Achtung! Diese Fragensammlung ist nicht vollständig und basiert nur auf wenigen Prüfungen. Die einzelnen Fragen stehen daher eher stellvertretend für ein Fragengenre (statt BLAST könnte z.B. auch ein anderer Algorithmus geprüft werden). Normalerweise kommen offenbar 5 – 7 Fragen mit 3 – 5 Unterpunkten die sich auf die Hauptfrage oft direkt beziehen („Warum ist es so wie sie es vorhin beschrieben haben?“). Der Stoff wird stark auf Verständnis geprüft, Formeln aus den Folien werden lt. Prof. Oostenbrink nicht verlangt.

Wie funktioniert BLAST? [L2]

= “Basic Local Alignment Search Tool”
Split query up in words (3 amino acids or 11 base pairs)
Derive all similar words (based on BLOSUM62 score)
Search for words in the sequence database
Extend the alignment in both directions until the score drops below a threshold
Alignments with a high enough score remain

Unterschied zwischen BLASTN/TBLASTX (mit der Tabelle aus den Folien); was sind die Vorteile? [L2]

BLASTN: Vergleicht eine Nukleotidsequenz gegen eine Nukleotidsequenzdatenbank
TBLASTX: Vergleicht die six-frame-Translation einer Nukleotidsequenz gegen die six-frame-Translationen einer Nukleotidsequenzdatenbank.
Vorteil von TBLASTX: „Structure is more conserved than sequence“, daher: das Alignment mit AA liefert genauere Informationen als jenes mit Nukleotiden.

Difference between heuristic methods and dynamic programming - give one example for a dynamic programming algorithm. [L2]

Faster, heuristic, algorithms for database searches: BLAST and FASTA
Dynamic programming Algorithmen: Global: Needleman-Wunsch (bei sehr ähnlichen Sequenzen), Local: Smith-Waterman
Unterschied:
Exhaustive (like dynamic programming) rigorous, exact solution
Heuristic (rule based) empirical, reasonable solution

Zwei wichtige Anwendungen von supervised learning tasks [L6]

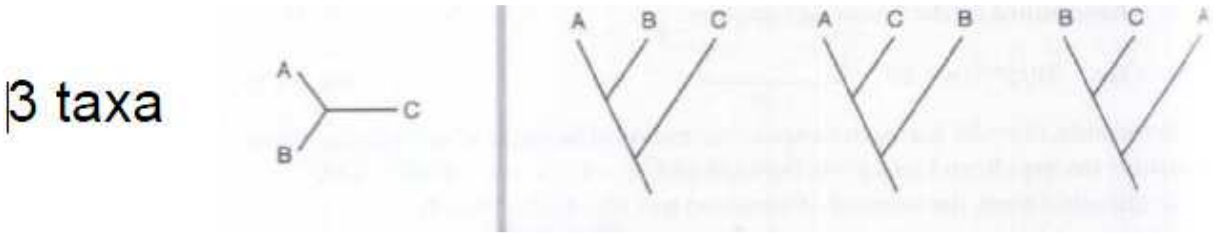
All data analysis problems can be grouped into two categories:
1. Supervised Learning methods are used for regression problems.
2. Unsupervised Learning methods are used for exploratory data analysis.

predict continuous y, from input data - > Regression
predict discrete y from, input data - > Classification

Zwei wichtige Anwendungen von unsupervised learning tasks [L6]

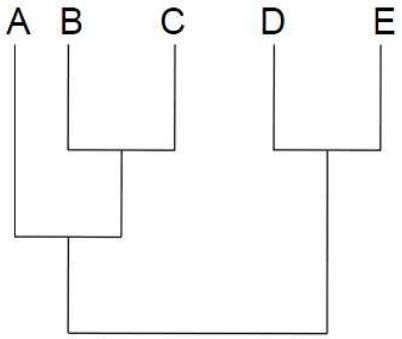
find unknown groups in input data - > Clustering (e.g. kmeans, mixture models)
find low dimensional representation for input data - > Dimensionality reduction PCA, independent component analysis (ICA)

