Proteins: Structure & Function

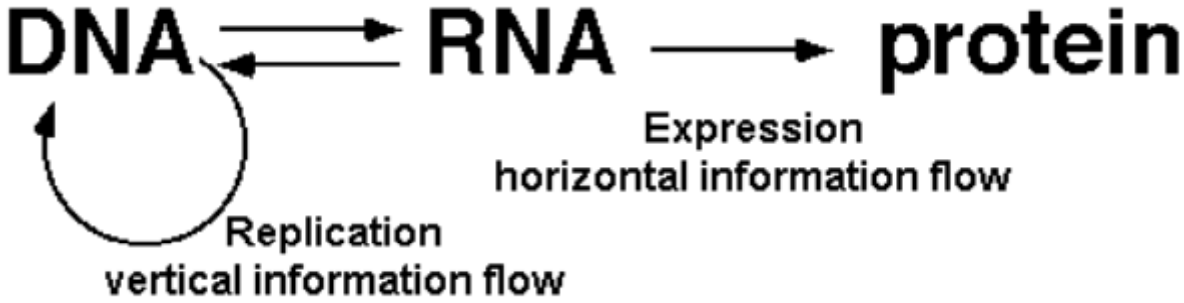
Ulf Leser

This Lecture

- Introduction
  - Structure
  - Function
  - Databases
- Predicting Protein Secondary Structure

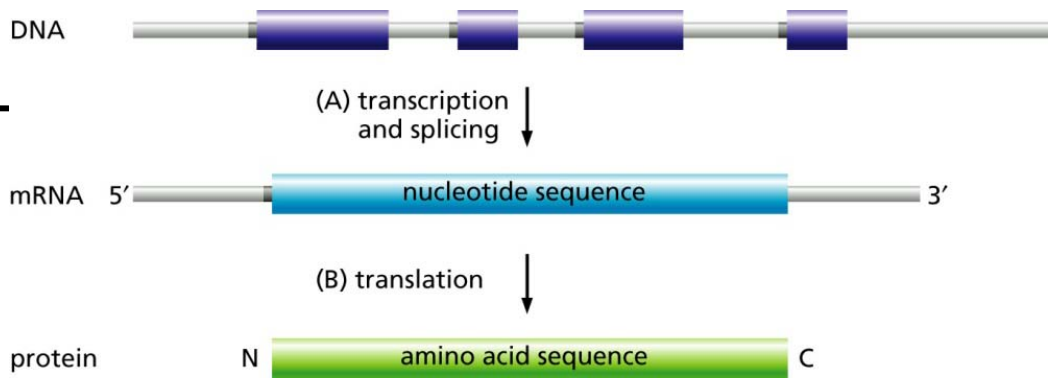
- Many figures from Zvelebil, M. and Baum, J. O. (2008). "Understanding Bioinformatics", Garland Science, Taylor & Francis Group.
- Examples often from O. Kohlbacher, Vorlesung Strukturvorhersage, WS 2004/2005, Universität Tübingen

Central Dogma of Molecular Biology



	U		C		A		G	
U	UUU UUC	Phenyl-alanine	UCU UCC UCA UCG	Serine	UAU UAC	Tyrosine	UGU UGC	Cysteine	U C A G
	UUA UUG	Leucine			UAA	Stop codon	UGA	Stop codon	
					UAG	Stop codon	UGG	Tryptophan	
C	CUU CUC CUA CUG	Leucine	CCU CCC CCA CCG	Proline	CAU CAC	Histidine	CGU CGC CGA CGG	Arginine	U C A G
					CAA CAG	Glutamine			
A	AUU AUC AUA	Isoleucine	ACU ACC ACA ACG	Threonine	AAU AAC	Asparagine	AGU AGC	Serine	U C A G
	AUG	Methionine; initiation codon			AAA AAG	Lysine	AGA AGG	Arginine	
G	GUU GUC GUA GUG	Valine	GCU GCC GCA GCG	Alanine	GAU GAC	Aspartic acid	GGU GGC GGA GGG	Glycine	U C A G
					GAA GAG	Glutamic acid			

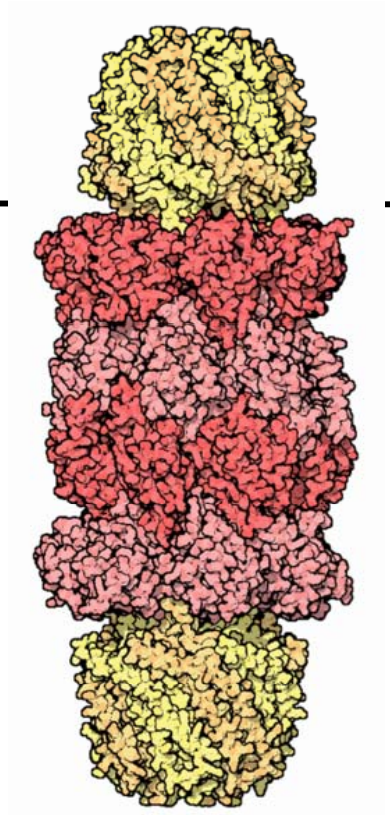
The Real Picture

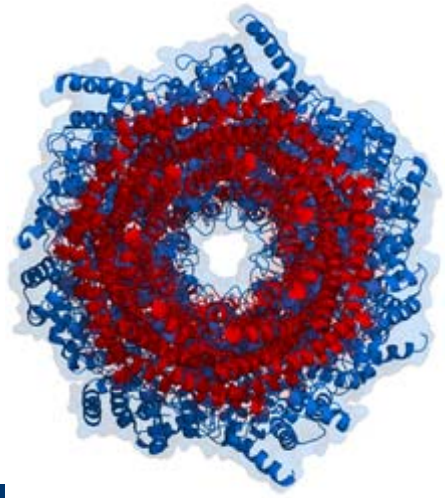


- Alternative Splicing
  - "One gene – one protein" is wrong
  - Exons may be spliced from the protein sequence
  - Human: ~ 6 times more proteins than genes
- Post-translational modifications
  - (De-)Phosporylation, glycolysation, cleavage of signals, …
  - Estimates: 100K proteins, 500K protein forms
- Complexes
  - Proteins gather together to perform specific function

