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

DNA \rightarrow RNA \rightarrow \text{protein}

- Replication: vertical information flow
- Expression: horizontal information flow

<table>
<thead>
<tr>
<th></th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>UUU</td>
<td>UUC</td>
<td>UAU</td>
<td>UGA</td>
</tr>
<tr>
<td></td>
<td>UUA</td>
<td>UUG</td>
<td>UAG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Phenylalanine</td>
<td>Leucine</td>
<td>Serine</td>
<td>Tyrosine</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>CUU</td>
<td>CUC</td>
<td>CAU</td>
<td>CUG</td>
</tr>
<tr>
<td></td>
<td>CUA</td>
<td>CUG</td>
<td>CAG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Leucine</td>
<td>Leucine</td>
<td>Proline</td>
<td>Histidine</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>AUU</td>
<td>AUC</td>
<td>AAA</td>
<td>AUG</td>
</tr>
<tr>
<td></td>
<td>AUG</td>
<td></td>
<td>AAG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Isoleucine</td>
<td>Methionine; initiation codon</td>
<td>Threonine</td>
<td>Asparagine</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>GUU</td>
<td>GUC</td>
<td>GAU</td>
<td>GUG</td>
</tr>
<tr>
<td></td>
<td>GUA</td>
<td>GUG</td>
<td>GAG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Valine</td>
<td>Alanine</td>
<td>Aspartic acid</td>
<td>Glycine</td>
</tr>
</tbody>
</table>

Ulf Leser: Bioinformatics, Summer Semester 2012
Details

- **Alternative Splicing**
  - “One gene – one protein” is wrong
  - Exons may be spliced from the protein sequence
  - Human: \(~6\) times more proteins than genes

- **Post-translational modifications**
  - (De-)Phosporylation, glycolysation, cleavage of signals, …
  - Rough estimates: 100K proteins, 500K protein forms

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

- **Very large complexes** present in all eukaryotes (and more species)
  - >2000 kDa, made of **dozens of single protein chains**
  - Formation of the complex is a complex process only partly understood yet

- **Breaks** (mis-folded, broken, superfluous, ...) proteins into small peptides for reuse
  - Tagging with ubiquitin which binds to the proteasome
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 ~1/3 of all human gene, no function is known
- Describing function
  - **Gene Ontology**: 3 branches, >30,000 concepts
  - Used world-wide to describe gene/protein function
Protein Interactions and Networks

- Function usually is carried out by a complex interplay between different proteins and other molecules
- **Pathways**: (artificial) fragments of the cellular network associated to a certain function
  - See lectures later
Function and Motifs

• Protein often have multiple functions
  – Avg. number of GO terms assigned to a human protein: ~6
• 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
Protein Families and Classification

- Several DBs classify proteins according to their **overall structure** (CATH, SCOP, FSSP)
- Highly related to **evolutionary relationships** (and function)
- Example: CATH
  - Class (all $\alpha$, all $\beta$, $\alpha$-$\beta$, other)
  - Architecture
  - Topology (similar folds)
  - Homologous superfamily (sets of concrete proteins)
- Helpful to characterize novel proteins
Functional Classification

- Folds correlate with function, but many exceptions
- **Enzyme classification** (EC-numbers)
  - 4-level hierarchy
  - Based on chemical reactions that are catalyzed
  - Closer related to function than classes of folds
  - Relation protein:EC-number is mostly 1:1
- EC-number and GO-annotation are highly correlated
  - But >30,000 concepts versus <4000 EC-numbers
Proteomics – Large Scale Protein Identification

• Measuring gene expression: RNA-Seq, microarrays, PCR, ...
• Measuring proteins is much harder
  – Isolating proteins is very complex
  – Sequencing a protein is very slow
• Options (next lecture)
  – Isolation: 2D-Page, chromatography, ...
  – Identification: Mass spectrometry
  – De-dovo sequencing with MS/MS
  – Quantification is very difficult
• Some classes of proteins are particularly hard to handle
  – Membrane proteins, non-soluble proteins
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 annotation set

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

- 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)

...
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

• An AA consists of a common core and a specific residue
  - Amino group – NH₂
  - Central C\textsubscript{α} - Carbon – CH
  - Carboxyl group – COOH
• Residues (side chains) vary greatly between AA
• Residues determine the specific properties of a AA
Structure of a Protein

- The core forms the backbone of a protein (AA chain)
- 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 proteins things are much more restricted
  - Peptide bound 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-Helixes**

- 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
β-Sheets

- Two linear and parallel stretches (β-strands)
- Strands are bound together by hydrogen bonds
- Can be parallel or anti-parallel (wrt. N/C terminus)
Other Substructures