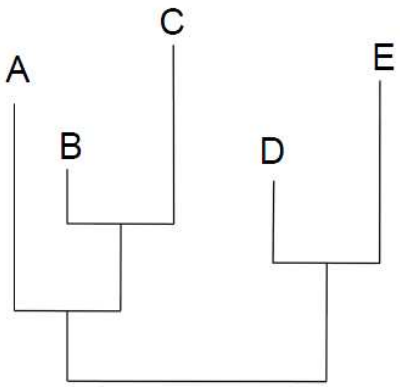
Mögliche Topologien eines rooted tree mit 3 Taxa aufzeichnen [L4]



Nur die zeichnen, die nicht nur durch Drehung um eine Achse entstehen. In diesem Beispiel also z.B. nicht den Baum mit C B A weil der ja ident ist mit dem B C A.

Ein gegebenes Cladogram im Newick-Format darstellen (oder umgekehrt) [L4]





(((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

Molecular Clock Hypothese erklären [L4]

MC is used to estimate the time of occurrence of speciation/mutation events (divergence time) by using fossil evidence or DNA/protein sequences

if molecular sequences evolve at constant rates, the amount of accumulated mutations is proportional to evolutionary time
problem: uniformity of evolutionary rates is rarely found because of:
1. changing generation times
2. population size (in larger populations the fitness advantage of any one mutation becomes smaller)
3. species-specific differences (metabolism, ecology, evolutionary history)
4. evolving functions of the encoded protein
5. changes in the intensity of natural selection
rates of molecular evolution
regions of low selection (silent substitutions) have high rates of variations:
0.7-0.8% per Myr (in bacteria, invertebrates and plants)
regions of very high selection (encoding rRNA) show low rates:
0.02% per Myr (million years)
calibration: individual molecular clocks can be tested for accuracy – they need to be calibrated against material evidence such as fossils
over long time spans, estimates can be off by 50% or more.

Welche Möglichkeiten zur Überprüfung von phylogenetischen Bäumen gibt es? [L4]

Jacknifing: half of the sites in a data set are randomly deleted – the new data set is subjected to phylogenetic tree construction using the same method as for the full tree and compared to the original one
problem: the dataset is no replicate of the original one

Bootstraping: relies on the perturbation of the original sequence dataset
1. nonparametric BS : random replacement of sites in the sequence
problem: certain sites may appear multiple times
2. parametric bootstrapping uses altered datasets with random sequences confined within a particular sequence distribution according to a given substitution model;
for publication: BS is performed 1000 times

4 Aminosäuren die in Helices vorkommen, welche Eigenschaften haben diese? [L9]

Leu , Ala, Glu, Met, Gln, Lys, Arg

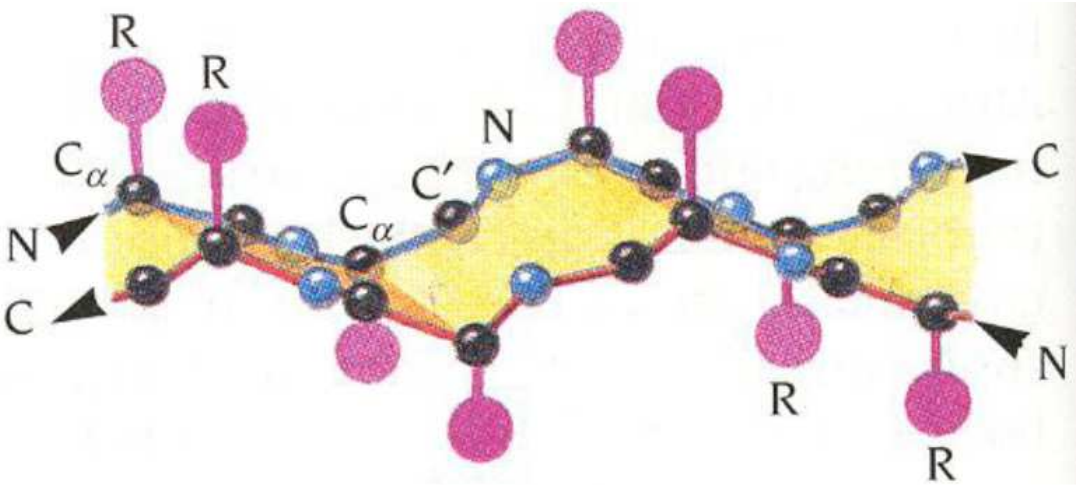
hydrogen bond between C'=O of residue n and NH of n+4 Thus all NH and C'=O groups are joined by H bonds except the first NH groups and last C'=O groups. Ends of helices are polar.