Example: Proteasome

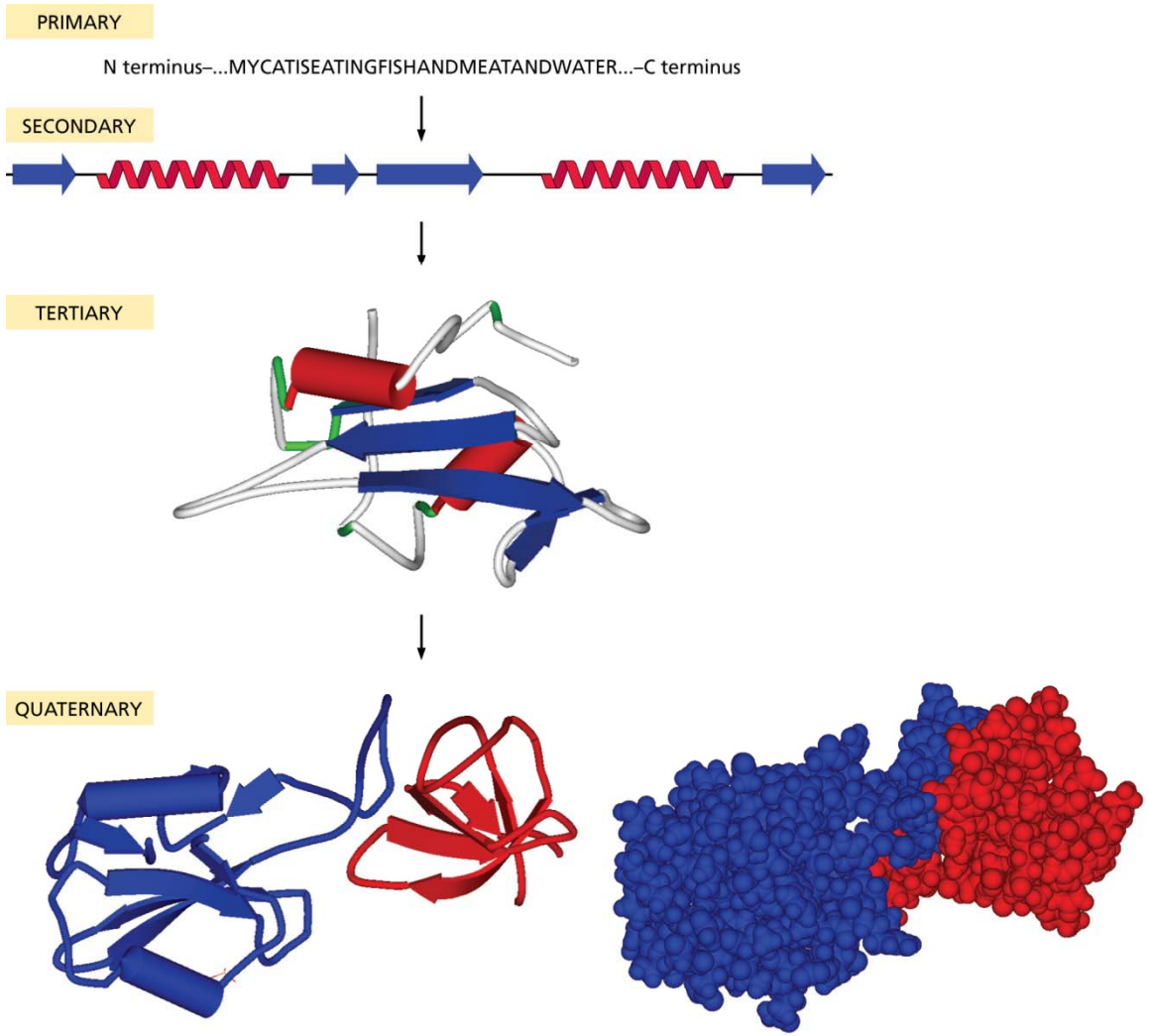
- Very large complexes present in all eukaryotes (and more)
  - >2000 kDa, made of dozens of single protein chains
  - Formation of the complex is a very complex process only partly understood yet
- Breaks (mis-folded, broken, superfluous, …) proteins into small peptides for reuse
  - Suspicious proteins are tagged with ubiquitin which binds to the proteasome





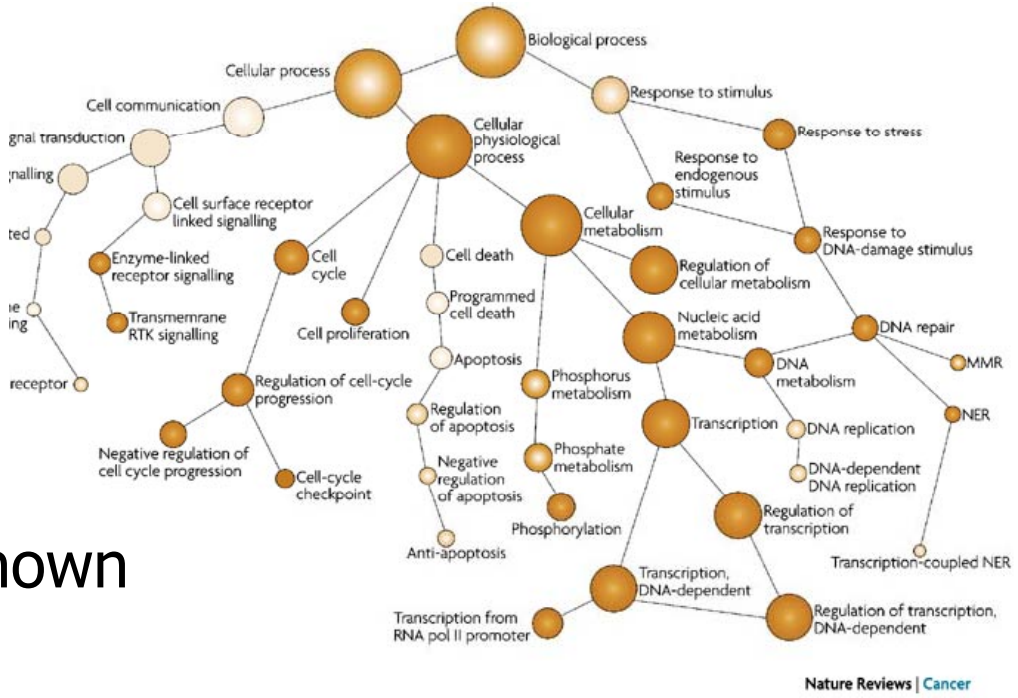
Protein Structure

- Primary
  - 1D-Seq. of AA
- Secondary
  - 1D-Seq. of "subfolds"
- Tertiary
  - 3D-Structure
- Quaternary
  - 3D-Structure of assembled complexes



