Proteins: Structure & Function

Ulf Leser
This Lecture

• Proteins
  - Structure
  - Function
  - Databases

• Predicting Protein Secondary Structure

• Examples often from O. Kohlbacher, Vorlesung Strukturvorhersage, WS 2004/2005, Universität Tübingen
Central Dogma of Molecular Biology

![Diagram showing the flow of information in molecular biology]

- **DNA** → **RNA** → **protein**
- **Expression** horizontal information flow
- **Replication** vertical information flow

<table>
<thead>
<tr>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>UU</td>
<td>CU</td>
<td>AU</td>
<td>UG</td>
</tr>
<tr>
<td>UU</td>
<td>UC</td>
<td>UU</td>
<td>UC</td>
</tr>
<tr>
<td>UA</td>
<td>UC</td>
<td>UC</td>
<td>UA</td>
</tr>
<tr>
<td>UU</td>
<td>UC</td>
<td>UC</td>
<td>UA</td>
</tr>
<tr>
<td>UG</td>
<td>CC</td>
<td>CA</td>
<td>CG</td>
</tr>
<tr>
<td>UU</td>
<td>CC</td>
<td>CC</td>
<td>CC</td>
</tr>
<tr>
<td>UC</td>
<td>CC</td>
<td>CC</td>
<td>CC</td>
</tr>
<tr>
<td>UG</td>
<td>CC</td>
<td>CC</td>
<td>CC</td>
</tr>
<tr>
<td>AU</td>
<td>CA</td>
<td>AC</td>
<td>AG</td>
</tr>
<tr>
<td>AU</td>
<td>AC</td>
<td>AC</td>
<td>AG</td>
</tr>
<tr>
<td>AC</td>
<td>AC</td>
<td>AC</td>
<td>AG</td>
</tr>
<tr>
<td>AU</td>
<td>AC</td>
<td>AC</td>
<td>AG</td>
</tr>
<tr>
<td>AA</td>
<td>GA</td>
<td>AG</td>
<td>GG</td>
</tr>
<tr>
<td>AA</td>
<td>GA</td>
<td>AG</td>
<td>GG</td>
</tr>
<tr>
<td>GA</td>
<td>GA</td>
<td>AG</td>
<td>GG</td>
</tr>
<tr>
<td>GA</td>
<td>GA</td>
<td>AG</td>
<td>GG</td>
</tr>
<tr>
<td>GA</td>
<td>GA</td>
<td>AG</td>
<td>GG</td>
</tr>
<tr>
<td>GA</td>
<td>GA</td>
<td>AG</td>
<td>GG</td>
</tr>
<tr>
<td>GA</td>
<td>GA</td>
<td>AG</td>
<td>GG</td>
</tr>
</tbody>
</table>

- Phenylalanine
- Leucine
- Serine
- Tyrosine
- Stop codon
- Stop codon
- Cysteine
- Stop codon
- Tryptophan
- Leucine
- Proline
- Histidine
- Glutamine
- Arginine
- Isoleucine
- Methionine
- Threonine
- Asparagine
- Lysine
- Serine
- Arginine
- Valine
- Alanine
- Aspartic acid
- Glutamic acid
- Glycine
- Valine
- Alanine
- Aspartic acid
- Glutamic acid
- Glycine
Details

- **Alternative Splicing**
  - “One gene – one protein” is wrong
  - Exons may be spliced out from the mRNA
  - Human: at least 6 times more unique proteins than genes

- **Post-translational modifications**
  - (De-)Phosphorylation, glycolysation, cleavage of signal peptides, ...
  - Human: At least 5 times more protein forms than proteins

- **Complexes**: Proteins **physically group together** to perform specific function
Example: Proteasome

- Function: Breaks (mis-folded, broken, superfluous, …) proteins into small peptides for reuse
- Very large complexes present in all eukaryotes (and more species)
  - >2000 kDa, consists of dozens of single proteins
  - Formation of the complex is a complex process only partly understood yet
Protein Structure

- **Primary**
  - 1D-Seq. of AA

- **Secondary**
  - 1D-Seq. of “subfolds”

- **Tertiary**
  - 3D-Structure

- **Quaternary**
  - Assembled complexes
Protein Function

- Proteins perform essentially everything that makes an organism alive
  - Metabolism
  - Signal processing
  - Gene regulation
  - Cell cycle
  - …
- For ~20% of all human gene, no function is known yet
- Describing function
  - Gene Ontology: 3 branches, >30,000 concepts
  - Used world-wide to describe gene/protein function
„Known“ Protein Functions

