Introduction:
Principles of protein structures

Lausanne, February 19, 2007

Torsten Schwede
Lorenza Bordoli
Simon Bernèche

Torsten.Schwede@unibas.ch
Lorenza.Bordoli@unibas.ch
Simon.Berneche@unibas.ch
Overview

- Amino Acids and Polypeptides (1)
- Physicochemical Properties of Amino Acid Side Chains (2)
- Protein Folding and Physicochemical Interactions (3)
- Principles of Protein Structure (4)
- Structure Comparison (5)
- Protein Structure Databases (6)

Amino Acids and Polypeptides (1)
### Amino Acids

Three and one letter code:

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Three-letter code</th>
<th>One-letter code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>Ala</td>
<td>A</td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg</td>
<td>R</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asn</td>
<td>N</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>Asp</td>
<td>D</td>
</tr>
<tr>
<td>Asparagine or aspartic acid</td>
<td>Asx</td>
<td>B</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys</td>
<td>C</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Gln</td>
<td>Q</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>Glu</td>
<td>E</td>
</tr>
<tr>
<td>Glutamine or glutamic acid</td>
<td>G1x</td>
<td>Z</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>Leucine</td>
<td>Leu</td>
<td>L</td>
</tr>
<tr>
<td>Lysine</td>
<td>Lys</td>
<td>K</td>
</tr>
<tr>
<td>Methionine</td>
<td>Met</td>
<td>M</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Phe</td>
<td>F</td>
</tr>
<tr>
<td>Proline</td>
<td>Pro</td>
<td>P</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>Tryptophan</td>
<td>Trp</td>
<td>W</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Tyr</td>
<td>Y</td>
</tr>
<tr>
<td>Valine</td>
<td>Val</td>
<td>V</td>
</tr>
</tbody>
</table>

### Amino Acids with aliphatic Side-Chains

- **Ala (A)**: \( \text{CH}_3\text{CH}(-\text{COO}^-)\text{NH}_3^+ \)
- **Val (V)**: \( \text{H}_3\text{C}\text{CH}(-\text{COO}^-)\text{NH}_3^+ \)
- **Ile (I)**: \( \text{H}_3\text{CCH}_2\text{CH}(-\text{COO}^-)\text{NH}_3^+ \)
- **Leu (L)**: \( \text{H}_3\text{CCH}_2\text{CH}_2\text{CH}(-\text{COO}^-)\text{NH}_3^+ \)
Sidechains with hydroxyl (-OH) groups

- Ser (S) \( pK_a = 13 \)
- Thr (T) \( pK_a = 13 \)

Sidechains containing sulphur

- Cys (C) \( pK_a = 8.3 \)
- Met (M)
Acidic amino acids

Asp (D) $pK_a=3.9$

Glu (E) $pK_a=4.1$

Amides of acidic amino acids

Asn (N)

Gln (Q)
Basic Amino Acids

Arg (R)  
\[ \text{pK}_a = 12.5 \]  

Lys (K)  
\[ \text{pK}_a = 10.8 \]  

His (H)  
\[ \text{pK}_a = 6.0 \]  

Side-chains with aromatic rings

Phe (F)  

Tyr (Y)  
\[ \text{pK}_a = 10.1 \]  

Trp (W)
Special cases ...

Imino acid

Pro (P)

Gly (G)

Side Chain Structures
Physicochemical Properties of Amino Acid Side Chains (2)

**Neutral Hydrophobic**
- Alanine
- Valine
- Leucine
- Isoleucine
- Proline
- Tryptophane
- Phenylalanine
- Methionine

**Neutral Polar**
- Glycine
- Serine
- Threonine
- Tyrosine
- Cysteine
- Asparagine
- Glutamine

**Basic**
- Lysin
- Arginine
- (Histidine)

**Acidic**
- Aspartic Acid
- Glutamic Acid

---

**pH and pKa**

- **pH**
  \[ pH = -\log [H^+] \]

- **Water ion product**
  \[ Kw = [H^+] [OH^-] = 10^{-14} \]
  \[ \log[H^+] + \log[OH^-] = \log(10^{-14}) \]
  \[ pH + pOH = 14 \]
pH and pKa

- **Dissociation of weak acids**

\[
\text{HA} \leftrightarrow \text{H}^+ + \text{A}^-
\]

\[
K_a = \frac{[\text{H}^+][\text{A}^-]}{[\text{HA}]} \quad [\text{H}^+] = K_a \frac{[\text{HA}]}{[\text{A}^-]}
\]

\[
\log [\text{H}^+] = \log K_a + \log \frac{[\text{HA}]}{[\text{A}^-]}
\]

\[
-\log [\text{H}^+] = -\log K_a + \log \frac{[\text{A}^-]}{[\text{HA}]}
\]

- **Henderson - Hasselbach Equation**

\[
pH = pK_a + \log \frac{[\text{A}^-]}{[\text{HA}]}
\]
Glu
**pH and pKa**

- **Lys**

  Enzymatic reactions often require proton transfer.

