sequence reflects these differences
- representations of structures
- structures in PDB are experimental – have errors
some questions

- $\text{(Asp)}_{100}$
  - is it soluble? Is it acidic / basic?
  - would it form a compact regular structure?
- How big is sequence space? How much has been tried by evolution?
- if you have a protein of poly-trp, would it form a specific structure? How would it behave in solution?
- for length $n$, do all / many / few of the $n^{20}$ sequences form specific structures?
- how would a Darwinist explain the uneven distribution of amino acid usage?
- why would you want to represent a protein by its surface?
- why might you draw it as a series of helices and strands?
- what is the biggest chain in the protein data bank? Examples
  - fatty acid synthase $> 2 \times 10^3$ residues/chain
  - dynein heavy chain motor domain $> 4 \times 10^3$ residues/chain