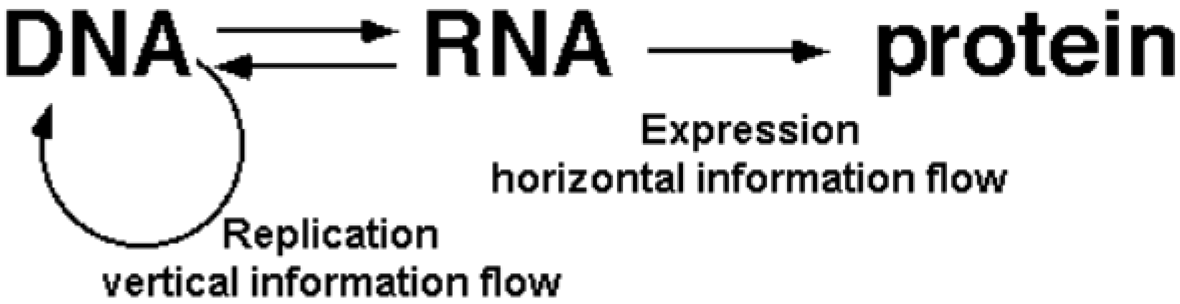
Proteins: Structure & Function

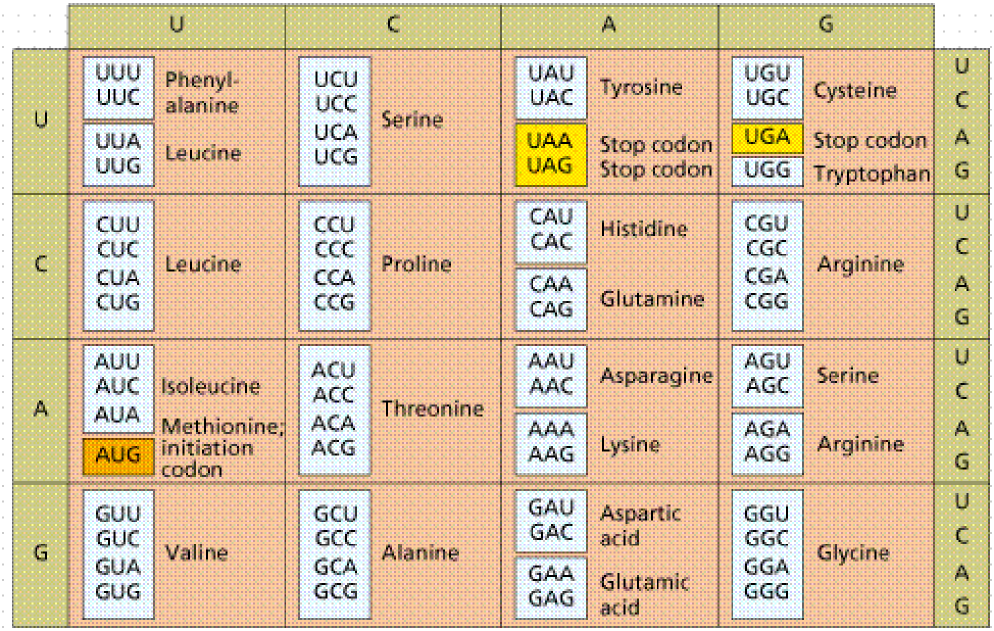
Ulf Leser

This Lecture

- Proteins
  - Structure
  - Function
  - Databases
- Predicting Protein Secondary Structure

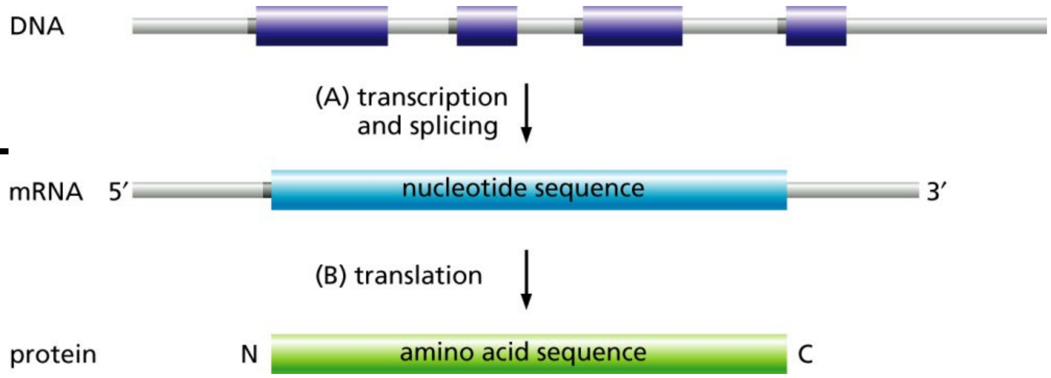
Central Dogma of Molecular Biology





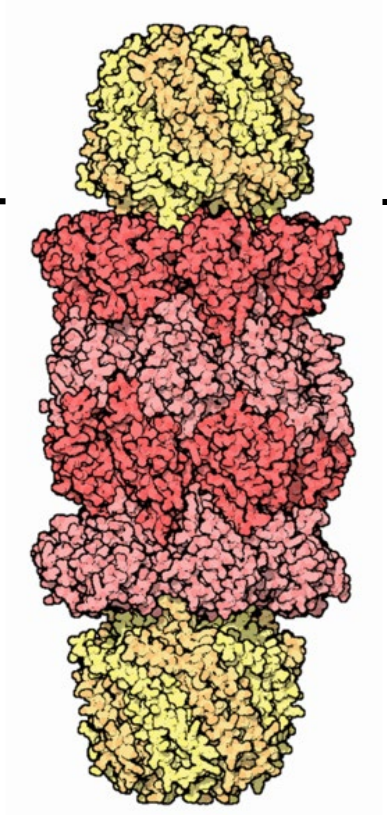
	U	C	A	G	
U	UUU UUC Phenylalanine UUA UUG Leucine	UCU UCC UCA UCG Serine	UAU UAC Tyrosine UAA Stop codon UAG Stop codon	UGU UGC Cysteine UGA Stop codon UGG Tryptophan	U C A G
C	CUU CUC CUA CUG Leucine	CCU CCC CCA CCG Proline	CAU CAC Histidine CAA CAG Glutamine	CGU CGC CGA CGG Arginine	U C A G
A	AUU AUC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

Details

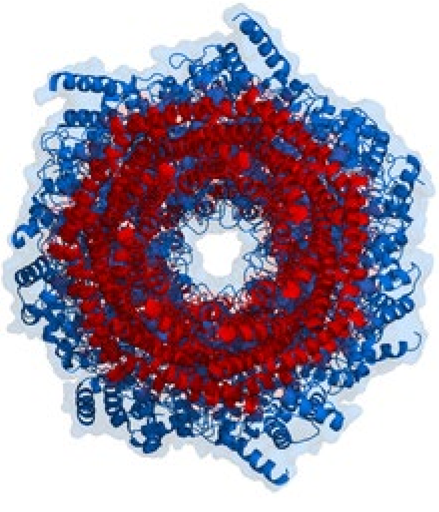


- Alternative Splicing
  - "One gene – one protein" is wrong
  - Exons may be spliced out from the mRNA
  - Human: at least 6 times more unique proteins than genes
    - Also called isoforms
- Post-translational modifications
  - (De-)Phosporylation, glycolysation, cleavage of signal peptides, …
- Complexes: Proteins physically and permanently grouping together to perform a specific function