Function and Motifs

- Proteins usually have multiple functions
  - Avg. n# of GO terms assigned to a human protein: 6-10
- Functions are associated to motifs or domains
- There probably exist only 4000-5000 motifs
  - Proteins as assemblies of functional motifs
- Performing a function often requires binding to another protein or molecule
  - The binding requires a certain constellation of the protein structure
  - Major target of pharmacological research
UniProt

- “Standard” database for protein sequences and annotation
  - Original name: SwissProt
  - Started at the Swiss Institute of Bioinformatics, now mostly EBI
  - Other: PIR, HPRD

- Continuous growth and curation
  - >30 „Scientific Database Curators“
  - Quarterly releases
  - Very rich set of annotations

- Actually two databases
  - SwissProt: Curated, high quality, versioned
  - TrEMBL: Automatic generation from (putative) coding genomic sequences, low quality, redundant, much larger

Def. and removal of „redundant“ sequences

20258  Homo sapiens (Human)
16327  Mus musculus (Mouse)
  9842  Arabidopsis thaliana (Mouse-ear cress)
   7560  Rattus norvegicus (Rat)
   6582  Saccharomyces cerevisiae (Baker's yeast)
  5803  Bos taurus (Bovine)
   ...

Bacteria (60%)
Archaea (4%)
Viruses (3%)
Eukaryota (33%)

Other Vertebrata (10%)
Homo (11%)
Other (8%)
Nematoda (3%)
Insecta (5%)
Fungi (18%)
Other Mammalia (29%)
PDB – Protein Structure Database

- Oldest protein database, evolved from a book
- Contains experimentally obtained protein 3D-structures
  - Plus some DNA, protein-ligand, complexes, …
  - X-Ray (~75%), NMR (nuclear magnetic resonance, ~23%)
- Costly and rather slow techniques
  - Growth much smaller than that of sequence-related DBs
- Many problems with legacy data and data formats

InterPro

- Integrated database of protein signatures, classifications, and motifs
  - Currently ~21,000 signatures
- Associates signatures with function (GO term)
- InterProScan – quick identification of signatures in a protein sequence
  - For a fast, first functional annotation
This Lecture

• Introduction

• Predicting Protein Secondary Structure
  - Secondary structure elements
  - Chou-Fasman
  - GOR IV
  - Other methods
Amino Acids (AA)

- AA consist of a common core and a specific residue
  - Amino group - NH$_2$
  - Central C$_\alpha$ - Carbon - CH
  - Carboxyl group - COOH

- Residues (side chains) vary greatly between AA
- Residues determine the specific properties of a AA
Side Chains

**BASIC SIDE CHAINS**
- lysine (Lys, or K)
- arginine (Arg, or R)
- histidine (His, or H)

**NONPOLAR SIDE CHAINS**
- alanine (Ala, or A)
- valine (Val, or V)
- leucine (Leu, or L)
- isoleucine (Ile, or I)

**ACIDIC SIDE CHAINS**
- aspartic acid (Asp, or D)
- glutamic acid (Glu, or E)

**UNCHARGED POLAR SIDE CHAINS**
- asparagine (Asn, or N)
- glutamine (Gln, or Q)
- serine (Ser, or S)
- threonine (Thr, or T)
- tyrosine (Tyr, or Y)
- glycine (Gly, or G)
- cysteine (Cys, or C)
Structure of a Protein

- Concatenation of **cores**: Backbone of AA chain (a protein)
- Covalent **peptide bonds** between carboxyl and amino group (with loss of H₂O)
Flexibility

• In principle, every chemical bond can rotate freely
  - Would allow arbitrary backbone structures

• In real proteins, things are more restricted
  - Peptide bound (B) is “flat” – almost no torsion possible
  - Flexibility only in the $C_\alpha$-flanking bonds $\phi$ and $\psi$
Ramachandran Plots

- Combinations of $\phi$ and $\psi$ are highly constrained
  - Due to chemical properties of the backbone / side chains
- Two combinations are favored: $\alpha$-helixes and $\beta$-sheets
  - More detailed classifications exist
  - Angels lead to specific structures
  - Secondary structure
α-Helix

- Sequence of angles forming a regularly structured helix
- Additional bonds between amino and carboxyl groups
  - Very stable structure
- May have two orientations
  - Most are right-handed
- 3.4 AA per twist
- Often short, sometimes very long
\textbf{\(\beta\)-Sheet}

- Two linear and \textit{parallel stretches} (\(\beta\)-strands)
- Strands are bound together by hydrogen bounds
- Can be parallel or anti-parallel (wrt. N/C terminus)