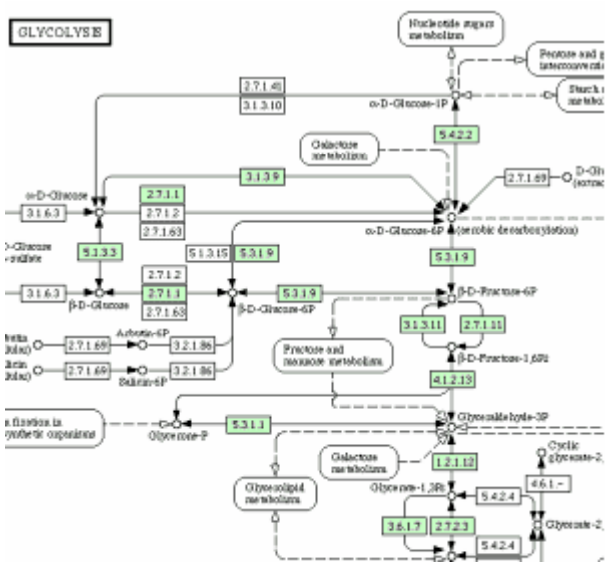
Protein Function

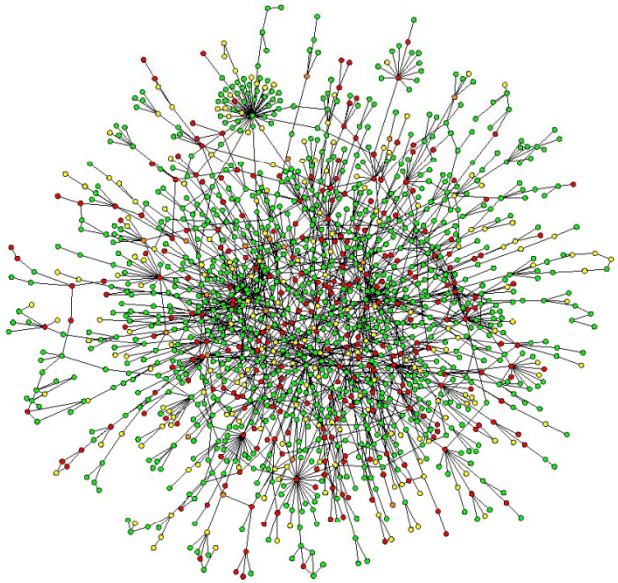
- Proteins perform essentially everything that makes an organism alive
  - Metabolism
  - Signal processing
  - Gene regulation
  - Cell cycle
  - …
- For ~1/3 of all human gene, no function is known
- Describing function
  - Gene Ontology: 3 branches, >30.000 concepts
  - Used world-wide to describe gene/protein function



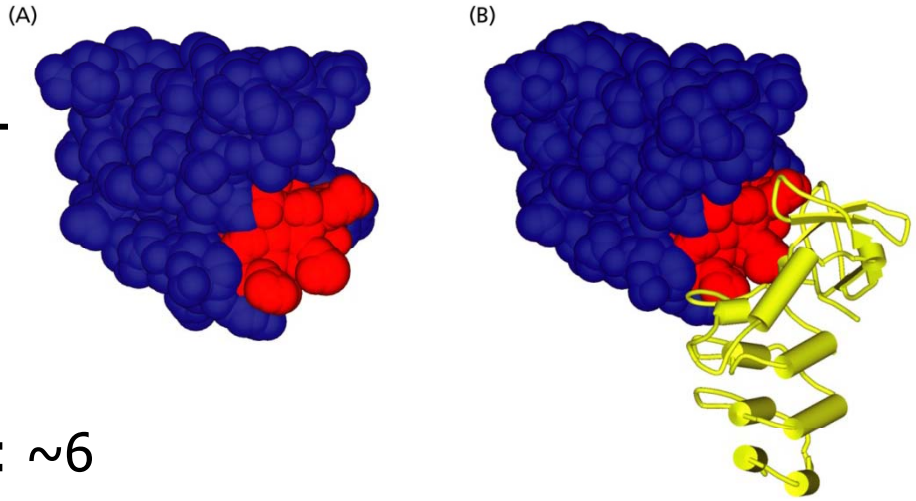
Protein Interactions and Networks

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





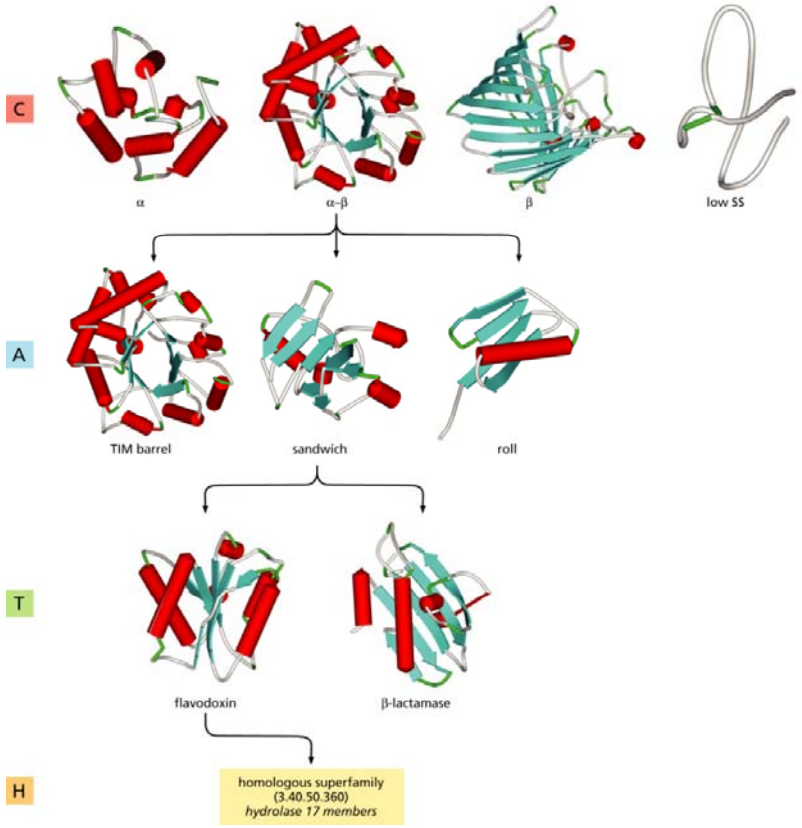
Function and Motifs



- A protein may "have" many functions
  - Avg. n# of GO terms assigned to a human protein: ~6
- Functions are carried out by substructures of a protein
  - Called motifs or domains
- There probably exist only 4000-5000 motifs
  - Proteins: "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
  - Blocking such bindings is a major goal in 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 α, all β, α-β, 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

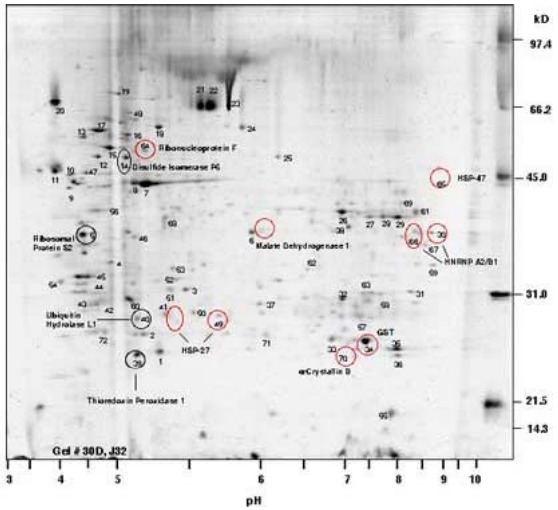
CLASSIFICATION OF ENZYMES

Group of Enzyme	Reaction Catalysed	Examples
1. Oxidoreductases	Transfer of hydrogen and oxygen atoms or electrons from one substrate to another.	Dehydrogenases Oxidases
2. Transferases	Transfer of a specific group (a phosphate or methyl etc.) from one substrate to another.	Transaminase Kinases
3. Hydrolases	Hydrolysis of a substrate.	Estrases Digestive enzymes
4. Isomerases	Change of the molecular form of the substrate.	Phospho hexo isomerase, Fumarase
5. Lyases	Nonhydrolytic removal of a group or addition of a group to a substrate.	Decarboxylases Aldolases
6. Ligases (Synthetases)	Joining of two molecules by the formation of new bonds.	Citric acid synthetase