Example Complex: Proteasome

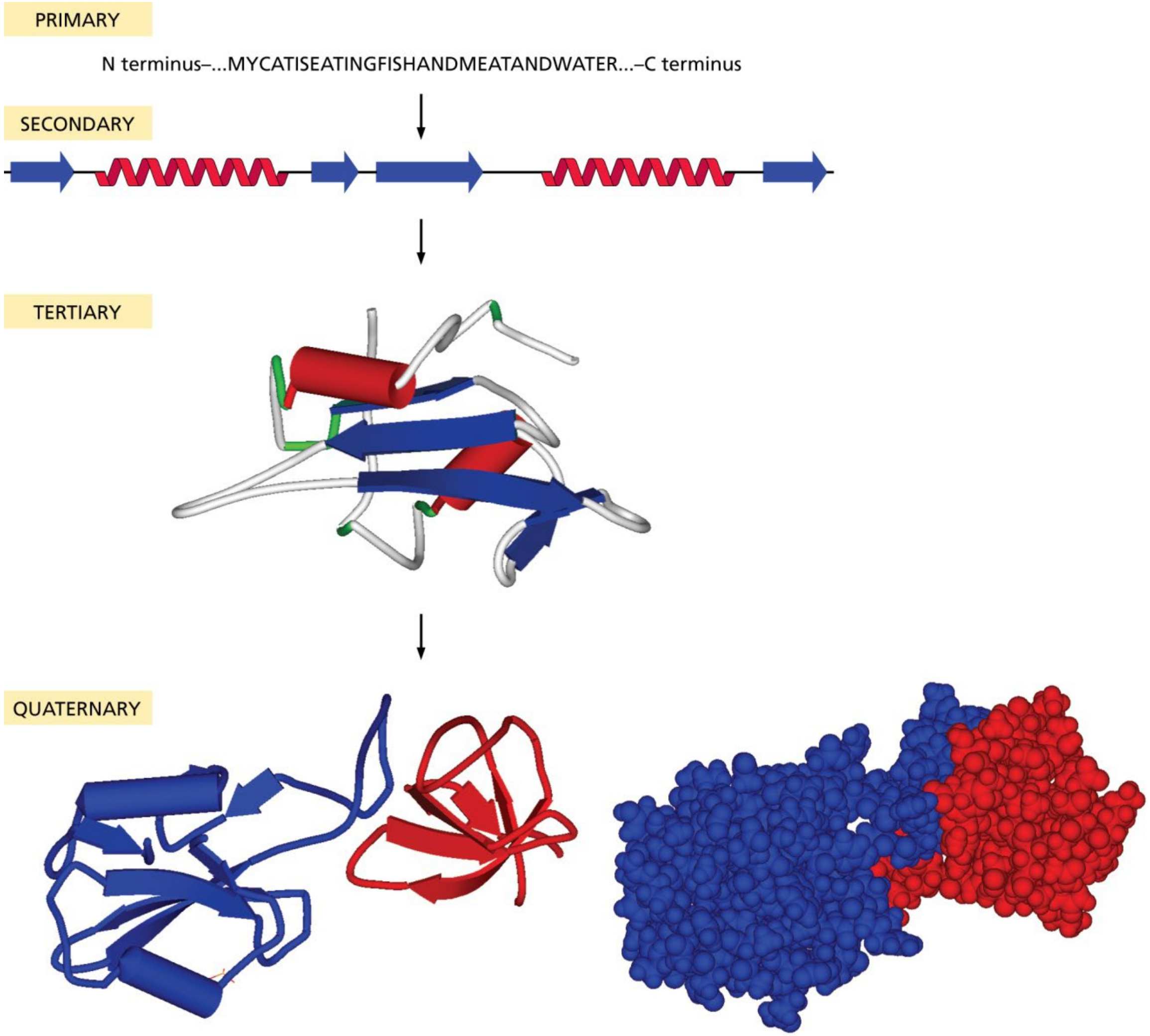


- Function: Breaks (mis-folded, broken, superfluous, …) proteins into small peptides for reuse
- Very large complex present in all eukaryotes (and more species)
  - >2000 kDa, consists of dozens of individual proteins
  - Formation of the complex is a complex process only partly understood yet