Other Substructures

• $\alpha$-helixes and $\beta$-sheets cover 50-80% of most proteins
• Other parts are called loops or coils
  – Usually not very important for the structure of the protein
  – But very important for its function
  – Often exposed on the surface; important for binding to other molecules
Importance of Secondary Structure Prediction (SSP)

- Secondary structure elements (SSE) are vital for the overall structure of a protein
- Often evolutionary well conserved
- SSE can be used to classify proteins
  - Such classes are highly correlated with function
- SSE gives important clues to protein structure
- SSP much simpler than 3D structure prediction
  - And 3D structure prediction can benefit a lot from a good SSP
Predicting Secondary Structure

• SSP: Given a protein sequence, assign each AA in the sequence to one of the three classes Helix (H), Strand (E), or Coil (-)

KGYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESNFTNTHATNRNTD
GSTDYGILQINSRWWCNDGRTPGSKNLCNIPCSALLSSDITASVNCAK
KIASSGNGMNAWVAWRNRCKGTDVHAWirGCRL

KGYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESNFTNTHATNRNTD
-----HHHHHHHHH----------------------EEEEEE----HHHHHHHHH--
GSTDYGILQINSRWWCNDGRTPGSKNLCNIPCSALLSSDITASVNCAK
-----EEEEEEEEEEEEEEEEEEEEEEEE----------------------------HHHHHHH
KIASSGNGMNAWVAWRNRCKGTDVHAWirGCRL
HHH-------EEE------------------------EEEE-----
Classification

• **Classification**: Classify each AA into one of three classes

• Classification is a **fundamental problem**
  - Classify the readout of a microarray as diseased / healthy
  - Classify a subsequence of a genome as coding / non-coding
  - Classify an email as spam / no spam

• Many **different techniques**: Naïve Bayes, Regression, Decision Trees, SVMs, Neural Networks, …
  - Classification function learned from properties of known objects
  - Often use same representation (feature vectors) of objects – methods exchangeable

• The following is a rather unsystematic approach
  - But simple to explain and classical for this application
This Lecture

• Introduction
• Predicting Protein Secondary Structure
  – Secondary structure elements
  – Chou-Fasman
  – Other methods
Chou-Fasmann Algorithm

- Observation: Different AA favor different folds
  - Different AA are more or less often in H, E, C
  - Different AA are more or less often within, starting, or ending a stretch of H, E, C

- Chou-Fasman algorithm (rough idea)
  - Classifies each AA into E or H; unclassified AA are assigned C
  - Compute a score for the probability of any AA to be E (H)
  - Basis: Relative frequencies in a set of sequences with known SSE
  - In principle, assigns each AA its most frequent class
  - Add several heuristic constraints like minimal length of stretches or tendency to build ungapped sequences
    - Rather CEEEEEECC, not CCECEECCECE
Details [sketch, some heuristics omitted]

• Let \( f_{j,k} \) be the relative frequency of observing \( AA_j \) in class \( k \)
• Let \( f_k \) be the average over all 20 \( f_{j,k} \) values
• Compute the propensity \( P_{j,k} \) of \( AA_j \) to be part of class \( k \) as
  \[
  P_{j,k} = \frac{f_{j,k}}{f_k}
  \]
• Using \( P_{j,k} \), classify each \( AA_j \) for every class \( k \) into
  - Strong, normal, weak builder \((H_\alpha, h_\alpha, l_\alpha, H_\beta, h_\beta, l_\beta)\)
    • Tendency to build a SS-element
  - Strong, weak breaker \((B_\alpha, b_\alpha, B_\beta, b_\beta)\)
    • Tendency to stop a SS-element
  - Indifferent \((i_\alpha, i_\beta)\)
Concrete Values

- Originally computed on only 15 proteins (1974)

