

Bei meiner Prüfung kamen im Prinzip nur die unten ausgearbeiteten Fragen, aber teilweise anders formuliert/anderer Fokus. Deswegen würde ich auf jeden Fall empfehlen, die Folien auch mitanzuschauen und auf Verständnis zu lernen. In den Folien zu Machine Learning gibt es noch einen Link vom Professor, wo er Beispielfragen + dazugehörige Antworten geschrieben hat, so wie er sie zur Prüfung geben würde – das zusätzlich anschauen ist sicherlich auch nicht schlecht. Sorry for any typos und viel Erfolg!

Nur ein kurzes Update: es werden nach wie vor die Altfragen geprüft, das einzige was sich geändert hat, ist dass man statt dass man alle möglichen phylogenetischen trees mit 4 taxa zeichnet, von 6 gegebenen phylogenetischen trees auswählen musste, welche gleich sind und welche verschieden und dann noch zwei zeichnen, die nicht dabei sind in der liste, aber die gleichen taxa haben.

- 1)
a.) Mention at least 4 amino acids that can be found on the surface of globular proteins. Describe their function.

Glutamic acid / glutamate	Glu	Ionic interactions with NH_3^+ groups
Aspartic acid / aspartate	Asp	
Arginine	Arg	Forms salt bridges with $-\text{COO}^-$ functions
Lysine	Lys	Participates in catalysis or metal binding
Histidine	His	

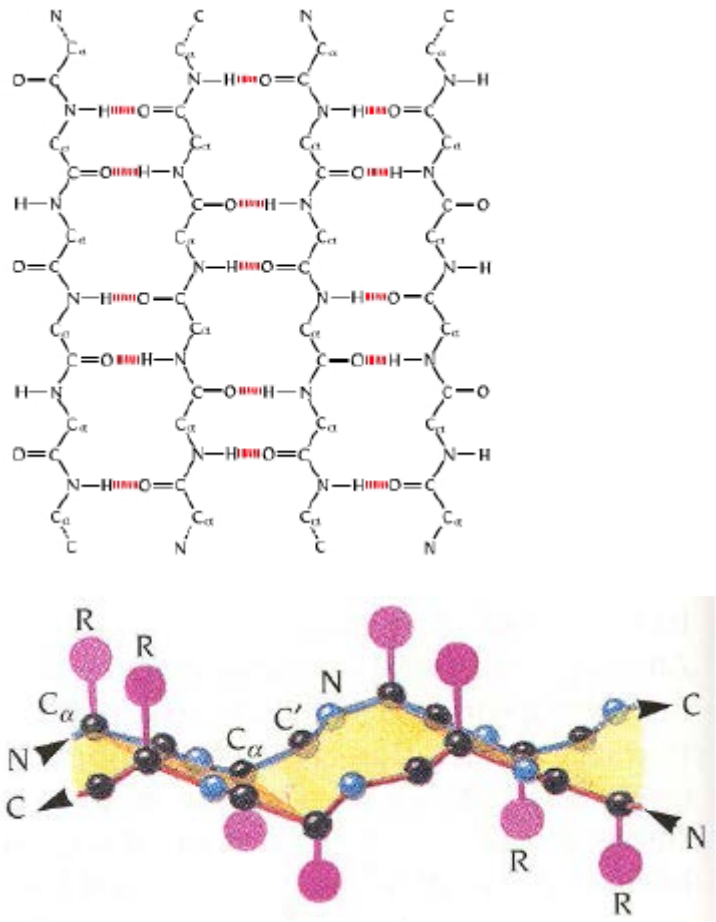
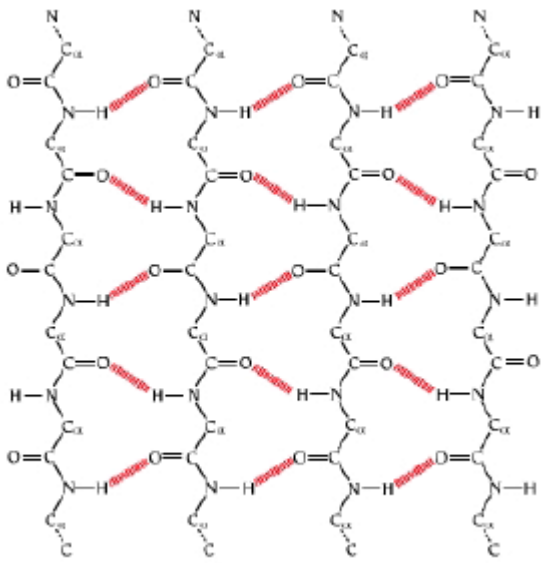
- b.) Picture of a Beta-Barrel was given. Describe the sequence pattern knowing that the core is hydrophobic and the surface hydrophilic.