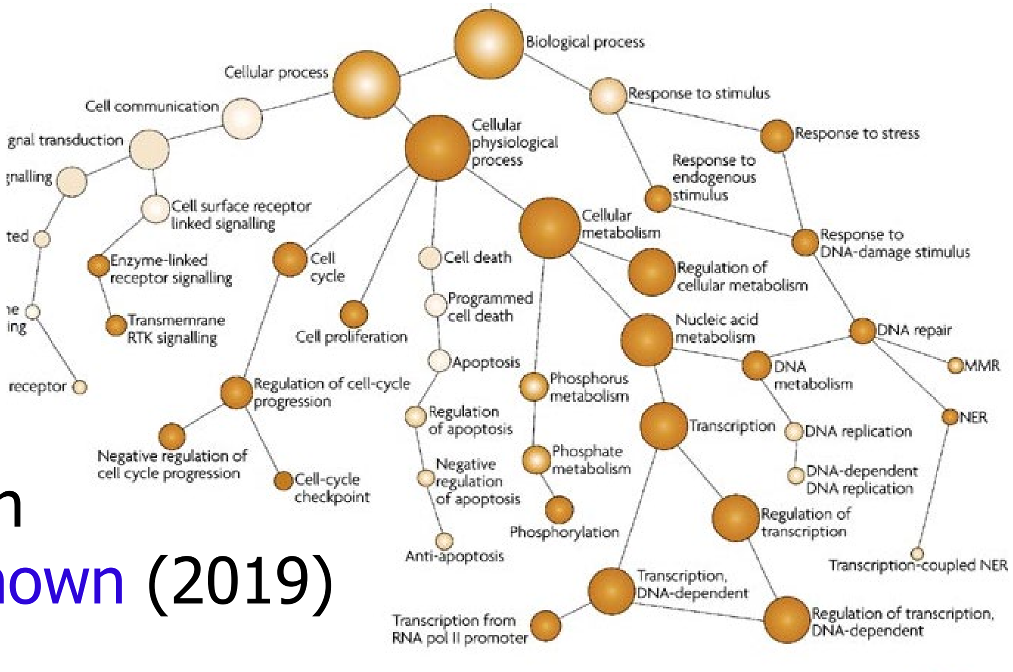
Protein Structure

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


PRIMARY
N terminus–...MYCATISEATINGFISHANDMEATANDWATER...–C terminus
SECONDARY
TERTIARY
QUATERNARY

Protein Function

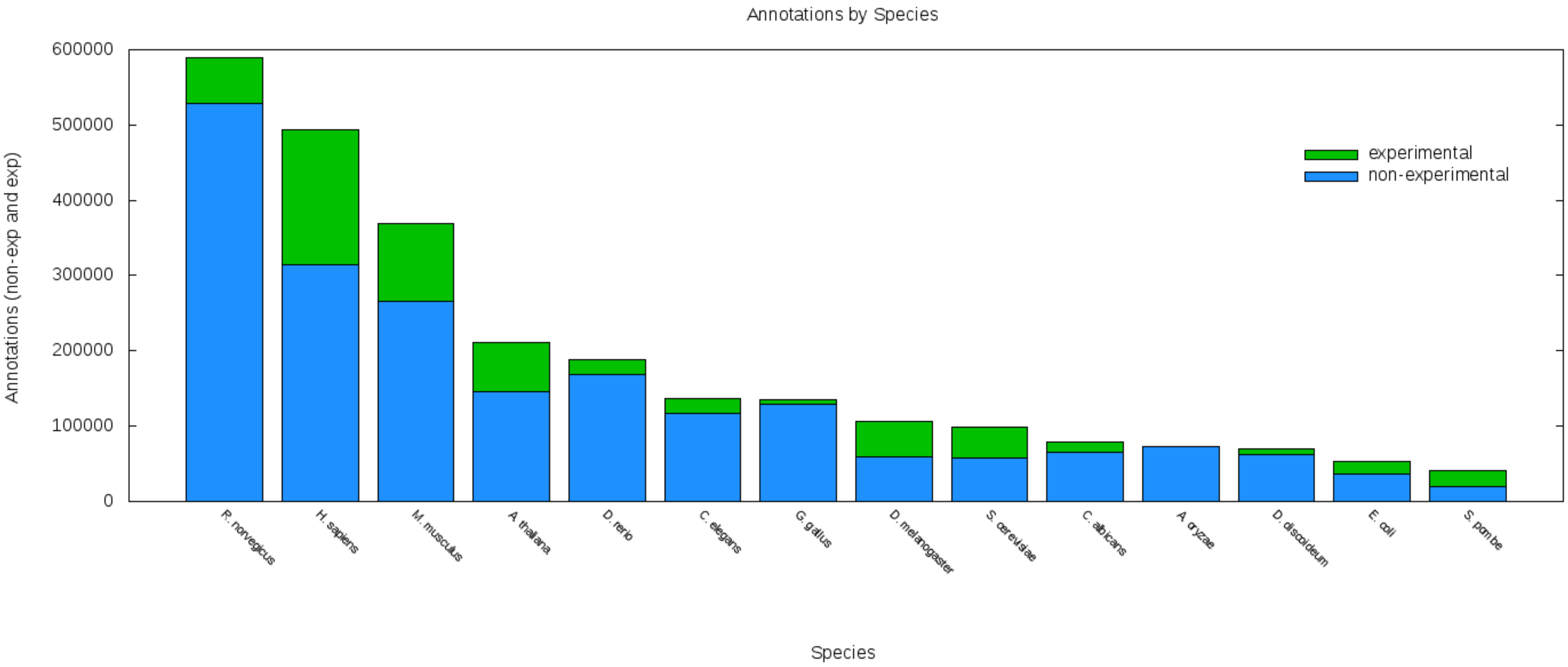
- Proteins perform many functions in living organisms
  - Metabolism
  - Signal processing
  - Gene regulation
  - Cell cycle
  - …


Biological process
Cellular process
Response to stimulus
Cell communication
gnal transduction
Response to stress
Cellular physiological process
Response to endogenous stimulus
gnalling
Cell surface receptor linked signalling
Cellular metabolism
ted
Response to DNA-damage stimulus
Enzyme-linked receptor signalling
Cell cycle
Cell death
Regulation of cellular metabolism
ne ing
Transmemrane RTK signalling
Cell proliferation
Programmed cell death
Nucleic acid metabolism
DNA repair
Apoptosis
MMR
receptor
Regulation of cell-cycle progression
Phosphorus metabolism
DNA metabolism
Regulation of apoptosis
NER
Transcription
DNA replication
Negative regulation of cell cycle progression
Phosphate metabolism
Negative regulation of apoptosis
DNA-dependent DNA replication
Cell-cycle checkpoint
Regulation of transcription
Phosphorylation
Anti-apoptosis
Transcription-coupled NER
Transcription, DNA-dependent
Regulation of transcription, DNA-dependent
Transcription from RNA pol II promoter

Nature Reviews | Cancer

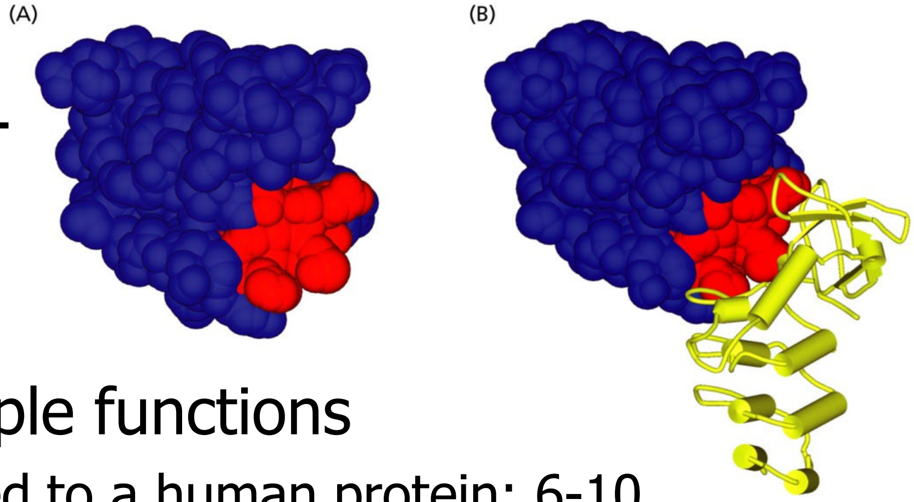
- For ~20% of all human gene, no function is known (2019)
- Describing function
  - Gene Ontology: 3 branches, >40.000 concepts
  - Used world-wide to describe gene/protein function

„Known“ Protein Functions



http://geneontology.org/page/current-go-statistics, June 2016

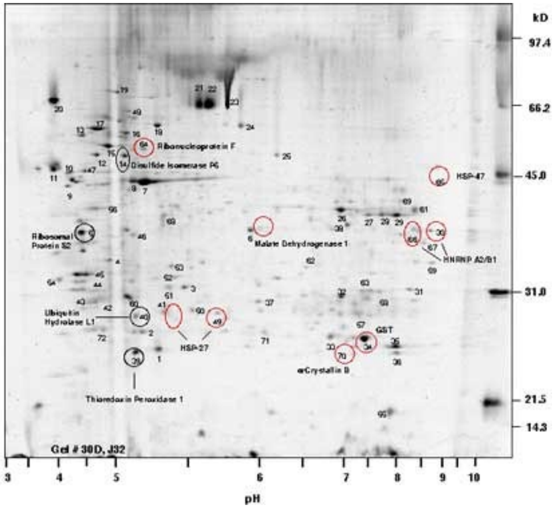
Function and Motifs



- 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

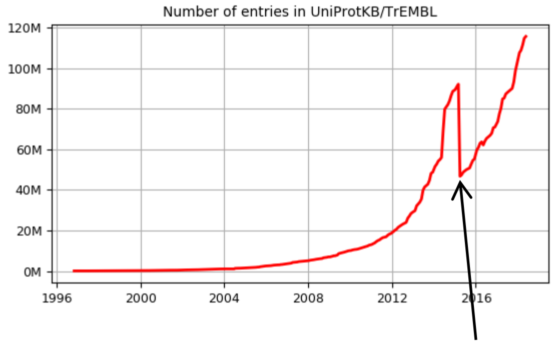
Proteomics – Large Scale Protein Identification