- EC-number and GO-annotation are highly correlated
  - But >30.000 concepts versus <4000 EC-numbers

Proteomics – Large Scale Protein Identification

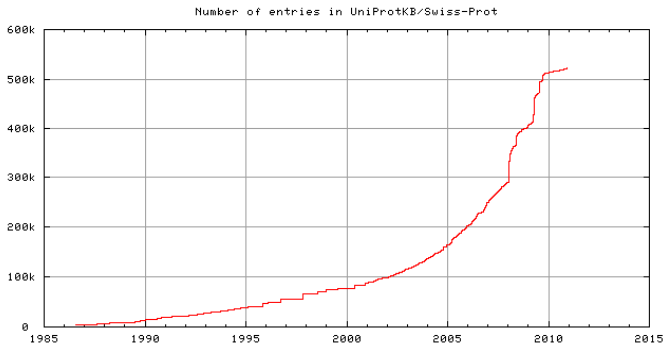
- Measuring gene expression is comparably simple
  - Sequencing mRNA, microarrays (based on hybridization)
- Measuring proteins is much harder
  - Isolating proteins is very complex
  - Sequencing a protein is very slow
- Options (see lecture later)
  - Using 2D-Page
  - Using mass spectrometry
  - De-dovo sequencing with MS/MS
  - Quantification is very difficult



- Some classes of proteins are particularly hard to handle
  - Membrane proteins, non-soluble proteins

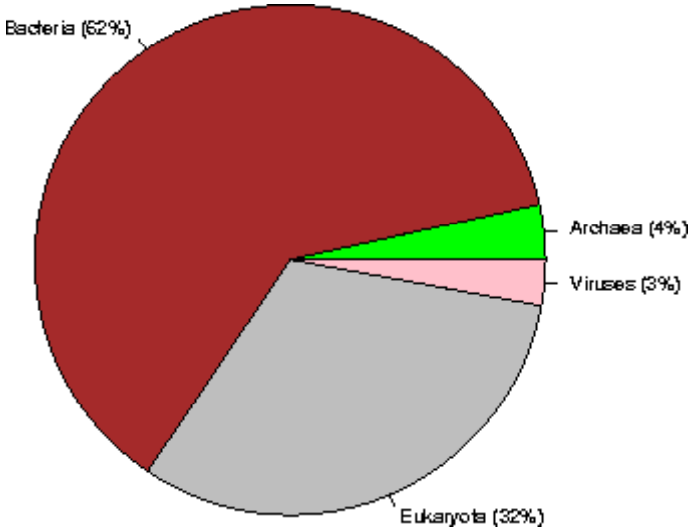
UniProt

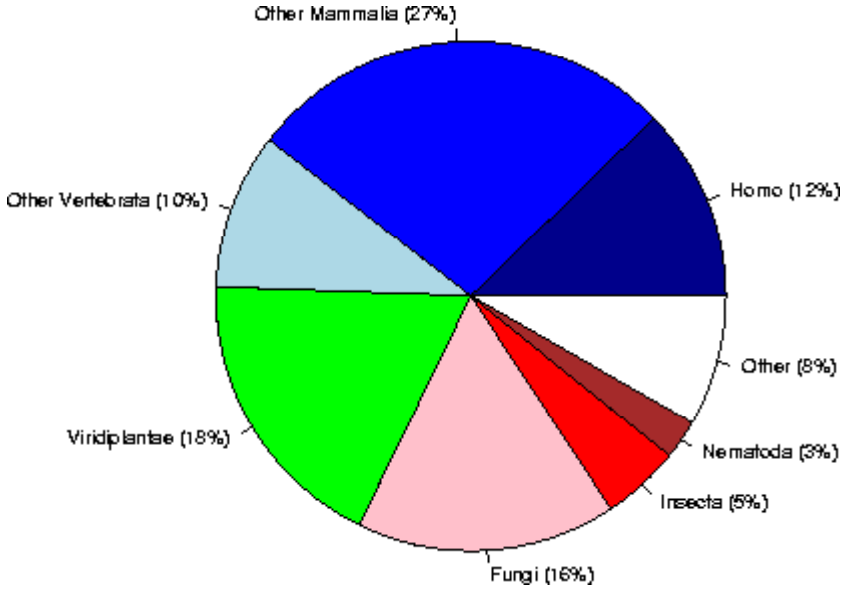
- "Standard" database for protein sequences and annotation
  - Original name: SwissProt
  - Started at the Swiss Institute of Bioinformatics, now mostly EBI
  - Other: PIR, HPRD
- Continuous growth and curation
  - >30 „Scientific Database Curators"
  - Quarterly releases
  - Very rich annotation set
  - Redundancy-free



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

UniProt: Species [http://www.expasy.org/sprot/relnotes/relstat.html]



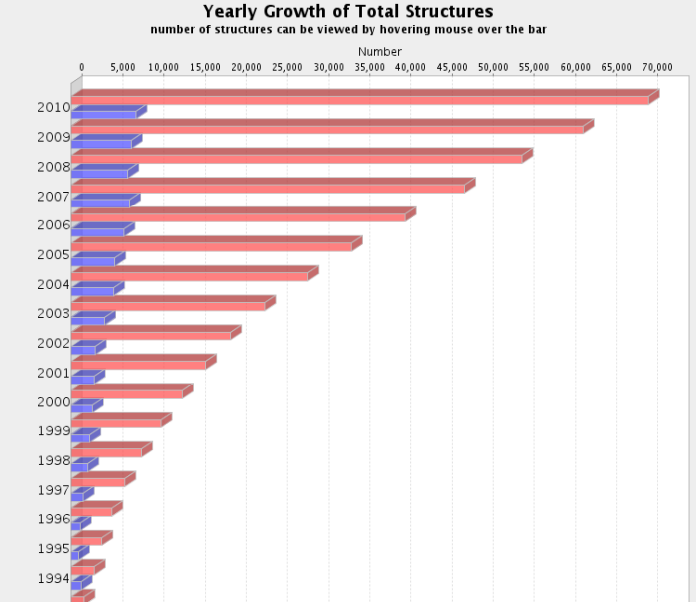


20258	Homo sapiens (Human)
16327	Mus musculus (Mouse)
9842	Arabidopsis thaliana (Mouse-ear cress)
7560	Rattus norvegicus (Rat)
6582	Saccharomyces cerevisiae (Baker's yeast)
5803	Bos taurus (Bovine)

...