The barrel consists of alternating polar and hydrophobic amino acids, so that the hydrophobic residues are oriented toward the core and the hydrophilic residues toward the outside of the barrel. Moreover, the beta-sheets in beta-barrels are usually antiparallel.

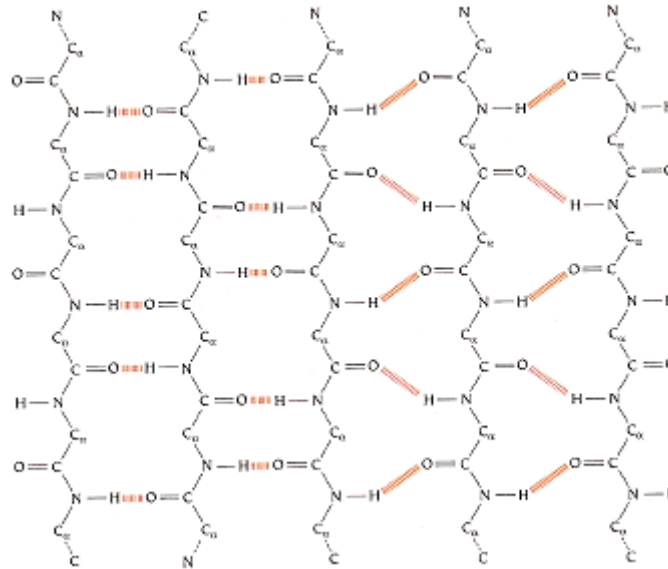
- c.) Describe a Beta-sheet structure. Shape of backbones, where are the side chains located, which forces stabilize the structure?

Very common in globular proteins. The beta-sheet consists of a combination of several regions of the poly-peptide chain. One beta-strand is typically 5 to 10 residues long or 3.4 Å/amino acid; the ideal torsion angles are: $\phi = -140^\circ$, $\psi = 130^\circ$. The strands are aligned adjacent to each other, connected by hydrogen bonds between $\text{C}'=\text{O}$ and NH . The residues of C_{α} atoms are vertically protruding on both sides of the pleated beta-sheet.

There are three types of beta-sheets:

<p>Antiparallel</p>		<p>Successive strands have alternating directions (N- to C- terminal) and a distinct pattern of H-bonds (narrowly-widely-spaced). All possible H-bonds are formed, except for the two flanking beta-strands. Antiparallel beta-sheets are more twisted, which allows for distortions and solvent exposition.</p>
<p>Parallel</p>		<p>Successive strands have the same direction. H-bonds are evenly spaced and bridge the strands at an angle. Parallel sheets are less twisted and always buried in the protein.</p>

Mixed



Adjacent strands may have the same or opposite direction to the preceding strand. However, there is a strong bias against mixed beta-sheets (only about 20%)

d.) How can proteins be classified based on their secondary structure?

Protein secondary structure can take one of these forms:

Alpha-Helix: right-handed helix of 4-40 residues length. Left-handed helices are rarely found (not in L-amino acids) and are very short. One turn has 3.6 residues. The turns are crosslinked by H-bonds between $C'=O$ of residue n and NH of residue $n+4$. All NH and $C'=O$ groups are joined by H-bonds except for the first NH group and the last $C'=O$ group. Thus, the ends are polar, which is important for biological function and binding of ligands.

Other forms of helices are albeit existent very rare (3(10)-helix, π -helix)

Strand: Beta-strands are 5-10 residues long (3.4 Å per amino acid). The ideal torsion angles are: $\phi = -140^\circ$ ($N-C_{\alpha}$), $\psi = 130^\circ$ ($C_{\alpha}-C'$). Adjacent beta-strands form a beta-sheet.

Turn: Tight turns on the surface of the molecule connecting adjacent antiparallel beta-strands.

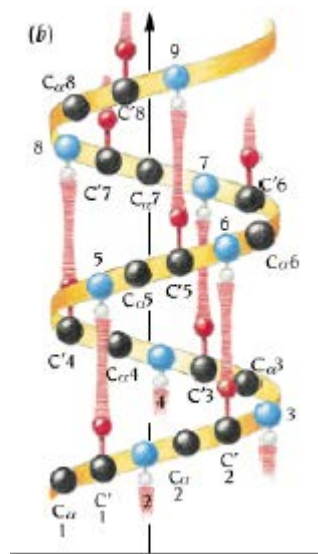
Unfavorable positions in the Ramachandran plot can often be attributed to turns.