- Measuring gene expression: RNA-Seq, microarrays, PCR, …
- Measuring protein abundance 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



UniProt

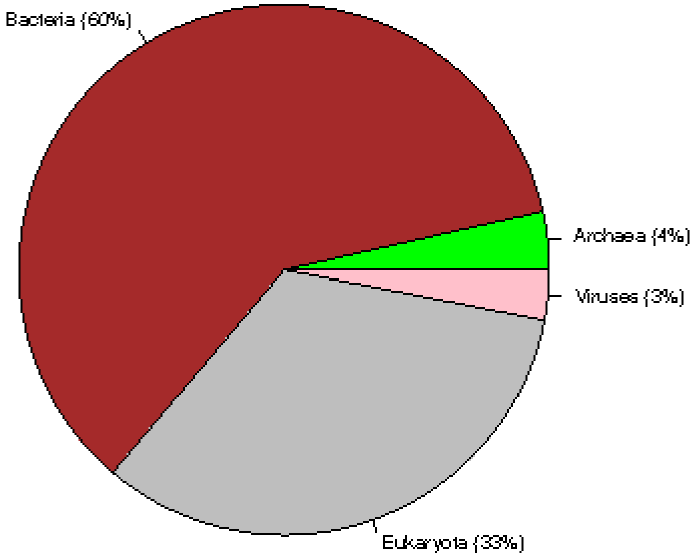
- “Standard” database for protein sequences and annotation
  - Original name: SwissProt
  - Started at the Swiss Institute of Bioinformatics, now mostly EBI
  - Other: PIR, HPRD
- Continuous growth and curation
  - >30 „Scientific Database Curators“
  - Quarterly releases
  - Very rich set of annotations

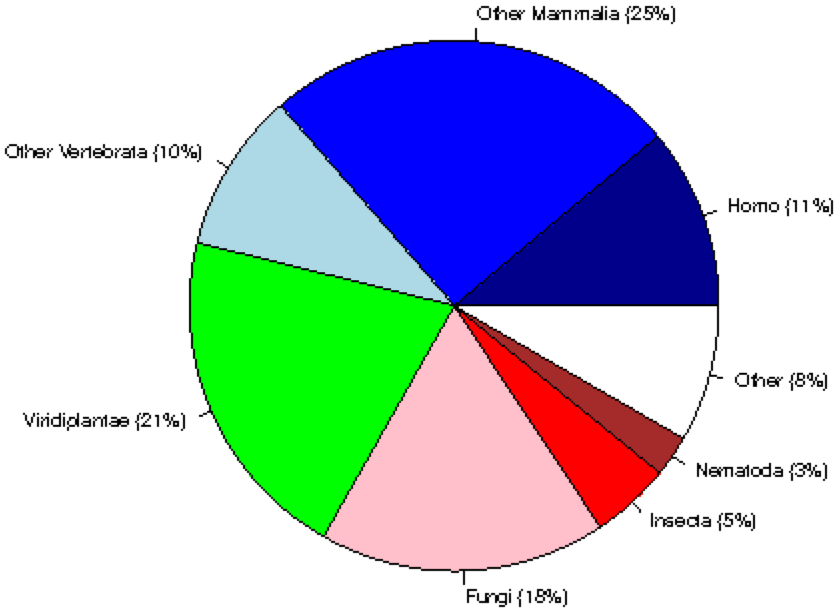


Def. and removal of „redundant“ sequences

- 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, June 2016]



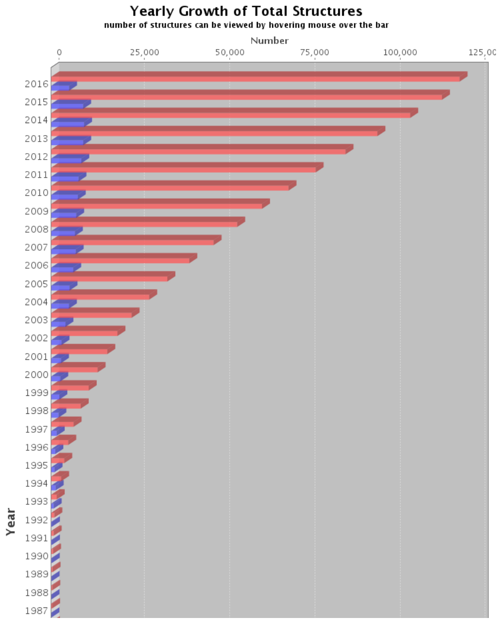


20258	Homo sapiens (Human)
16327	Mus musculus (Mouse)
9842	Arabidopsis thaliana (Mouse-ear cress)
7560	Rattus norvegicus (Rat)
6582	Saccharomyces cerevisiae (Baker's yeast)
5803	Bos taurus (Bovine)

...

PDB – Protein Structure Database

- Oldest protein database, evolved from a book
- Experimentally determined 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



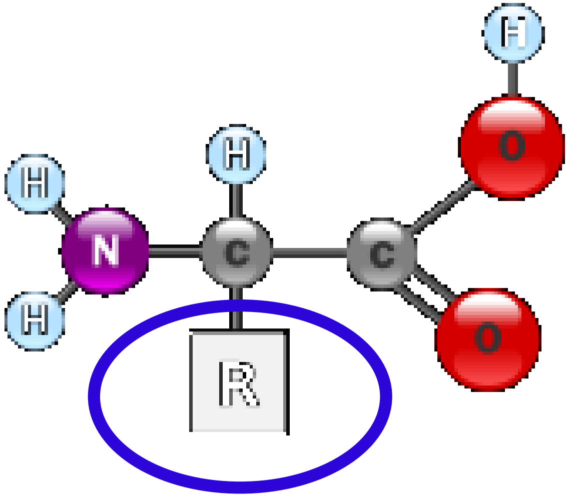
http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total, June 2016