PDB – Protein Structure Database

- Oldest protein database, evolved from a book
- Contains all experimentally obtained protein 3D-structures
  - Plus DNA, protein-ligand, complexes, …
  - X-Ray (~75%), NMR (Nuclear magnetic resonance, ~23%)
- Still costly and 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

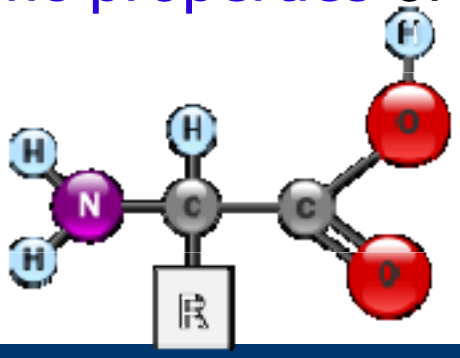
InterPro 30.0

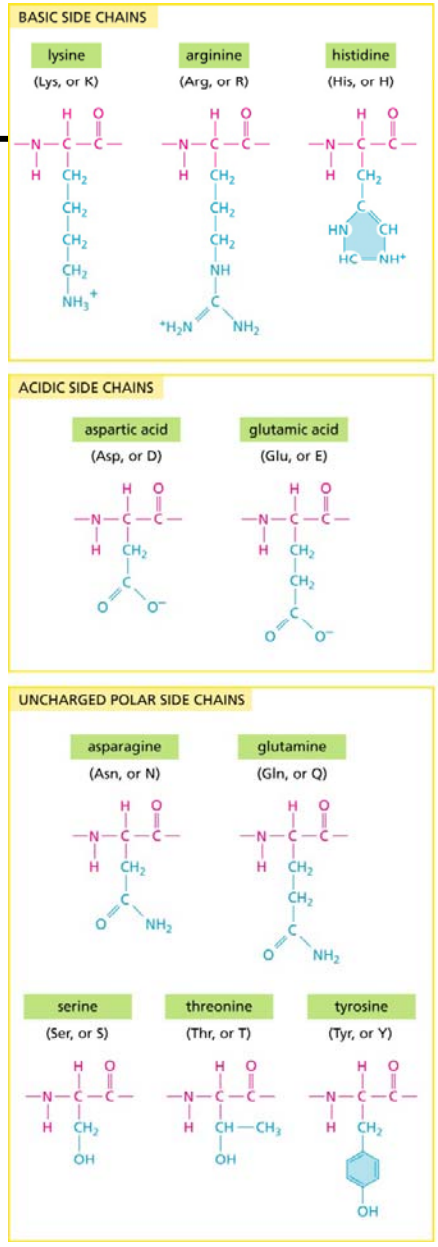
This Lecture

- Introduction
- Predicting Protein Secondary Structure
  - Secondary structure elements
  - Chou-Fasman
  - GOR IV
  - Other methods

Amino Acids

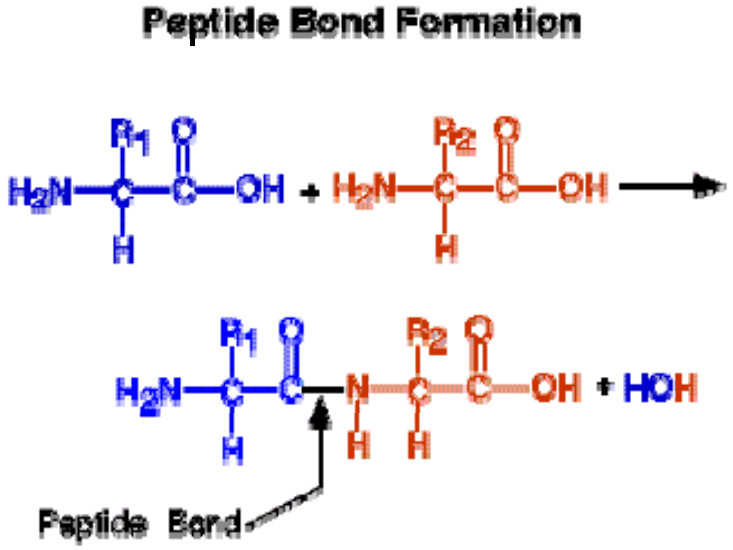
- An AA consists of a common „core" and a specific residue
  - Amino group – NH_2
  - Central C_α - Carbon – CH
  - Carboxyl group – COOH
- Residues (side chains) vary greatly between AA
- Residues determine the specific properties of a AA

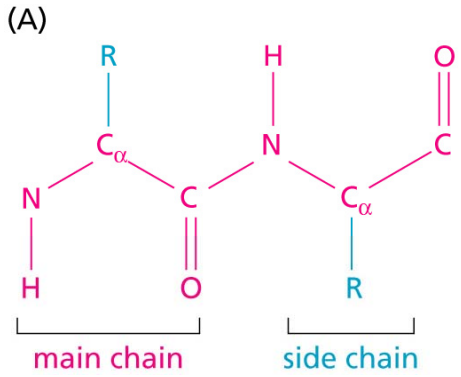




Structure of a Protein

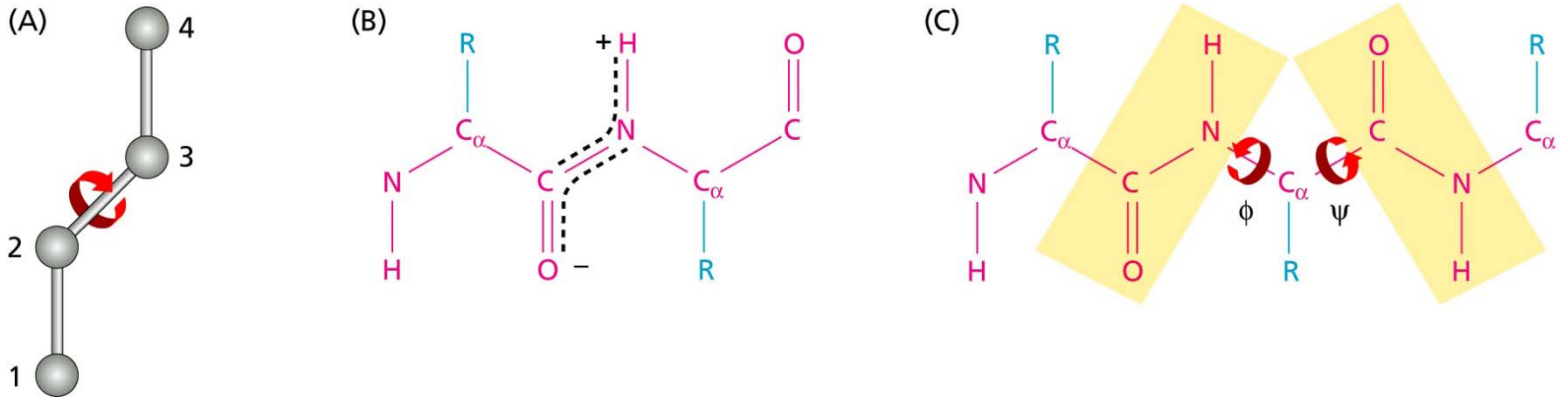
- The core forms the backbone of a protein (AA chain)
- Main structure: Covalent peptide bond between carboxyl and amino group
  - Loss of H_2O