  Q: Which amino-acid(s) are able to change their protonation state under physiological conditions?
Protein Folding and Physicochemical Interactions (3)

Why do proteins fold?

Anfinson’s paradigm

- 1957, Nobel Prize 1972
All the necessary information for the 3-dimensional structure of an enzyme is contained in the primary structure or sequence of the amino acids.

Levinthal's Paradox (1968)

If a chain of a hundred amino acids is considered and it assumed each amino acid can exist in one of three conformations, extended, helical or loop, then there are $3^{100}$ possible ways to arrange this chain.

This is roughly $10^{48}$ conformations. Bond rotation can be estimated to occur at a rate of roughly $10^{14}$ s$^{-1}$. This means that search for the right conformation through random searching alone would take the order of $10^{34}$s or $10^{26}$ years, several orders of magnitudes greater than the age of the universe!

The protein sequence contains all information needed to create a correctly folded protein.

- Many proteins fold spontaneously to their native structure
- Protein folding is relatively fast (nsec – sec)
- Chaperones speed up folding, but do not alter the structure

The protein sequence contains all information needed to create a correctly folded protein.

Why do proteins fold?

- N - Cα(HR₁) - CO - N - Cα (HR₂) - CO - N - Cα(HR₁) - CO -
Side Chain Properties

**Neutral Hydrophobic**
- Alanine
- Valine
- Leucine
- Isoleucine
- Proline
- Tryptophane
- Phenylalanine
- Methionine

**Neutral Polar**
- Glycine
- Serine
- Threonine
- Tyrosine
- Cysteine
- Asparagine
- Glutamine

**Basic**
- Lysin
- Arginine
  - (Histidine)

**Acidic**
- Aspartic Acid
- Glutamic Acid

Hydrophobic Effects
- **main driving force for protein folding**
Hydrophobic Effects

- main driving force for protein folding

Water molecules in bulk water are mobile and can form H-bonds in all directions.

Hydrophobic surfaces don't form H-bonds. The surrounding water molecules have to orient and become more ordered.

The entropy loss can be minimized by gathering the hydrophobic surfaces together in the core of a protein and separating them from the solvent.

1D-Structure prediction

- Projection onto strings of structural assignments

  E.g. “Solvent Accessibility”

  A B C D E F G ...
  | | | | | | | |
  e e b b e e e ...
Surface Definitions

- Van der Waals Radius
- Molecular Surface
- Solvent Accessible Surface

Hydrogen Bonds

- H-atoms bound to electronegative atoms (e.g. N, O) are polarized and can form H-bonds
- H-bonding partners include:
  - main chain atoms
  - side chain atoms
  - water molecules
  - ligands, etc...

\[ \text{H-atoms bound to electronegative atoms (e.g. N, O) are polarized and can form H-bonds} \]

\[ \text{H-bonding partners include:} \]

\[ \text{main chain atoms} \]

\[ \text{side chain atoms} \]

\[ \text{water molecules} \]

\[ \text{ligands, etc…} \]
Energetics of protein folding

\[ \Delta G = \Delta H - T \Delta S \]

- Difference of two very large energetic terms
- Low overall stabilization energy

Hydrogen bonding amino acids

- Peptide
- Glut / Asp Acid
- Threonine
- Glutamine / Asparagine
- Lysine
- Histidine
- Tryptophan
- Arginine
- Tyrosine
- Serine
Why do proteins fold?

- Change of energy state from unfolded to folded
- Folded state must have overall lower energy
- Let’s assume the folded state is the lowest possible state for this polypeptide

Protein Sequence Space

- How many different proteins are theoretically possible?
- How many of these have been tested during evolution?
Protein sequence space

- Assuming a peptide of length 100 aa

Possible combinations: \( n_c = 20^{100} \approx 1.27 \times 10^{130} \)

Volume of one peptide:

- \( r_{\text{atom}} \approx 2\text{Å} \)
- \( v_{\text{atom}} \approx 35\text{Å}^3 \)
- \( \text{packing} \approx 75\% \)
- \( v_{\text{peptide}} \approx 1.3 \times 10^5 \text{Å}^3 \)

Protein sequence space

- 1.27\( \times 10^{130} \) combinations. For comparison ...