This Lecture

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

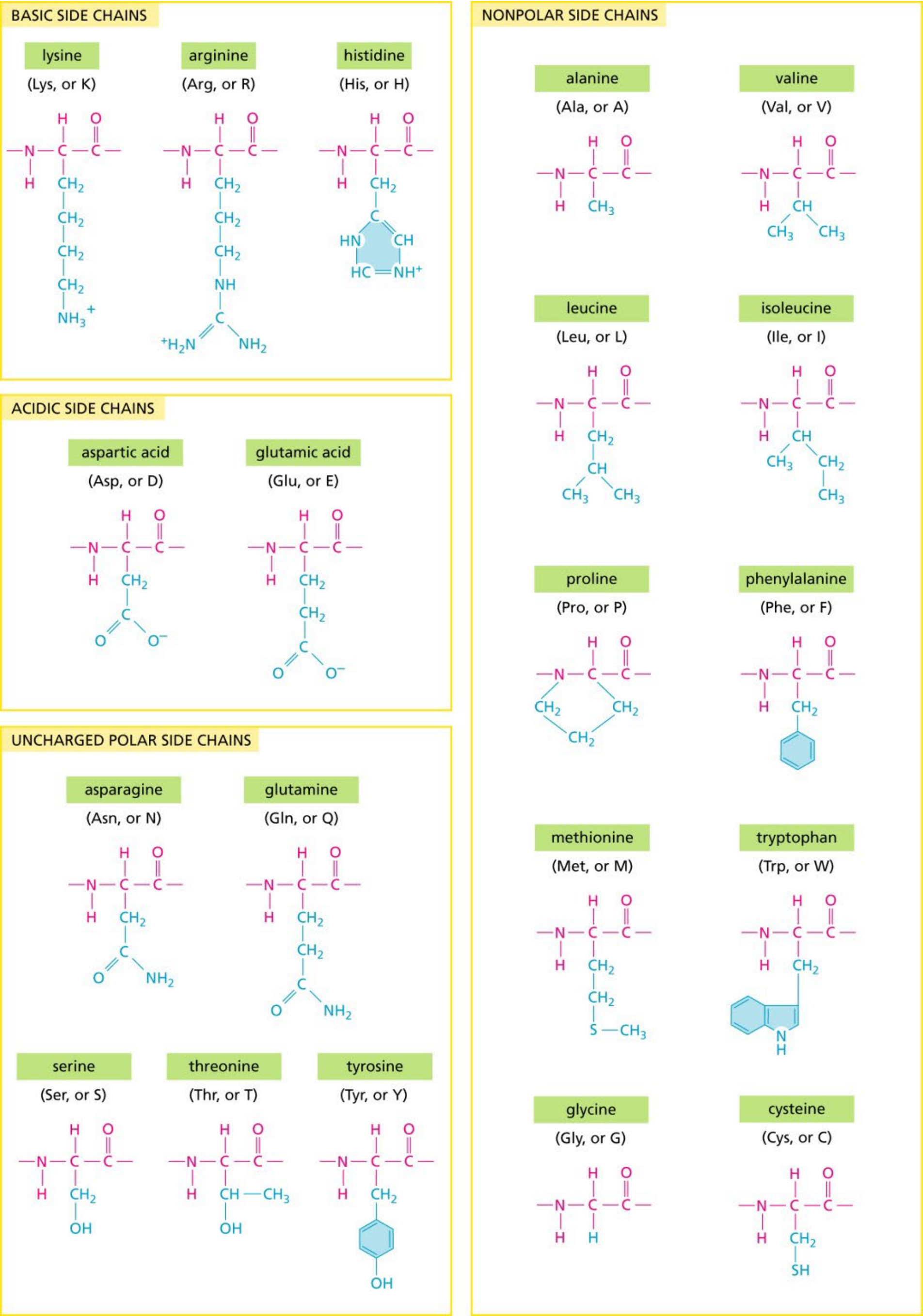
Amino Acids (AA)

- Amino acids: Common core and specific residue
  - Core
    - Amino group – NH_2
    - Central C_α - Carbon – CH
    - Carboxyl group – COOH
  - Residue: AA-specific



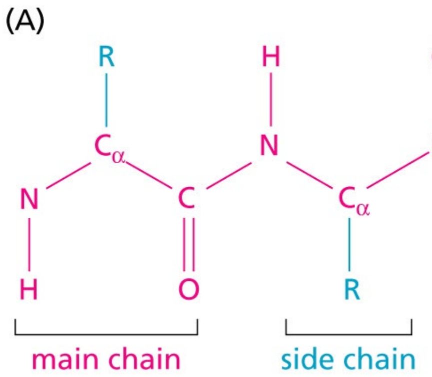
- Core: Chaining AA to protein sequences
- Residues (side chains): Specific properties of a AA
  - Vary greatly between AA

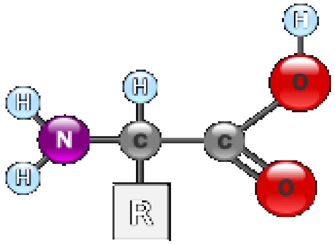
Side Chains

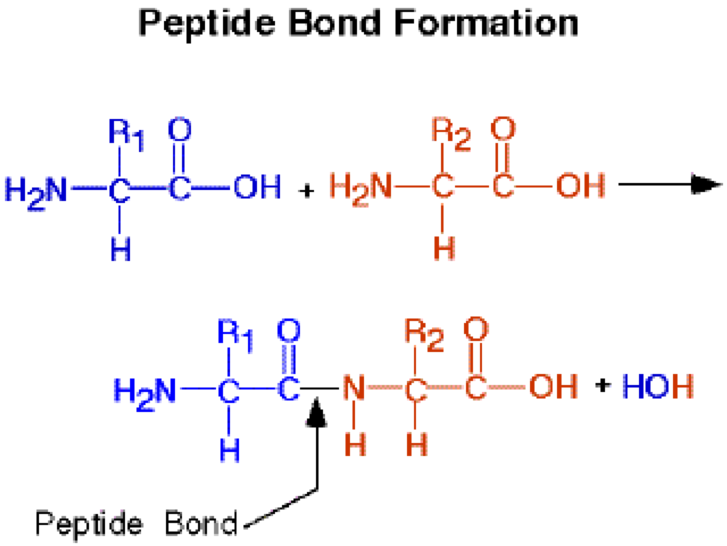

BASIC SIDE CHAINS
lysine
(Lys, or K)
arginine
(Arg, or R)
histidine
(His, or H)
ACIDIC SIDE CHAINS
aspartic acid
(Asp, or D)
glutamic acid
(Glu, or E)
UNCHARGED POLAR SIDE CHAINS
asparagine
(Asn, or N)
glutamine
(Gln, or Q)
serine
(Ser, or S)
threonine
(Thr, or T)
tyrosine
(Tyr, or Y)
NONPOLAR SIDE CHAINS
alanine
(Ala, or A)
valine
(Val, or V)
leucine
(Leu, or L)
isoleucine
(Ile, or I)
proline
(Pro, or P)
phenylalanine
(Phe, or F)
methionine
(Met, or M)
tryptophan
(Trp, or W)
glycine
(Gly, or G)
cysteine
(Cys, or C)

Structure of a Protein

- Concatenation of cores: Backbone of AA chain (= protein)
  - Covalent peptide bonds between carboxyl and amino group
  - Loss of a H_2O