- Beta-hairpin: loop of variable length; 70% consist of less than 7 amino acids
- Gamma-turn: only one residue is not involved in the sheet; unfavorable for adjacent H-bond in sheet.
- Beta-turn: with two residues; there are three types of which type I is the most common. Glycine, asparagine and proline are often found in reverse turns.

Loops: connects different types of secondary structure. Loops are often flexible and can adopt different conformations, and therefore open and close the active site of an enzyme. Loop regions are at the surface of the molecule, where $C'=O$ and NH are typically H-bonded to water and not to each other. They often contain hydrophilic amino acids. Loop regions frequently participate in binding sites and enzyme active sites; antigen binding sites of antibodies are built up from six loop regions. In regard to evolution, cores were more stable and changes almost exclusively occur in loops. The length of loop is apparently not important for the protein core.

If combined, secondary structures form motifs (supersecondary structures).

Coiled coil: (fibrous proteins, e.g. fibrinogen, myosin and keratin) Left-handed supercoiles consist of 2 or 3 right-handed alpha-helices. The number of residues per turn are reduced to 3.5 and thus the pattern of side-chain interactions between the 2 helices is repeated every 7 residues = heptad repeat.



Helix-loop-helix: helix-turn-helix (specific for DNA binding), calcium-binding motif

Beta-hairpin: present in most antiparallel beta-structures. The length of the loop is 2-5 residues.

Greek key motif: formed from 4 adjacent beta-strands. One long antiparallel structure with loops in the middle of both beta-strands.

Beta-alpha-beta motif: While this motif can have two different “hands”, it is almost always found right-handed in proteins. It is often found in parallel beta-sheets to connect beta-strands. The alpha-helix shields hydrophobic residues of beta-strands from water.

Classification by comparison of protein structures and identification of relationships among structures:

SCOP – Structural Classification of Proteins curated, manual comparison

CATH – Class, Architecture, Topology and Homology curated, automated procedure, manual comparison

Prediction of secondary structure

Structural data is usually not available for new sequences, unknown folds or artificial sequences. The statistical evaluation of many sequences shows a good correlation between amino acid sequence and 3D-structure. More than 90% of all amino acids in proteins can be assigned to secondary structures, less than 10% are random coils.

Ab initio-based makes use of sequence information only. The prediction is based on statistical calculations.

Short-range interactions like alpha-helices are easy to predict, but the prediction of long-range interactions like beta-sheets is extremely difficult.

Homology-based does rely on secondary structural patterns conserved among multiple homologous sequences rather than statistics.

2)

a.) Write the consensus sequence of 5 (given) sequences using regular expression. b.) What is the difference of regular expression and profiles or HMMs?

Pattern notation to describe a motif using regular expression

- Every amino acid or nucleotide has one letter: ACDEFGHIKLMNPQRSTVWY
- X non-specified amino acid
- () repetitive stretches of the same symbol; e.g. V(2)=2 valines, X(2,4)=2, 3 or 4 amino acids
- [] alternative amino acids; e.g. [R,K]=arginine or lysine
- { } disallowed amino acids; e.g. {F,Y,W}=anything but phenylalanine, tyrosine or tryptophan
-

E-X(2)-[FHM]-X(4)-{P}-L = E – two non-specified aa – either F, H or M – 4 non-specified aa – anything but P – L

Regular expression, Profiles and HMMs are used to identify new domains and motifs based on a multiple alignment. Regular expression is similar to a consensus sequence, but contains more information (e.g. exclusion of certain aa or non-specific aa) and is therefore ideal for database searches and discovery of distant homologs. Profiles and HMMs are both statistical approaches. Profiles are PSSM that take gap information into account. By calculating the likelihood that a sequence belongs to the same class, distant homologs are identified. A Markov Model describes a sequence of events occurring one after another, where each event determines the probability of the next event. HMMs (Hidden Markov

Models) include unobserved states (state=nucleotide or amino acid). Observed and hidden states are connected with a transition probability. The probability of each state is called emission probability. A multiple alignment based model is trained to obtain the calculated optimal statistical parameters in the HMM and can thereby be used to assess how well a sequence matches the model (obtain further alignments).

c.) Mention one method to perform multiple sequence alignment of more sequences simultaneously.