Volume of the Earth:

\[ R \approx 6.4 \times 10^3 \text{km} \approx 6.4 \times 10^{16} \text{Å} \]
\[ V = \frac{4}{3} \pi R^3 \approx 1.1 \times 10^{51} \text{Å}^3 \]

Peptides/Earth:

\[ n_p \approx \frac{1.1 \times 10^{51}}{1.3 \times 10^5} \approx 7.7 \times 10^{45} \]
Protein sequence space

- 1.27*10^{130} combinations. For comparison...

Age of the Earth: $3 \times 10^9 \text{years} \approx 2.6 \times 10^{13} \text{hours}
\approx 9.5 \times 10^{16} \text{sec}
\approx 9.5 \times 10^{28} \text{psec}$

- If the whole planet consisted of peptides, and peptides were renewed every psec...

$$n_T \approx \left(7.7 \times 10^{45}\right) \times \left(9.5 \times 10^{28}\right) \approx 7.3 \times 10^{74}$$

Protein sequence space

- Assuming a peptide of length 100 aa

Possible combinations: $n_c = 20^{100} \approx 1.27 \times 10^{130}$

- If the whole planet consisted of peptides, and peptides were renewed every psec...

$$n_T \approx \left(7.7 \times 10^{45}\right) \times \left(9.5 \times 10^{28}\right) \approx 7.3 \times 10^{74}$$

$$10^{130} - 10^{75} \approx 10^{130}$$
Principles of Protein Structure (4)

- Primary Structure
- Secondary Structure
- Tertiary Structure
- Quaternary Structure
Stereochemistry: L- and D-amino acids

Geometry of a peptide bond
Dihedral angles

\( \Phi, \Psi, \) and \( \omega \)

Q: Which values would you expect for \( \omega \)?

Dihedral angles \( \Phi \) and \( \Psi \)

\( \phi = 0^\circ, \psi = 180^\circ \) \hspace{1cm} \( \phi = 180^\circ, \psi = 0^\circ \)
Dihedral angles $\Phi$ and $\Psi$

$\Phi = -60^\circ, \psi = 180^\circ$

$\Phi = 0^\circ, \psi = 0^\circ$

Ramachandran Plots

$\Psi$ (deg)

$\Phi$ (deg)
Ramachandran Plots

$\Psi$ (deg)

$\Phi$ (deg)

Amino acid preferences

- Alanine and Arginine

ALA    ARG

ALI    ARG
Amino acid preferences

- Amino acid with special preferences:
  - GLY
  - PRO

Right-handed (clockwise) helical conformation

Right-handed alpha-helix. White dots show the hydrogen bonds.

Alpha helices
Alpha helices

Beta strands / beta sheets
Anti-parallel beta sheet

Parallel and anti-parallel beta sheets
Left-handed twist in beta-sheets

The left-handed twist of beta-sheets.

0° - 30° per aa

Turns and loops

Schematic diagram showing the interresidue backbone hydrogen bonds that stabilize the reversal of the chain direction. Side chains are depicted as large light purple spheres. Due to the tight geometry of the turn, some residues are found more commonly in turns than others.
Turns and loops

- Hairpin loops

```
<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Preference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha$-helix</td>
</tr>
<tr>
<td>Glu</td>
<td>1.59</td>
</tr>
<tr>
<td>Ala</td>
<td>1.41</td>
</tr>
<tr>
<td>Leu</td>
<td>1.34</td>
</tr>
<tr>
<td>Met</td>
<td>1.38</td>
</tr>
<tr>
<td>Gin</td>
<td>1.27</td>
</tr>
<tr>
<td>Lys</td>
<td>1.23</td>
</tr>
<tr>
<td>Arg</td>
<td>1.21</td>
</tr>
<tr>
<td>His</td>
<td>1.05</td>
</tr>
<tr>
<td>Val</td>
<td>0.90</td>
</tr>
<tr>
<td>Ile</td>
<td>1.09</td>
</tr>
<tr>
<td>Tyr</td>
<td>0.74</td>
</tr>
<tr>
<td>Cys</td>
<td>0.66</td>
</tr>
<tr>
<td>Trp</td>
<td>1.02</td>
</tr>
<tr>
<td>Phe</td>
<td>1.16</td>
</tr>
<tr>
<td>Thr</td>
<td>0.76</td>
</tr>
<tr>
<td>Gly</td>
<td>0.43</td>
</tr>
<tr>
<td>Asn</td>
<td>0.76</td>
</tr>
<tr>
<td>Pro</td>
<td>0.34</td>
</tr>
<tr>
<td>Ser</td>
<td>0.57</td>
</tr>
<tr>
<td>Asp</td>
<td>0.99</td>
</tr>
</tbody>
</table>
```