• α-helixes and β-sheets form around 50-80% of a protein
• 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

• 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 associated to function
• Knowing the SSE gives important clues to protein function
• Secondary structure prediction (SSP) is 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 (_)

\[
\text{KVYGRCELAAMKRLGLDNYRGYS} \text{LGNWVCAAKFESNFNTATNRNTD}
\]

\[
\text{GSTDYGILQINSRWCC} \text{NDGRTPGSKNCINPC} \text{SA} \text{LLSSDI} \text{TASVNCAK}
\]

\[
\text{KIAGGNGMN} \text{AWA} \text{RNRC} \text{KGTDVHAWIRGCR} \text{L}
\]
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, ...
  - Based on same principles can be exchanged easily
  - Classification function learned from properties of known objects
• 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 constraints about minimal length of E (H) stretches
  - Several further heuristics
Some 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 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 (\( H_\alpha, h_\alpha, I_\alpha \))
  - Strong, weak breaker (\( B_\alpha, b_\alpha \))
  - Indifferent (\( i_\alpha \))
- For now, context (neighboring AAs) is ignored completely.
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>0.53</td>
<td>$H_\alpha$</td>
<td>Met</td>
<td>0.67</td>
<td>$H_\beta$</td>
</tr>
<tr>
<td>Ala</td>
<td>1.45</td>
<td>$H_\alpha$</td>
<td>Val</td>
<td>1.65</td>
<td>$H_\beta$</td>
</tr>
<tr>
<td>Leu</td>
<td>1.34</td>
<td>$H_\alpha$</td>
<td>Ile</td>
<td>1.60</td>
<td>$H_\beta$</td>
</tr>
<tr>
<td>His</td>
<td>1.24</td>
<td>$h_\alpha$</td>
<td>Cys</td>
<td>1.30</td>
<td>$i_\alpha$</td>
</tr>
<tr>
<td>Met</td>
<td>0.20</td>
<td>$h_\alpha$</td>
<td>Tyr</td>
<td>1.29</td>
<td>$i_\beta$</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>$H_\alpha$</td>
<td>Gln</td>
<td>1.23</td>
<td>$h_\beta$</td>
</tr>
<tr>
<td>Val</td>
<td>1.14</td>
<td>$H_\alpha$</td>
<td>Leu</td>
<td>1.22</td>
<td></td>
</tr>
<tr>
<td>Phe</td>
<td>1.12</td>
<td>$H_\alpha$</td>
<td>Thr</td>
<td>1.20</td>
<td></td>
</tr>
<tr>
<td>Lys</td>
<td>1.07</td>
<td>$I_\alpha$</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>$I_\alpha$</td>
<td>Ala</td>
<td>0.93</td>
<td>$I_\beta$</td>
</tr>
<tr>
<td>Asp</td>
<td>0.98</td>
<td>$I_\alpha$</td>
<td>Arg</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>Thr</td>
<td>0.82</td>
<td>$I_\alpha$</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>$i_\alpha$</td>
<td>Lys</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>Cys</td>
<td>0.77</td>
<td>$i_\alpha$</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>$b_\alpha$</td>
<td>Asn</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>Pro</td>
<td>0.59</td>
<td>$B_\alpha$</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 ($l_\alpha, i_\alpha$), 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
  - 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”]

\[ \Sigma = 5 \]

Helixstart
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: Only look at properties of single AA
  - 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%
  - E.g. PHD (1994)

• Forth generation: Build ensembles of good methods
  • Accuracy: ~80%
  • E.g. Jpred (1998)
GOR Algorithm


• In principle, GOR uses $P_k$ values for **16-grams of AAs**
  - Recall: Helices have a “reach” of ~4 AA
  - But neighbors in a strand can be arbitrarily far away from each other (but they are not in practice)

• These cannot be learned from counting
  - There are $16^{20} \sim 1.2E24$ different 16-grams

• Different versions of GOR use different methods to estimate these values

• Other difference
  - Use of negative information (chances of AA j not being in class k)
  - No cores+extension: Each AA is classified based on its 16-context
PHD [Rost et al.: PHD-an automatic mail server for protein SSP, Bioinformatics 10(1), 1994]

- **Uses two neural networks**
- Input is not only the AA and its context, but its *profile*
  - Given the input sequence X, PHD first search *homologous sequences* (using PSI-BLAST)
  - All these are subjected to a multiple sequence alignment
  - The “column” of an AA in X is its profile
- **Rationale**
  - On average, one can exchange ~65% of a protein without changing its secondary structure notably
  - Thus, the concrete AA is not as important as one might think
  - Homologous sequences are believed to have the same function and the same secondary structure
  - We do not classify the AA, but the list of all its replacements
Further Reading