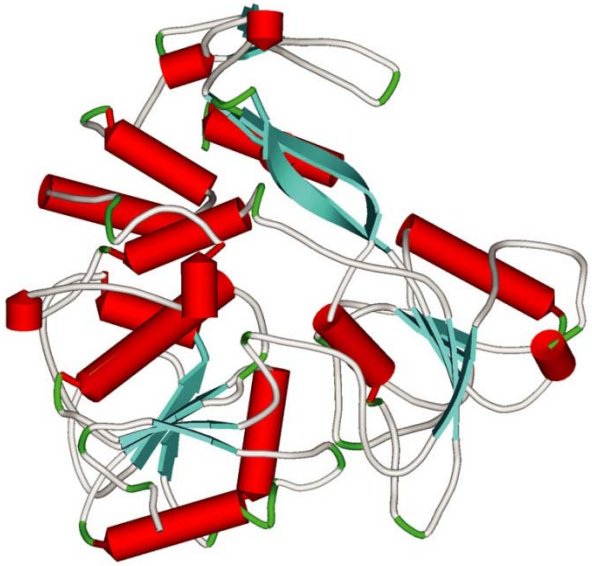
Structure Flexibility

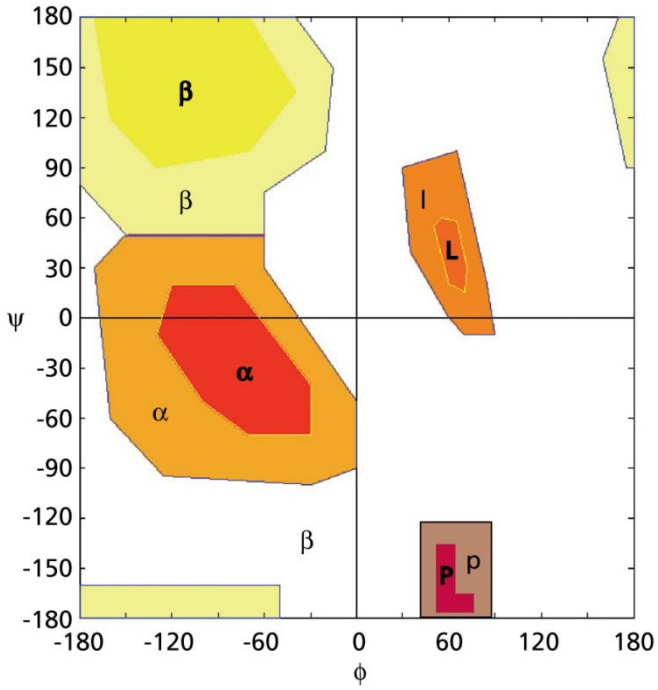
- In principle, every chemical bond can rotate freely, which would allow arbitrary structures to be formed
- In proteins things are much more restricted
  - Peptide bound is "flat" – almost no torsion possible
  - Flexibility only in the C_α-flanking bonds ϕ and ψ



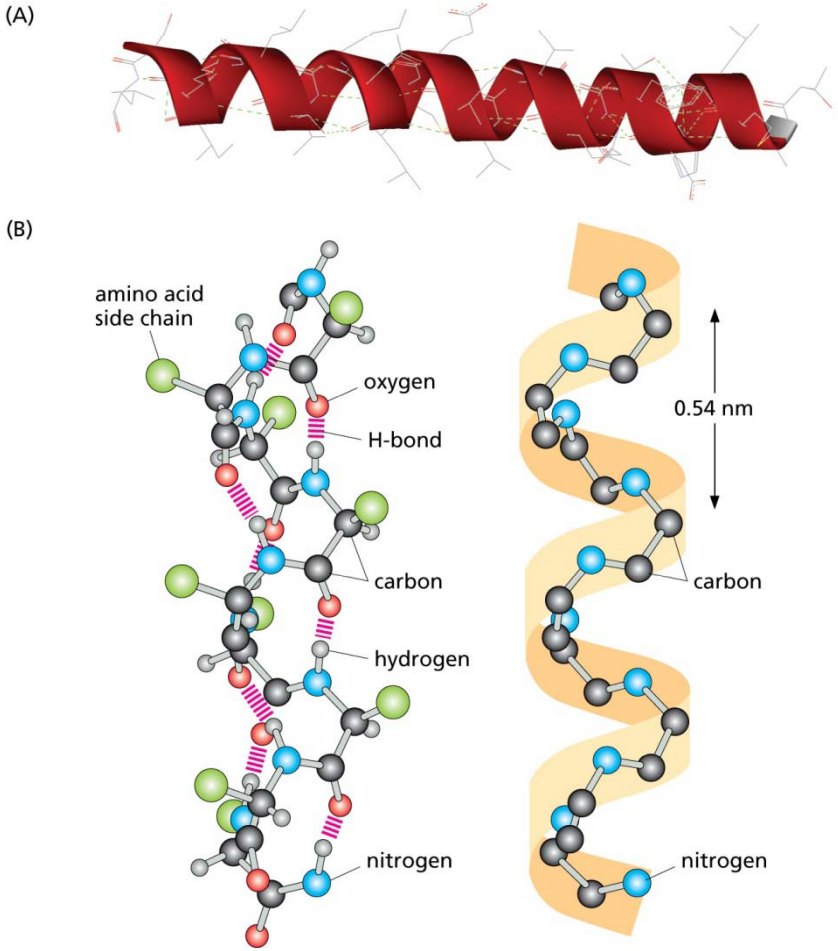
Ramachandran Plots

- Combinations of ϕ and ψ are highly constrained
  - Due to chemical properties of the backbone / side chains
- Two combinations are favored: α-helixes and β-sheets
  - More detailed classifications exist
  - Angels lead to specific structures
  - Secondary structure