Pair Wise Structure Comparison (5)

pairwise

Pair Wise Structure Comparison

- Root mean square deviation
  - Comparing two structures A and B
  - $R_{i,A}$ = Position of atom $i$ in structure A
  - $n$ = Number of equivalent atoms

$$r.m.s.d. = \sqrt{\frac{\sum_{i=0}^{n} (R_{i,A} - R_{i,B})^2}{n}}$$
RMSD

- Comparing two structures

Min: \( r.m.s.d. = \sqrt{\frac{\sum_{i=1}^{n} (R_{i,A} - R_{i,B})^2}{n}} \)
Structural Alignments

- **Protein Structure is better conserved than sequence**
- Structural alignments establish equivalences between amino acid residues based on the 3D structures of two or more proteins
- Structure alignments therefore provide information not available from sequence alignment methods
- Structural alignments can be used to guide sequence alignments (see: T_COFFEE / SAP)
Protein Structure Databases (6)

- PDB
- EBI-MSD
- SCOP
- CATH
PDB Holdings List: 13-Apr-2004

<table>
<thead>
<tr>
<th>Exp. Tech.</th>
<th>Molecule Type</th>
<th>Proteins, Peptides, and Viruses</th>
<th>Protein/Nucleic Acid Complexes</th>
<th>Nucleic Acids</th>
<th>Carbohydrates</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-ray Diffraction and other</td>
<td></td>
<td>19696</td>
<td>969</td>
<td>722</td>
<td>14</td>
<td>21406</td>
</tr>
<tr>
<td>NMR</td>
<td></td>
<td>9031</td>
<td>97</td>
<td>577</td>
<td>4</td>
<td>3708</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>22727</td>
<td>1066</td>
<td>1304</td>
<td>18</td>
<td>25115</td>
</tr>
</tbody>
</table>

Fold Classification Databases

- Number of folds is limited
  - according to C. Chothia ca. 1000 – 5000 folds

Fold databases systematically classify protein structure folds

- **FSSP**
  - Fold classification based on Structure-Structure alignment of Proteins
  - http://www2.ebi.ac.uk/dali/fssp/
  - Contains structurally aligned proteins based in DALI

- **SCOP**

- **CATH**
**Topology Cartoons**

- **TOPS: Topology of Protein Structure**
  - http://www.tops.leeds.ac.uk/

---

**CATH - Protein Structure Classification**

- Hierarchical classification of protein domain structures
- UCL, Janet Thornton & Christine Orengo
- clusters proteins at four major levels:
  - Class(C)
  - Architecture(A)
  - Topology(T)
  - Homologous superfamily (H)

[ http://www.biochem.ucl.ac.uk/bsm/cath_new/ ]
**CATH - Protein Structure Classification**

- **Class (C)**
  
  derived from secondary structure content is assigned automatically.

- **Architecture (A)**

  describes the gross orientation of secondary structures, independent of connectivity.

- **Topology (T)**

  clusters structures according to their topological connections and numbers of secondary structures.

  [http://www.biochem.ucl.ac.uk/bsm/cath_new/]

**SCOP - Structural Classification of Proteins**

- MRC Cambridge (UK), Alexey Murzin, Brenner S. E., Hubbard T., Chothia C.
- hierarchical classification of protein domain structures created by manual inspection
- comprehensive description of the structural and evolutionary relationships
- organized as a tree structure
  - Class
  - Fold
  - Superfamily
  - Family
  - Species

[http://scop.mrc-lmb.cam.ac.uk/scop/]

Protein Structure / Fold Databases

- PDB: http://www.pdb.org
- EBI-MSD http://www.ebi.ac.uk/msd/
- SCOP http://scop.mrc-lmb.cam.ac.uk/scop/
- CATH http://www.biochem.ucl.ac.uk/bsm/cath_new/

References

- P.E.Bourne, H. Weissig. Structural Bioinformatics, Wiley-Liss and Sons.