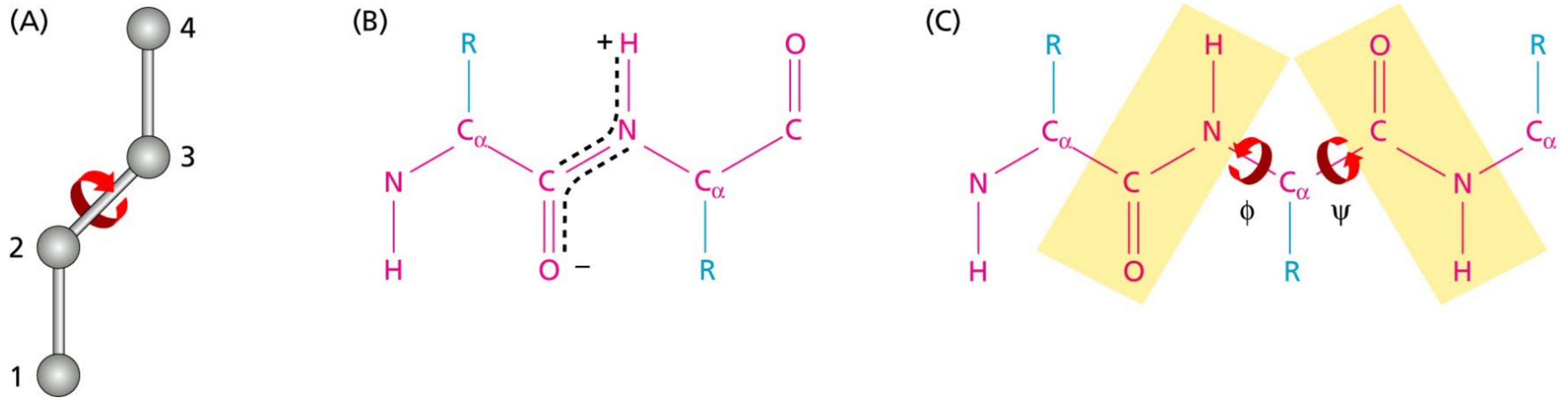






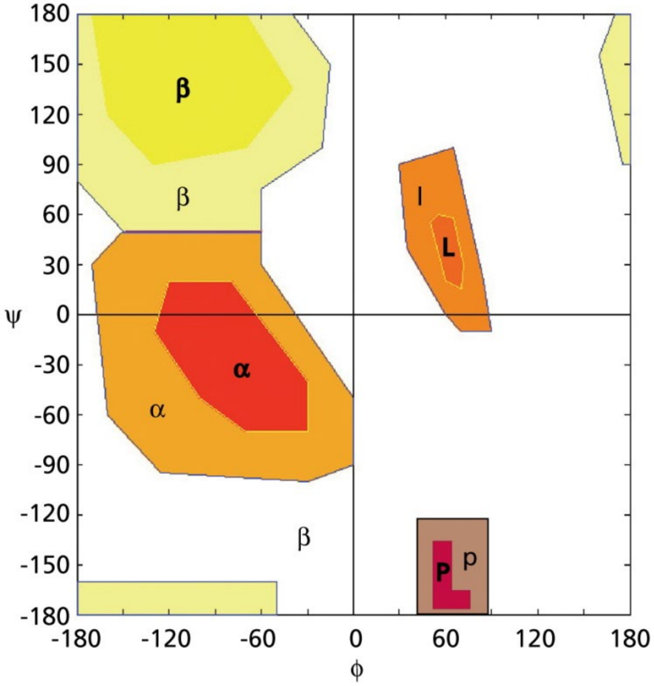
Flexibility

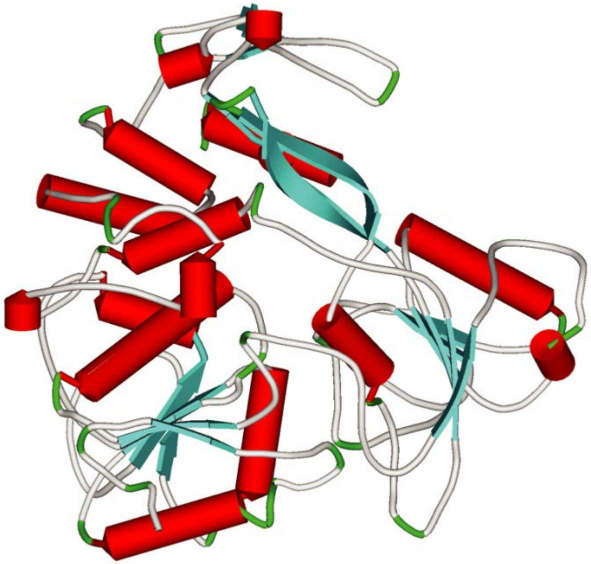
- In principle, every chemical bond can rotate freely
  - Would allow arbitrary backbone structures
- In real proteins, observed angels are strongly constrained
  - Peptide bound (B) is "flat" – almost no torsion possible
  - Flexibility only in the C_α-flanking bonds ϕ and ψ



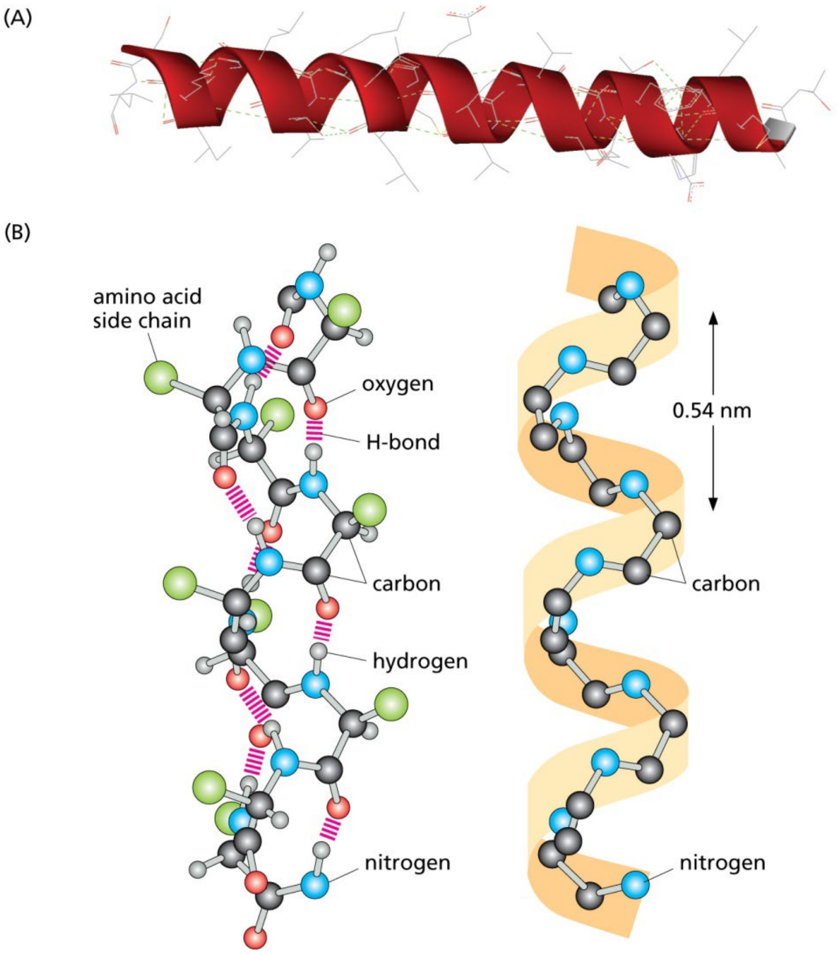
Ramachandran Plots

- 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 3D structures
  - Secondary structure