- Progressive** All possible pairs are aligned by global alignment and a distance matrix between sequences as well as a distance tree is created. The closest aligned pairs are transformed to consensus sequence, which is then aligned to the next closest sequence and so on. Progressive alignment cannot handle variations in length of the sequences. Once gaps are introduced in a pairwise alignment, it is fixed. The best known programs using progressive alignment are Clustal (Global alignment, uses different scoring matrices depending on the similarity of aligned sequences) and T-Coffee (global AND local pairwise alignment combined, better but also slower than Clustal)
- Iterative** Iterative corrections are made to the alignment until convergence of the score
- Outer iteration: Tree based on pairwise alignments, weights are given to the alignment scores
 - Inner iteration: Sequences are randomly divided into two groups and the two derived consensus sequences are globally aligned. This process is repeated until there is no more improvement.
- Blockwise** Block-based alignments are best for more divergent sequences of varying lengths. Blocks of un-gapped alignments are identified and aligned locally. First the sequence is split up into small segments and all possible pairwise alignments are made. High scoring blocks can be progressively aligned. There is no gap penalty.

d.) PSSM and 5 sequences from which it was constructed were given.

Write the consensus sequence of the PSSM, explain how to do that and compare the consensus to every single sequence above.

5 Sequences + tables see L5, p. 9-10

Position-Specific Scoring Matrices give a statistical representation of a multiple sequence alignment. To construct a PSSM the observation frequencies of all residues at all positions and in total are calculated [$1/\text{amount of sequences} * \text{amount of residue on at this position}$], then the frequency of each residue at every position is normalized with the occurrence of the residue in general [$\text{frequency}/\text{occurrence}$]. Subsequently the $^2\log$ is taken of the probabilities [$\ln(\text{probability})/\ln(2)$]. The residues with the highest values per position are used for the consensus sequence.

Now to calculate the likelihood of one sequence, the sum of the log odd scores of the correct residue at every position is taken as power of 2 (2^{sum}) → it is 2^{sum} times more likely than by random chance that this sequence fits to the PSSM.

3) Describe the origin of noise. (20 pt)

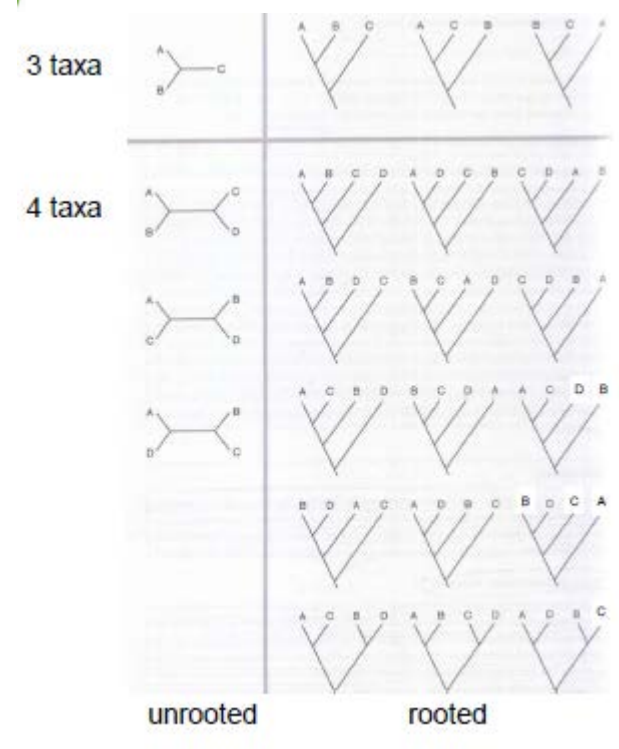
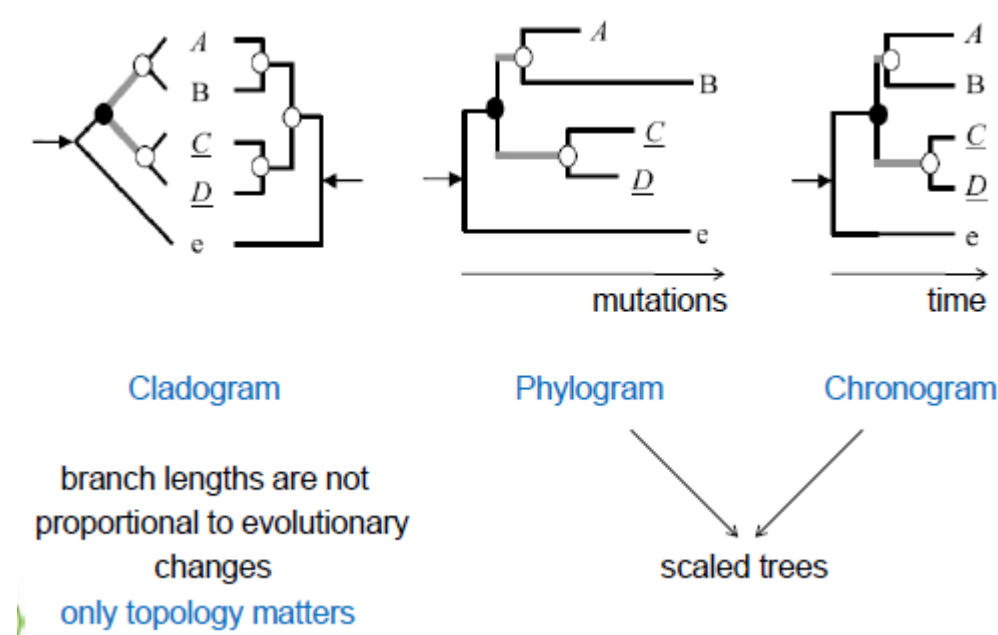
Measurement processes involve errors which arise from noise. Reasons for noise can be:

- **measurement noise** random errors
- **misclassifications** e.g. wrong phenotype
- **simplified methods** certain influences are ignored e.g. because they are too difficult to control. However, they still alter the output.

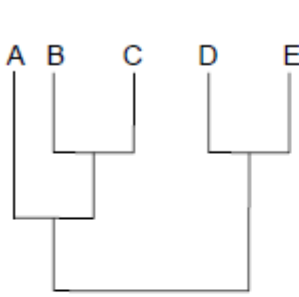
To remove noise, data analysis uses replicates. The remaining aspects are modeled as well as possible. An optimally

chosen model should have appropriate abstraction – not too simple and not too complex – and explain new observations reasonably well while avoiding fitting to the noise.

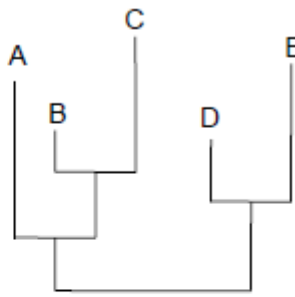
- 4)
a. Draw rooted tree for 3 taxa A,B and C (only the relevant topologies).



- b.) Represent the following clades using the Newick format (tree was given).



(((B,C),A),(D,E))



(((B:1,C:2):1,A:2):1,(D:1.2,E:2.5):1.5)

the numbers are relative units
representing sequence divergence

<http://marvin.cs.uidaho.edu/Teaching/CS515/newickFormat.html>

c.) Mention the 4 steps for tree construction.

1. **Choose molecular markers:** nucleotide sequences for closely related organisms, protein sequences for far related sequences (protein sequences are more conserved).
2. **Perform multiple sequence alignment:** most critical step – it establishes the positional correspondence of sequences. Therefore, alignments must be chosen carefully and often have to be edited manually.
3. **Choose a tree building method:** In general, tree-building methods are either based on distance between pairs of sequence (amount of dissimilarity) → assume that all involved sequences are homologous, or based on discrete characters → assume that characters at corresponding positions are homologous among the sequences.
4. **Assess tree reliability:** How reliable is a tree and is it significantly better than another?
 - **Jackknifing** half of the sites in a data set are randomly deleted – the new dataset is subjected to phylogenetic construction using the same method as for the full tree and compared to the original one
 - **Bootstrapping** perturbation of the original sequence dataset by either random replacement of sites in the sequence (nonparametric BS) or altered datasets with random sequences confined within a particular sequence distribution according to a given substitution model (parametric BS). The latter prevents multiple repeat of certain sites which is a problem of the nonparametric BS.

For reliable results, bootstrapping replicates should be performed about 1000-times. A bootstrap value of 70% corresponds to an approximate 95% statistical confidence.

d.) Why is it difficult to find the correct tree?

“Essentially, all models are wrong but some are useful”, there are too many possibilities for phylogenetic trees.

Homoplasy (=number of observed substitutions does not represent the true evolutionary events: back mutation, unknown intermediate steps, parallel mutations in both sequences) leads to incorrect trees

5)

a.) Explain the 2 experimental methods to determine protein structure. What parameters are measured, how do we get to the structure?

X-Ray Crystallography Proteins need to crystallize to large crystals in which their positions are fixed. The x-rays are then

deflected by the electron clouds surrounding the atoms in the crystal, producing a regular pattern of diffraction. The diffraction can be converted to an electron density map using Fourier transformation.