<table>
<thead>
<tr>
<th>AS</th>
<th>$P_\alpha$</th>
<th>Klasse</th>
<th>AS</th>
<th>$P_\beta$</th>
<th>Klasse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glu</td>
<td>1.53</td>
<td>$h_\alpha$</td>
<td>Met</td>
<td>1.67</td>
<td>$h_\beta$</td>
</tr>
<tr>
<td>Ala</td>
<td>1.45</td>
<td></td>
<td>Val</td>
<td>1.65</td>
<td></td>
</tr>
<tr>
<td>Leu</td>
<td>1.34</td>
<td></td>
<td>Ile</td>
<td>1.60</td>
<td></td>
</tr>
<tr>
<td>His</td>
<td>1.24</td>
<td></td>
<td>Cys</td>
<td>1.30</td>
<td></td>
</tr>
<tr>
<td>Met</td>
<td>1.20</td>
<td></td>
<td>Tyr</td>
<td>1.29</td>
<td></td>
</tr>
<tr>
<td>Gln</td>
<td>1.17</td>
<td>h_\alpha</td>
<td>Phe</td>
<td>1.28</td>
<td></td>
</tr>
<tr>
<td>Trp</td>
<td>1.14</td>
<td></td>
<td>Gln</td>
<td>1.23</td>
<td>h_\beta</td>
</tr>
<tr>
<td>Val</td>
<td>1.14</td>
<td></td>
<td>Leu</td>
<td>1.22</td>
<td></td>
</tr>
<tr>
<td>Phe</td>
<td>1.12</td>
<td></td>
<td>Thr</td>
<td>1.20</td>
<td></td>
</tr>
<tr>
<td>Lys</td>
<td>1.07</td>
<td>$l_\alpha$</td>
<td>Trp</td>
<td>1.19</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AS</th>
<th>$P_\alpha$</th>
<th>Klasse</th>
<th>AS</th>
<th>$P_\beta$</th>
<th>Klasse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ile</td>
<td>1.00</td>
<td>$l_\alpha$</td>
<td>Ala</td>
<td>0.93</td>
<td>$l_\beta$</td>
</tr>
<tr>
<td>Asp</td>
<td>0.98</td>
<td></td>
<td>Arg</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>Thr</td>
<td>0.82</td>
<td></td>
<td>Gly</td>
<td>0.81</td>
<td>i_\beta</td>
</tr>
<tr>
<td>Ser</td>
<td>0.79</td>
<td>i_\alpha</td>
<td>Asp</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>Arg</td>
<td>0.79</td>
<td></td>
<td>Lys</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>Cys</td>
<td>0.77</td>
<td></td>
<td>Ser</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>Asn</td>
<td>0.73</td>
<td>b_\alpha</td>
<td>His</td>
<td>0.71</td>
<td>b_\beta</td>
</tr>
<tr>
<td>Tyr</td>
<td>0.61</td>
<td></td>
<td>Asn</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>Pro</td>
<td>0.59</td>
<td></td>
<td>Pro</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>Gly</td>
<td>0.53</td>
<td>B_\alpha</td>
<td>Glu</td>
<td>0.26</td>
<td>B_\beta</td>
</tr>
</tbody>
</table>
Algorithm for Helices

- Go through the protein sequence
- **Score each AA** with 1 ($H_\alpha$, $h_\alpha$), 0.5 ($I_\alpha$, $i_\alpha$), or -1 ($B_\alpha$, $b_\alpha$)
- Find **helix cores**: subsequences of length 6 with an aggregated AA score $\geq 4$
- Starting from the middle of each core, shift a **window of length 4** to the left (then to the right)
  - Compute aggregated score $A$ using values $P_{j,k}$ inside the window
    - Not the coarse-grained H/I/B/I values
    - If $A \geq 4$, continue; otherwise stop
- Similar method for strands
- **Conflicts** (regions assigned both H and E) are resolved based on aggregated scores
Example [Source: O. Kohlbacher, “Strukturvorhersage”]

\[ \sum = 5 \]

Helixstart

\[ 4.3 / 4 > 1.0 \]

\[ 3.6 / 4 < 1.0 \]

\[ 4.5 / 4 > 1.0 \]

\[ 4.1 / 4 > 1.0 \]

\[ 3.2 / 4 < 1.0 \]
Performance

• Accuracy app. 50-60%
  - Measured on per-AA correctness

• Prediction is more accurate in helices than in strands
  - Because helices build local bridges (hydrogen bounds between the turns; each AA binds to the $+4$ AA)

• General problem
  - Secondary structure is not only a local problem
  - Looking only at single AAs is not enough
    • Note: Scores are based on individual AA; aggregation by summation assumes statistical independence of pairs, triples ... in a class

• One needs to include the context of an AA
This Lecture

- Introduction
- Predicting Protein Secondary Structure
  - Secondary structure elements
  - Chou-Fasman
  - Other methods
Classes of Methods

• First generation: Properties of single AA only
  - Accuracy: 50-60%, e.g. Chou-Fasman (1974)

• Second generation: Include info. about neighborhood
  - Accuracy: ~65%, e.g. GOR (1974 – 1987)

• Third generation: Include info. from homologous seq’s
  - Accuracy: ~70-75%, w.g. PHD (1994)

• Forth generation: Build ensembles of good methods
  - Accuracy: ~80%, e.g. Jpred (1998)

• Current performance
  - Jpred 4 (2015): 82% overall, ~90% for certain other properties
  - Spine-X (2012): 84% overall
Further Reading