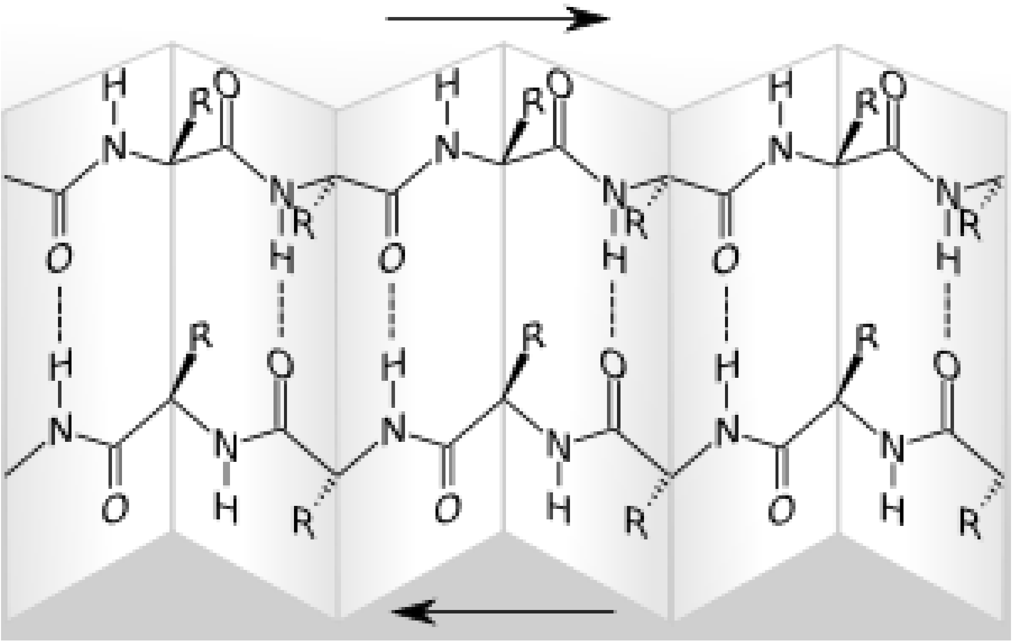
α-Helix



- 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

β-Sheet

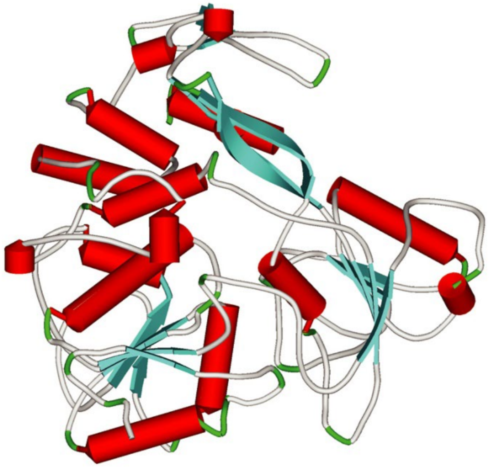
- Two linear and parallel stretches (β-strands)
- Strands are bound together by hydrogen bounds
- Can be parallel or anti-parallel (wrt. N/C terminus)