NMR Spectroscopy Detection via the spinning patterns of atomic nuclei in a magnetic field. Protein samples are labeled with C(13) and N(15) radioisotopes. A radio frequency radiation is used to induce transitions between nuclear spin states in a magnetic field. Interactions between isotope pairs produce radio signal peaks that correlate with the distances between them. By interpreting the signals observed by the NMR, proximity between atoms can be determined. Knowledge of distances between all labeled atoms in a protein allows a protein model to be built that satisfies all the constraints.

b.) Explain 2 methods to get protein models. Why do they work?

Homology modeling The prediction of the tertiary structure of an unknown protein using a known 3D structure of homologous proteins. Underlying is the assumption that structure is more conserved than sequence. Known homologs (templates) on which the prediction is based and which should have their 3D geometry determined by X-ray or NMR are prerequisite.

Threading/structural fold recognition Assessment of the compatibility of an amino acid sequence with a known structure in a fold library. There are two methods: 1. Pairwise energy-based method (threading) predicts the structural fold of a protein by fitting the protein sequence to all folds and selects the best-fitting fold by ranking the calculated energies. It can be used to identify structurally similar proteins even without detectable sequence similarity. 2. Profile-based method (fold recognition) generates a structural profile by superimposition. Profile scores are calculated for solvent exposure, hydrophobicity and polarity of the amino acids.

Ab initio protein structure prediction Attempt to produce protein models solely based on sequence information. This method tries to calculate the protein structure with the lowest global energy which is assumed to represent the native state. The validity of this assumption however is not confirmed. On the upside, this method is not limited to known protein folds, however, the physiochemical laws governing the folding process are not yet well understood – the used energy functions are inaccurate.

c.) Ramachandran plot: explain the dihedral angles and how it can be used to evaluate models.

The polypeptide backbone rotation is limited to the two torsion angles ϕ (N-C_{alpha}) and ψ (C'-C_{alpha}). These angles define the secondary structure. Since ϕ and ψ are the only degrees of freedom, the conformation of the whole main chain is determined when they are known. The Ramachandran plot writes ϕ and ψ against each other and maps the entire conformational space of a peptide and shows allowed and disallowed regions.

Welche AS kommen im hydrophoben core vor und welche Eigenschaften haben sie? Erklären wie das beta-sheet aufgebaut ist, welche Form das backbone hat (Winkel phi und psi miteinbeziehen!), wie die Seitenketten angeordnet sind und durch welche Kräfte ein beta-sheet zusammenhält. Es war die Skizze eines beta-barrels gegeben und man sollte erklären welche AS man darin am meisten findet.

Isoleucine	Ile	Found mostly in the hydrophobic core
------------	-----	--------------------------------------

Leucine	Leu	Participating in hydrophobic interactions	
Methionine	Met	Nonreactive side chains	
Valine	Val		
Phenylalanine	Phe	Found in the hydrophobic core	
Tyrosine	Tyr	Aromatic interaction (π -stacking)	Weakly polar – hydrogen bonding, phosphorylation sites
Tryptophan	Trp		

Mögliche Tertiärstrukturen von Proteinen aufzählen und erklären.

Domain structures can be classified into three main classes:

Alpha-domains core consists of mainly alpha-helices

- **Four-helix bundle:** simplest and most frequent alpha-helical domain. The helical axes are almost parallel to each other. The core is hydrophobic while the surface is hydrophilic.
- **Globin fold:** Globins are monomeric or oligomeric heme-containing proteins (e.g. myoglobin, hemoglobin, phytocyanins). They consist of a bundle of eight helices (A-H), connected by rather short loops. The helices form a pocket for the active site that binds the e.g. heme.

Beta-domains usually two antiparallel beta-sheets packed against each other

- **Closed barrel:** The last strand of a beta-sheet is joined by H-bonds to the first strand. The strands are antiparallel and connected by hairpins.
- **Jelly roll barrel:** Jelly roll motif wrapped around a barrel. Though there are also larger barrels, most consist of 8 beta-strands. The barrel has two connections across the top and the bottom.

Alpha/beta-domains combination of beta-alpha-beta motifs, parallel or mixed beta-sheets surrounded by alpha-helices; loops participate in catalytic action and binding, but do not contribute to the structural stability. These structures are the most frequent domain structures.

- TIM barrel
- Rossmann fold
- Horseshoe fold

BLAST (20P):

Difference between heuristic sequence alignment and dynamic programming, name 1 dynamic programming method

Dynamic programming is an exhaustive solution for local and global alignments. However, because of its complexity and computational intensity, it is too slow and therefore impractical for database searches. While less accurate, heuristic sequence alignment is less time consuming and thus also more commonly used.