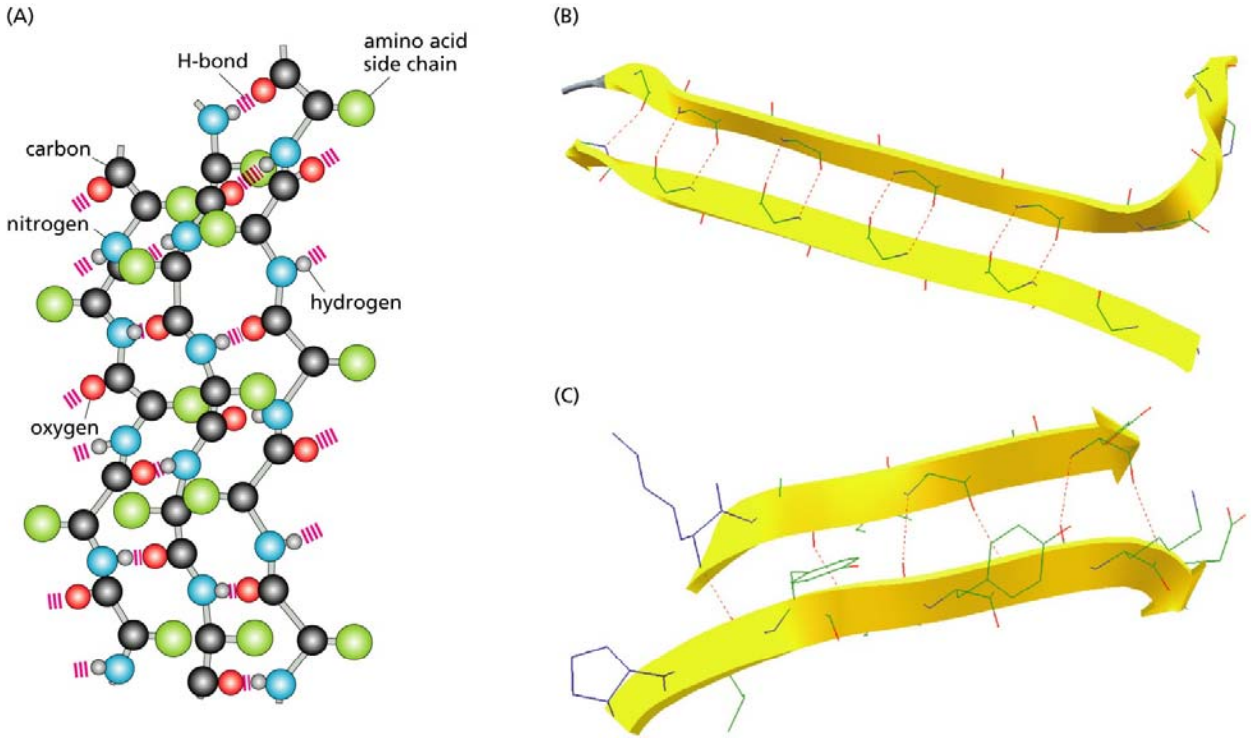
α-Helixes



- Comb. 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

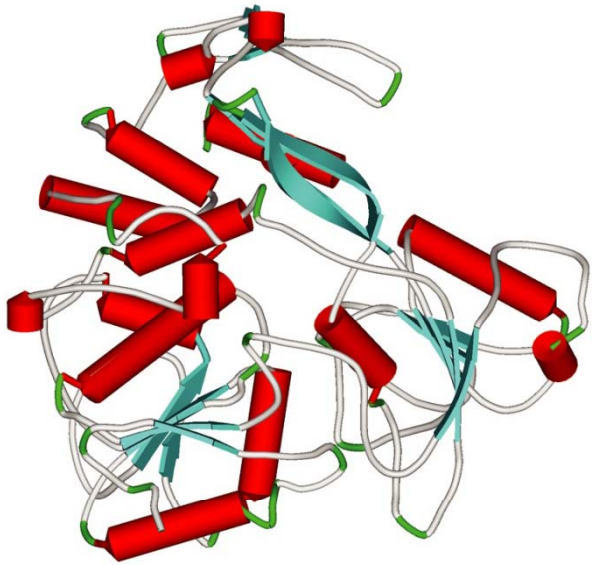
β-Sheets

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



Other Substructures

- α-helixes and β-sheets form around 50-80% of a protein
- All other „glue" parts are called loops or coils
  - Usually not very important for the structure of the protein
  - But very important for its function
  - Loops are often exposed on the surface and participate in binding to other molecules



Importance

- Secondary structure elements (SSE) are vital for the overall structure of a protein
- Thus, they often are evolutionary well conserved
- SSE can be used to classify proteins
  - Such classes are highly associated to function
- Knowing secondary structure thus 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 one of the three classes Helix, E (Strand), or Coil

```
KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESNFNTHATNRNTD
GSTDYGILQINSRWWCNDGRTPGSKNLCNIPCSALLSSDITASVNCAK
KIASGGNGMNAWVAWRNRCKGTDVHAWIRGCRL
```

```
KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESNFNTHATNRNTD
-----HHHHHHHHH-------------EEEEE---------------
GSTDYGILQINSRWWCNDGRTPGSKNLCNIPCSALLSSDITASVNCAK
----EEEEEE--------------------------------HHHHHH
KIASGGNGMNAWVAWRNRCKGTDVHAWIRGCRL
HHH-------EEE--------------------
```

Classification

- Classification: Classify each AA into one of three classes
- Classification is a fundamental problem (not only in bioinformatics)
  - 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
- A wealth of different techniques exist (Machine Learning)
  - Naïve Bayes, Regression, Decision Trees, Support Vector Machines, Neural Networks, Sequential Models, …
- Many of them use the same abstraction and can be exchanged easily
  - Describe objects by features
  - Learn properties of feature values in the different classes
  - Derive a classification function on the features
  - Pre-requisite: Distribution of feature values are at least slightly different in the different classes

This Lecture

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

Chou-Fasmann Algorithm

Chou & Fasman (1974). Prediction of protein conformation. Biochemistry 13

- 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 of assignments in a set of sequences with known secondary structure
  - In principle, assigns each AA its most frequent class
  - Add constraints about minimal length of E, H stretches and several other 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

$$P_{j,k}=f_{j,k}/f_k$$

- Using $P_{j,k}$, classify each AA j for class k according into
  - Strong, normal, weak builder (H_α, h_α, I_α)
  - Strong, weak breaker (B_α, b_α)
  - Indifferent (i_α)
- For now, context (neighboring AAs) is ignored completely

Concrete Values

- Originally computed on only 15 proteins (1974)

