Protein Struktur

- Biologen und Chemiker dürfen mit Handys spielen (leise)
- go home, go to sleep
- wake up at slide 39
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 ($O_2$, sugars, fats)
- anti-freeze ...
Proteins are easy

- data (protein data bank, www.rcsb.org)
  - $1.4 \times 10^5$ 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 (< ≈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</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>$\alpha$-helix spacing</td>
<td>5 ½ Å</td>
</tr>
<tr>
<td>$C_i^\alpha$ to $C_i^{\alpha+1}$</td>
<td>3.8 Å</td>
</tr>
</tbody>
</table>
Proteins are polymers

simple polymers \( \text{A} - \text{X} - \text{B} \)

many times gives

\[
\text{A} - \text{X} - \text{X} - \text{X} - \text{X} - \text{X} - \text{X} - \text{X} - \text{X} - \text{B}
\]

example

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 and 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
proteins chemistry

amino acids (monomers) all look like:

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

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

- 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

unhappy O atoms – high energy
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
*α* 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)

![Diagram of β-sheet with anti-parallel and parallel orientations]
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 side-chains 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
15.10.2018
Side chain properties

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

example
• 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
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)
- Arginim (Arg)
- Lysin (Lys)
- Histidin (His)
- Asparaginsäure (Asp)
- Glutaminsäure (Glu)
summary of amino acids (part 2)

Alanin (Ala)  Valin (Val)

Phenylalanin (Phe)  Methionin (Met)

Leucin (Leu)  Prolin (Pro)

Isoleucin (Ile)  Tryptophan (Trp)
Amino Acids by property

aromatic

tryptophan

phenylalanine

tyrosine
rather hydrophobic

leucine

isoleucine

cysteine

methionine

alanine

proline

glycine

valine
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

trp big … small gly ala
Proline

- only one rotatable angle!
- peptide bond sometimes \textit{cis}
- pro ramachandran plot
glycine
• no side chain
• can visit forbidden parts of phi-psi map

evolutionary consequences?

ala-ala-ala cannot adopt this conformation
cys residues can bind to each other

effect on conservation (mutation rates)?
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? $20^n$
Nomenclature

Some rules are unavoidable

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

Always write from N to C terminal (convention)
Definitions, primary, secondary ...

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

Interpretation / explanation
• too much trp is bad for you / expensive / dangerous

Possibilities
• metabolic cost issues
  • does it cost energy 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

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

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

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>U</td>
<td>22 %</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>30 %</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>22 %</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>26 %</td>
</tr>
</tbody>
</table>
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
**Representation**

Ultimately, our representation of a structure...

<table>
<thead>
<tr>
<th>ATOM</th>
<th>x</th>
<th>y</th>
<th>z</th>
<th>Temperature</th>
<th>Occupancy</th>
<th>Chain ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31.758</td>
<td>13.358</td>
<td>-13.673</td>
<td>1.00</td>
<td>18.79</td>
<td>1BPI 137</td>
</tr>
<tr>
<td>2</td>
<td>31.718</td>
<td>13.292</td>
<td>-12.188</td>
<td>1.00</td>
<td>14.26</td>
<td>1BPI 138</td>
</tr>
<tr>
<td>3</td>
<td>33.154</td>
<td>13.224</td>
<td>-11.664</td>
<td>1.00</td>
<td>18.25</td>
<td>1BPI 139</td>
</tr>
<tr>
<td>4</td>
<td>33.996</td>
<td>12.441</td>
<td>-12.225</td>
<td>1.00</td>
<td>20.10</td>
<td>1BPI 140</td>
</tr>
<tr>
<td>5</td>
<td>30.886</td>
<td>12.103</td>
<td>-11.724</td>
<td>1.00</td>
<td>16.74</td>
<td>1BPI 141</td>
</tr>
<tr>
<td>6</td>
<td>29.594</td>
<td>11.968</td>
<td>-12.534</td>
<td>1.00</td>
<td>15.96</td>
<td>1BPI 142</td>
</tr>
<tr>
<td>7</td>
<td>28.700</td>
<td>13.182</td>
<td>-12.299</td>
<td>1.00</td>
<td>15.45</td>
<td>1BPI 143</td>
</tr>
<tr>
<td>8</td>
<td>27.267</td>
<td>12.895</td>
<td>-12.546</td>
<td>1.00</td>
<td>12.82</td>
<td>1BPI 144</td>
</tr>
<tr>
<td>9</td>
<td>26.661</td>
<td>13.087</td>
<td>-13.727</td>
<td>1.00</td>
<td>17.38</td>
<td>1BPI 145</td>
</tr>
<tr>
<td>10</td>
<td>27.370</td>
<td>13.558</td>
<td>-14.735</td>
<td>1.00</td>
<td>18.38</td>
<td>1BPI 146</td>
</tr>
<tr>
<td>11</td>
<td>25.367</td>
<td>12.797</td>
<td>-13.838</td>
<td>1.00</td>
<td>25.73</td>
<td>1BPI 147</td>
</tr>
<tr>
<td>12</td>
<td>33.800</td>
<td>13.936</td>
<td>-10.586</td>
<td>1.00</td>
<td>17.07</td>
<td>1BPI 148</td>
</tr>
<tr>
<td>13</td>
<td>34.976</td>
<td>13.367</td>
<td>-9.840</td>
<td>1.00</td>
<td>14.99</td>
<td>1BPI 149</td>
</tr>
<tr>
<td>14</td>
<td>34.960</td>
<td>11.922</td>
<td>-9.660</td>
<td>1.00</td>
<td>13.11</td>
<td>1BPI 150</td>
</tr>
<tr>
<td>15</td>
<td>33.962</td>
<td>11.306</td>
<td>-9.391</td>
<td>1.00</td>
<td>10.57</td>
<td>1BPI 151</td>
</tr>
<tr>
<td>16</td>
<td>34.922</td>
<td>14.145</td>
<td>-8.523</td>
<td>1.00</td>
<td>15.81</td>
<td>1BPI 152</td>
</tr>
<tr>
<td>17</td>
<td>15.391</td>
<td>-8.737</td>
<td>1.00</td>
<td>18.91</td>
<td>1BPI 153</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>15.273</td>
<td>-10.096</td>
<td>1.00</td>
<td>19.41</td>
<td>1BPI 154</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>-9.707</td>
<td>8.73</td>
<td>1.00</td>
<td></td>
<td>1BPI 155</td>
<td></td>
</tr>
</tbody>
</table>

**Drawing the structure?**
Representations

- 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
Representations

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
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?
- Why does a protein fold to a specific structure, but polyethylene not?
- Glycine can reach parts of the angle space which other residues cannot. What are the evolutionary consequences?
- 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\times10^3 residues/chain
  - dynein heavy chain motor domain > 4\times10^3 residues/chain