Dynamic programming methods: Needleman-Wunsch algorithm (global), Smith-Waterman algorithm (local)

Heuristic: BLAST, FASTA

Explain how BLAST works

BLAST (Basic Local Alignment Searching Tool) is a heuristic database searching tool for multiple sequence alignment. It splits queries up into words (3-6 aa or 16-258 bp) and derives all similar words based on the BLOSUM62 score. It then searches for words in the sequence database. The alignments are extended into both directions until the score drops

below a threshold. Alignments with high enough score are kept.

Difference between BLASTN, TBLASTX; Why is this useful?

To overcome the degeneracy of the genetic code: Amino acids are more informative and sensitive. Nucleotide comparisons are only meaningful for very close homologs and non-coding sequences.

BLASTN uses nucleotides in its query and searches for nucleotides in the database. BLASTX compares the nucleotide query to the amino acid database.

Molecular clock hypothesis, how is it used

The Molecular Clock Hypothesis is used to estimate the time of occurrence of speciation/mutation events=divergence time by using fossil evidence or DNA/protein sequences for calibration.

*"If molecular sequences evolve at constant rates,
the amount of accumulated mutations is proportional to evolutionary time"*

BUT uniformity of evolutionary rates is rarely found, because of

- changing generation times
- population size (in larger populations, the fitness advantage of any mutation becomes smaller)
- species-specific differences (metabolism, ecology, evolutionary history)
- evolving functions of the encoded protein
- changes in the intensity of natural selection

Regions of low selection have **high rates** of variation: 0.7-0.8% /myr

Regions of high selection have **low rates** of variation: 0.02% /myr

Individual clocks can be tested for accuracy by calibrating them against material evidence. Over long time spans, estimates can be off by more than 50%!

List the two data analysis methods

Supervised learning Used for regression problem – known data, modelling of given target variable

Unsupervised learning Used for exploratory data analysis –unknown groups in data/explanation of reduced dimension

Link application scenarios to data analysis method

Task	– >	Method
predict continuous y from input data	– >	Regression
predict discrete y from input data	– >	Classification
find unknown groups in input data	– >	Clustering (e.g. k-means, mixture models)
find low dimensional representation for input data	– >	Dimensionality reduction (PCA, ICA)

BLOSUM

was sind scoring matrices? Beschreibung

Scoring matrices are a set of values for quantifying the likelihood of one residue at a certain position being substituted by another residue in an alignment. It is derived from statistical analysis of residue substitution data from sets of reliable

alignments of highly related sequences.

was soll die 62 bei BLOSUM 62

BLOSUM = Blocks of Amino Acid Substitution Matrix. It is derived from multiple sequence alignments of 500 groups of proteins. Blocks used are ungapped alignments of less than 60 aa length. In comparison to PAM, all BLOSUM matrices are observational. BLOSUM matrices with high numbers are suitable for alignments of closely related proteins of high sequence identity, those with low numbers for rather divergent proteins. For BLOSUM62, the sequences selected to derive the matrix had an average identity of 62%.

eine BLOSUM scoring matrix war gegeben, anhand derer 2 matrizen scoren finde die matrix die 0% identity mit der bei c gegebenen Matrix

1. look up the score for every pair of amino acids
2. add up the scores

A high score indicates a good alignment which is probably biologically meaningful. It is very unlikely that you get this score by random arrangement of amino acids. A low score indicates the opposite (duh).

Beispiel mit membrangebundenem Enzym mit h m-Gruppe - warum ist es schwer, die Struktur zu ermitteln + durch welche Mutationen kann man es doch schaffen

Membranprotein mit Heme Group und anchor- welche mutationen um es experimentell zu bestimmen (Anchor bindet das Globul re Protein an Membran > Anchor wegbringen > geht in L sung)

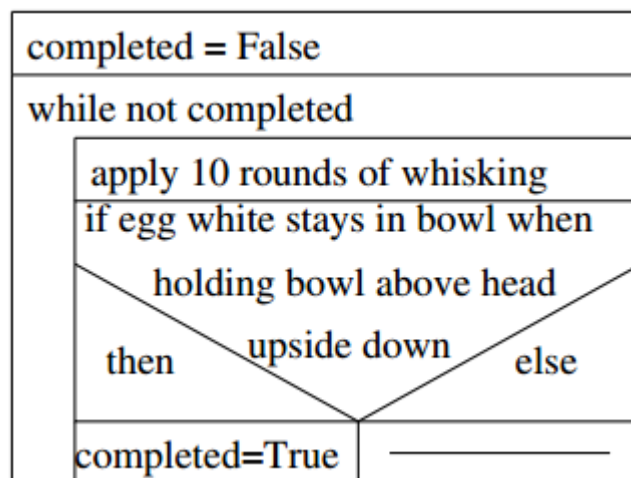
Machine Learning

was ist der unterschied zwischen anweisungen an menschen und computer?