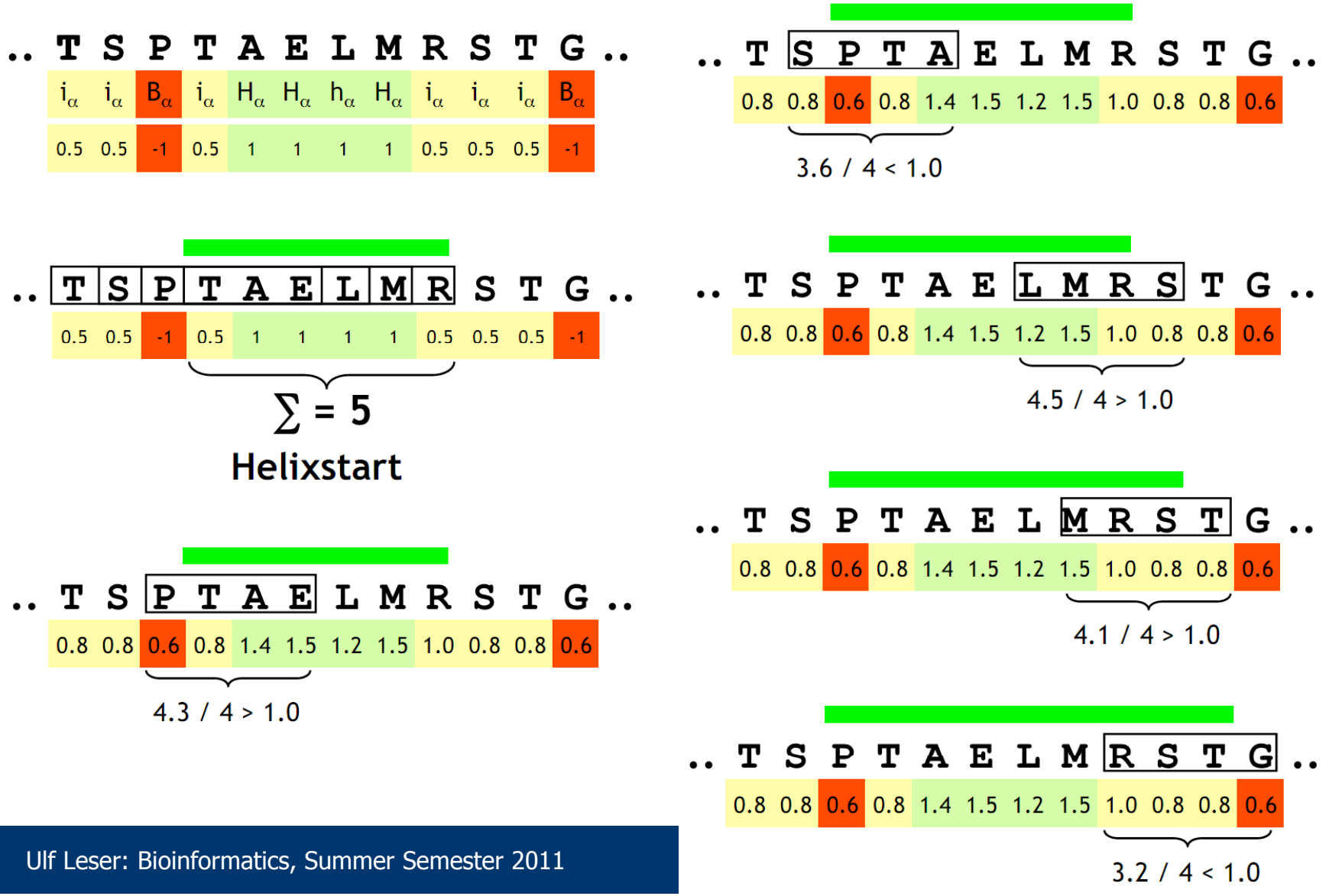
AS	P_α	Klasse	AS	P_β	Klasse
Glu	.53	H_α	Met	.67	H_β
Ala	1.45		Val	1.65	
Leu	1.34		Ile	1.60	
His	1.24	h_α	Cys	1.30	h_β
Met	.20		Tyr	1.29	
Gln	1.17		Phe	1.28	
Trp	1.14		Gln	1.23	
Val	1.14		Leu	1.22	
Phe	1.12		Thr	1.20	
Lys	1.07	I_α	Trp	1.19	

AS	P_α	Klasse	AS	P_β	Klasse
Ile	1.00	I_α	Ala	0.93	I_β
Asp	0.98	i_α	Arg	0.90	i_β
Thr	0.82		Gly	0.81	
Ser	0.79		Asp	0.80	
Arg	0.79		Lys	0.74	b_β
Cys	0.77		Ser	0.72	
Asn	0.73	b_α	His	0.71	
Tyr	0.61		Asn	0.65	
Pro	0.59	B_α	Pro	0.62	
Gly	0.53		Glu	.26	B_β

Algorithm for Helices

- Go through the protein sequence
- Score each AA with 1 (H_α, h_α), 0.5 (I_α, i_α), or -1 (B_α, b_α)
- Find helix cores: subsequences of length 6 with an aggregated AA score ≥ 4
- Starting from the middle of each core, shift a window of length 4 to the left (then to the right)
  - Compute the aggregated P_k-score P inside the window
  - If P ≥ 4, continue; otherwise stop
- Similar method for strands
- Conflicts (regions assigned both H and E) are resolved based on aggregated P_k-scores

Example


.. T S P T A E L M R S T G ..
i_α i_α B_α i_α H_α H_α h_α H_α i_α i_α i_α B_α
0.5 0.5 -1 0.5 1 1 1 1 0.5 0.5 0.5 -1
.. T S P T A E L M R S T G ..
0.5 0.5 -1 0.5 1 1 1 1 0.5 0.5 0.5 -1
∑ = 5
Helixstart
.. T S P T A E L M R S T G ..
0.8 0.8 0.6 0.8 1.4 1.5 1.2 1.5 1.0 0.8 0.8 0.6
4.3 / 4 > 1.0
.. T S P T A E L M R S T G ..
0.8 0.8 0.6 0.8 1.4 1.5 1.2 1.5 1.0 0.8 0.8 0.6
3.6 / 4 < 1.0
.. T S P T A E L M R S T G ..
0.8 0.8 0.6 0.8 1.4 1.5 1.2 1.5 1.0 0.8 0.8 0.6
4.5 / 4 > 1.0
.. T S P T A E L M R S T G ..
0.8 0.8 0.6 0.8 1.4 1.5 1.2 1.5 1.0 0.8 0.8 0.6
4.1 / 4 > 1.0
.. T S P T A E L M R S T G ..
0.8 0.8 0.6 0.8 1.4 1.5 1.2 1.5 1.0 0.8 0.8 0.6
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: 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

[Garnier et al. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins, Jounal of Molecular Biology 120(1), 1978]

- 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}~1.2E24 different such 16-grams
- Different versions of GOR use different methods to estimate these values
  - GOR IV uses propensities of all pairs in this window
- 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

- Gerhard Steger (2003). “Bioinformatik – Methoden zur Vorhersage von RNA- und Proteinstrukturen”, Birkhäuser, chapter 8,10,11,13
- Zvelebil, M. and Baum, J. O. (2008). "Understanding Bioinformatics", Garland Science, Taylor & Francis Group, chapter 2, 11, 12 (partly)