Quelle: Wikipedia

Other Substructures

- α-helixes and β-sheets cover 50-80% of most proteins
- Other parts are called loops or coils
  - Usually less important for the structure of the protein
  - But very important for its function
  - Often exposed on the surface
  - Determine 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
  - Mostly alpha, mostly beta, …
  - 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 (-)

```
KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESNFNTHATNRNTD
GSTDYGILQINSRWWCNDGRTPGSKNLCNIPCSALLSSDITASVNCAK
KIASGGNGMNAWVAWRNRCKGTDVHAWIRGCRL
```

```
KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESNFNTHATNRNTD
-----HHHHHHHHH-------------EEEEE------HHHHHHHH--
GSTDYGILQINSRWWCNDGRTPGSKNLCNIPCSALLSSDITASVNCAK
----EEEEEEEEEEEEEEEEEEE-------------------HHHHHH
KIASGGNGMNAWVAWRNRCKGTDVHAWIRGCRL
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 heuristic approach
  - Simple to explain, classical, no ML required, not too bad

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)
  - Compute a score for the probability of any AA to be E / H
    - When both are improbable: Assign C
  - Basis: Relative frequencies in a set of sequences with known SSE
  - First assign each AA its most frequent class
  - Then perform several heuristic tricks to change classes
    - E.g. minimal length of stretches
    - Example: CCEEEEEEECCECE, not CCEEECEEECCECE

Details [sketch, some heuristics omitted]

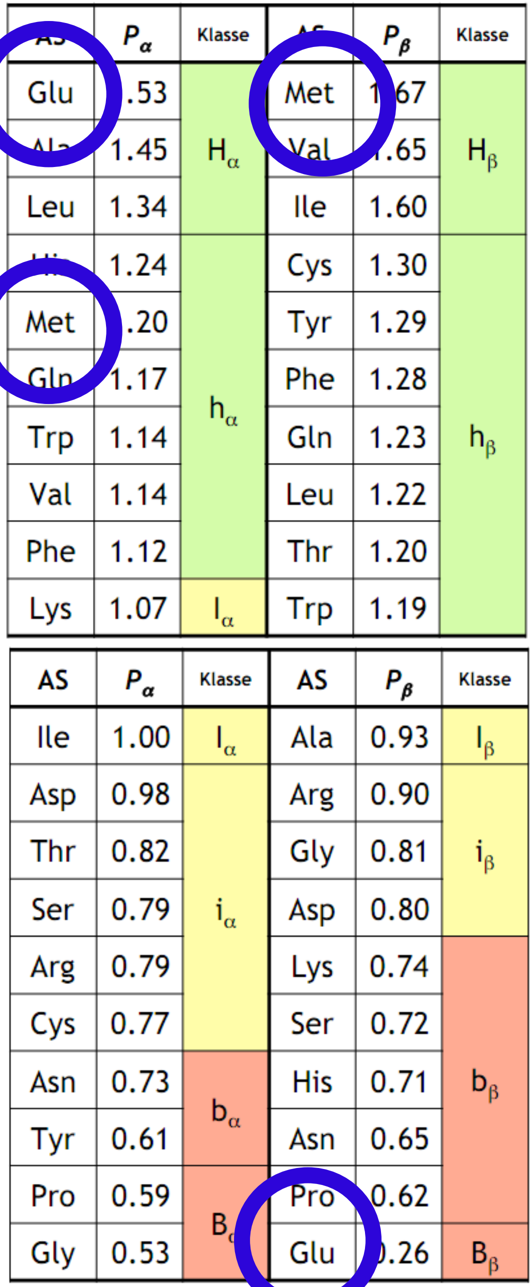
- Let $f_{j,k}$ be the relative frequency of observing AA j in class k
- Let f_k be the average over all 20 $f_{j,k}$ values
- Compute the propensity $P_{j,k}$ of AA j to be part of class k as

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

  - This is not a probability, rather an odds-score
- Using $P_{j,k}$, classify each AA j for every class k into
  - Strong, normal, weak builder (H_α, h_α, I_α, H_β, h_β, I_β)
    - Tendency to build a SS-element
  - Strong, weak breaker (B_α, b_α, B_β, b_β)
    - Tendency to stop a SS-element
  - Indifferent (i_α, i_β)
  - Thus, we actually have 12 (13) classes

Concrete Values

- Originally computed on only 15 proteins (1974)
- Read
  - Glu(tamate) often is at the start of a helix and often at the end of a strand
  - Met(hionine) often starts strands and regularly starts helices
  - …



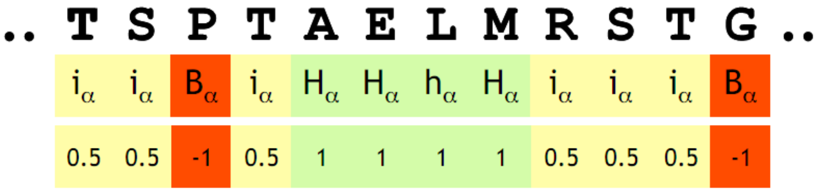
AS	P_α	Klasse	AS	P_β	Klasse
Glu	1.53	H_α	Met	1.67	H_β
Ala	1.45		Val	1.65	
Leu	1.34		Ile	1.60	
His	1.24	h_α	Cys	1.30	h_β
Met	1.20		Tyr	1.29	
Gln	1.17		Phe	1.28	
Trp	1.14		Gln	1.23	
Val	1.14		Leu	1.22	
Phe	1.12		Thr	1.20	
Lys	1.07	I_α	Trp	1.19	

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	0.26	B_β

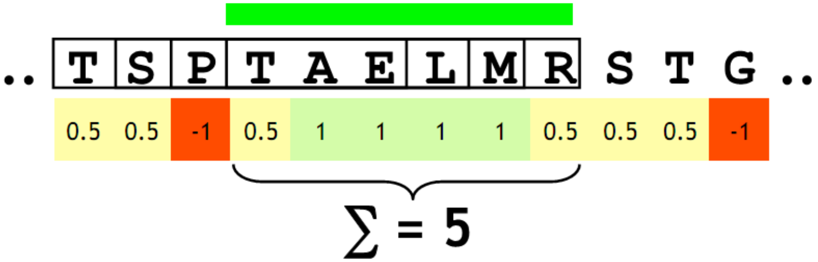
Algorithm for Helices

- Score each AA with 1 (H_α, h_α), 0.5 (I_α, i_α), or -1 (B_α, b_α)
  - Heuristic discretization – don't trust your counts too much
- 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 aggregated score A using original $P_{j,k}$ values inside each window
  - If A ≥ 4, continue the helix; otherwise stop
- Similar method for strands
- Conflicts (regions assigned both H and E) are resolved based on higher aggregated score

Example [Source: O. Kohlbacher, “Strukturvorhersage”]



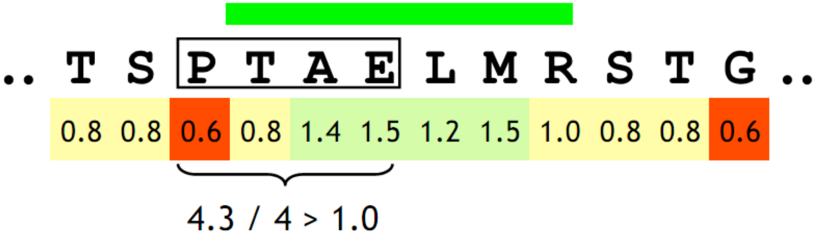
..	T	S	P	T	A	E	L	M	R	S	T	G	..
	i_α	i_α	B_α	i_α	H_α	H_α	h_α	H_α	i_α	i_α	i_α	B_α	
	0.5	0.5	-1	0.5	1	1	1	1	0.5	0.5	0.5	-1	



..	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	

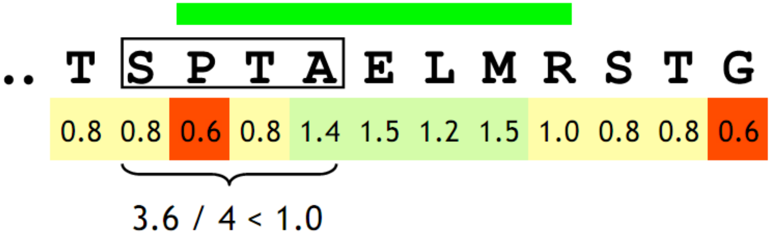
$\sum = 5$

Helixstart



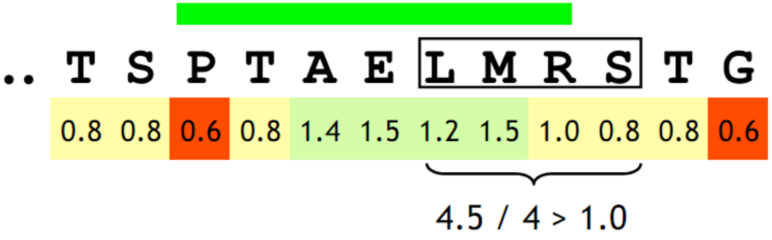
..	T	S	P	T	A	E	L	M	R	S	T	G	..
	0.8	0.8	0.6	0.8	1.4	1.5	1.2	1.5	1.0	0.8	0.8	0.6	

4.3 / 4 > 1.0



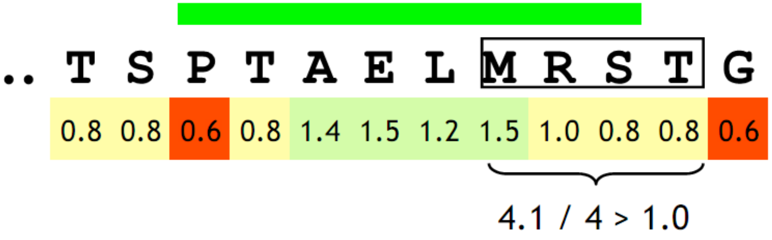
..	T	S	P	T	A	E	L	M	R	S	T	G	..
	0.8	0.8	0.6	0.8	1.4	1.5	1.2	1.5	1.0	0.8	0.8	0.6	

3.6 / 4 < 1.0



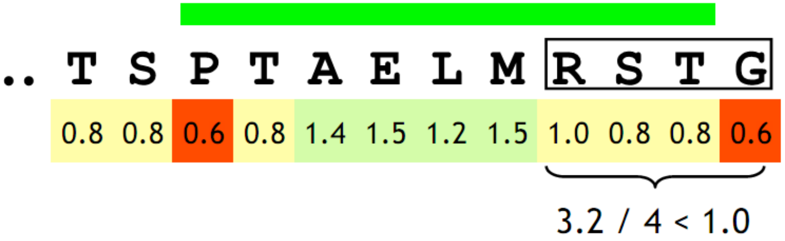
..	T	S	P	T	A	E	L	M	R	S	T	G	..
	0.8	0.8	0.6	0.8	1.4	1.5	1.2	1.5	1.0	0.8	0.8	0.6	

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
- General problem of Chou-Fasman
  - 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

- Gerhard Steger (2003). “Bioinformatik – Methoden zur Vorhersage von RNA- und Proteinstrukturen”, Birkhäuser, chapter 8,10,11,13
- Many figures from Zvelebil, M. and Baum, J. O. (2008). "Understanding Bioinformatics", Garland Science, Taylor & Francis Group, chapter 2, 11, 12 (partly)
- Many examples from O. Kohlbacher, Vorlesung Strukturvorhersage, WS 2004/2005, Universität Tübingen