A computer operates at a „simple“ input-output basis. Input is processed to output. The main difference between a computer and a human is the human creativity and ability of applying common sense; computers on the other hand will always operate exactly as specified in the program. This means that programmers have to carefully think through all steps of the program and avoid undesired side-effects.

welche paradigms kennst du?

1. **Imperative Programming** the solution is described using sequences of instructions, alternatives and loops



(“how to beat an eggwhite”)

2. **Object Oriented Programming** solution and data are tied together. Objects are containers of data and „methods“ which describe actions how to interact with the objects.
3. **Functional Programming** the solution is described by its properties and making use of induction

Clustering

Ist es schlimm wenn die werte mit unterschiedlichen scales gemessen wurden? muss man sie vereinheitlichen?

If data is measured on different scales, the results of the cluster analysis becomes biased. In order to prevent biased results, data has to be standardized.

k-means clustering, was muss man vorher angeben etc. etc.

The k-means algorithm is an iterative procedure and depends on the randomly chosen starting values. Beforehand k observations are chosen by random. These points represent the initial cluster centroids. Then the distance of each object to all centroids is calculated and the object is assigned to the cluster of the nearest centroid. When all objects have been assigned, the centroids of the k clusters are recalculated. The last two steps are repeated until a maximum number of iterations is reached or the centroids do no longer change.

Gemeinsamkeiten und Unterschiede zwischen den 2 Clusteringmethoden (hierachal, Partitionale Algorithmen)

The aim of cluster analysis is finding groups in data.

Hierarchical clustering algorithms: agglomerative algorithms (bottom-up), divisive algorithms (top-down) – results in a dendrogram = tree-like structure

Partitioning algorithms: k-means clustering algorithm (most common): you have to specify the number of clusters beforehand

Multiple sequence alignment
wie kann man das scoren?

PSSM

Profiles

HMM

Amino acids – code

AMINO ACID				AMINO ACID			
		SIDE CHAIN				SIDE CHAIN	
Aspartic acid	Asp	D	negative	Alanine	Ala	A	nonpolar
Glutamic acid	Glu	E	negative	Glycine	Gly	G	nonpolar
Arginine	Arg	R	positive	Valine	Val	V	nonpolar
Lysine	Lys	K	positive	Leucine	Leu	L	nonpolar
Histidine	His	H	positive	Isoleucine	Ile	I	nonpolar
Asparagine	Asn	N	uncharged polar	Proline	Pro	P	nonpolar
Glutamine	Gln	Q	uncharged polar	Phenylalanine	Phe	F	nonpolar
Serine	Ser	S	uncharged polar	Methionine	Met	M	nonpolar
Threonine	Thr	T	uncharged polar	Tryptophan	Trp	W	nonpolar
Tyrosine	Tyr	Y	uncharged polar	Cysteine	Cys	C	nonpolar

POLAR AMINO ACIDS

NONPOLAR AMINO ACIDS

AMBIGUOUS AMINO ACIDS

3-letter 1-letter

asparagine or aspartic acid

Asx B

glutamine or glutamic acid

Glx Z

leucine or Isoleucine

Xle J

unspecified or unknown amino acid

Xaa X

Functional features

glycine	{	nonreactive, Gly and Pro disrupt regular secondary structures
proline		Gly adopts flexible conformations, Pro – cyclic, rigid
cysteine	{	posttranslational modification sites
serine		participating in active sites of enzymes
threonine		metal binding
glutamine	{	hydrogen bonding partners
asparagine		in enzyme active sites
arginine	{	forming salt bridges with -COO^- functions
lysine		usually found on the surface of globular proteins
histidine		participates in catalysis or metal binding

Functional features

glutamate	{	on the surface of globular proteins
aspartate		ionic interactions with -NH_3^+ groups
Isoleucine	{	found mostly in the hydrophobic core
leucine		participating in hydrophobic interactions
methionine		nonreactive side chains
valine		
phenylalanine	{	aromatic interactions (π -stacking)
tyrosine		found in the hydrophobic core
tryptophan		Tyr and Trp are weakly polar – hydrogen bonding, phosphorylation sites