Viele nonpolar, Welche Eigenschaften sonst?

Nennen sie 4 Aminosäuren, die im hydrophobic core vorkommen und beschreiben sie deren Eigenschaften. [L9]

Isoleucine, Leucine, Methionine, Valine
found mostly in the hydrophobic core, participating in hydrophobic interactions, nonreactive side chains

Aufbau von Beta-Sheets erklären, Backbone, Konformation, in welche Richtung stehen die Seitenketten? [L9]



built up from combination of several regions of the poly-peptide chain very common in globular proteins.
strands are typically 5 to 10 residues long
ideal torsion angles: ϕ = -140°, ψ= 130°
strands are aligned adjacent to each other such that hydrogen bonds can form between C'=O of one b-strand and NH on an adjacent strand
residues of C_α atoms above and below the plane of the sheet
paralell/antiparalell sheets (Ein Strang C, einer N-terminal = antiparalell)

Welche Kräfte stabilisieren Beta-Sheets? [L9]

Hydrogen bonds between C'=O of one b-strand and NH on an adjacent strand.

Gegeben war ein wasserlösliches Beta Barrel/ Closed Barrel mit hydrophobem Core. Welche Aminosäuren-Abfolge erwartet man? [L9]

In many cases the strands contain alternating polar and hydrophobic amino acids, so that the hydrophobic residues are oriented into the interior of the barrel to form a hydrophobic core and the polar residues are oriented toward the outside of the barrel on the solvent-exposed surface. Porins and other membrane proteins containing beta barrels reverse this pattern,

with hydrophobic residues oriented toward the exterior where they contact the surrounding lipids, and hydrophilic residues oriented toward the interior pore.

In welche Klassen der Tertiärstrukturen können Proteine anhand ihrer Sekundärstruktur unterteilt werden? [L9]

alpha-domains: core built up exclusively from helices
beta-domains: usually two antiparallel sheets packed against each other
alpha/beta-domains: combination of motifs, parallel sheet surrounded by helices

Warum ist es nützlich ein 3D-Model/Proteinmodell zu haben? [L10/L8]

When Experiment is impossible, too dangerous, too expensive or blind (interesting properties cannot be observed within the parameters of the experiment) simulation is the solution. Simulation can also complement the experiment. Computational models can be useful to rationalize and predict experimental findings.

Mit welchen 2 Methoden kann man experimentell zu Proteinmodellen kommen? Auf welchen physikalischen Phänomenen beruht das Modeling jeweils? [L10]

We cannot „see“ proteins at an atomic resolution.
visible spectrum has a wavelength of 390 - 750 nm, bond lengths = 0.15 nm

1. X-Ray Crystallography:
Proteins may form crystals, X-rays have the correct wavelength to give a diffraction pattern
From the intensity of all signals, one can determine the 3D electron density.
The atomic structure is fit to the density.
Unclear why proteins crystallize and why not (trial and error)
Phase problem for back calculation to electron density need to be solved.
The rules for diffraction are given by Bragg’s law.

2. NMR Spectroscopy: Nuclear magnetic resonance
Nuclei with spin (H, 13C, 15N) can be aligned in a magnetic field.
Energy difference depends on: applied magnetic field, chemical environment and spin-spin interactions
Intensity of cross peaks indicates distance between nuclei
From this, protein models can be obtained

Electron microscopy:
Less purity of protein needed, more transient state analysis
Suited to membrane proteins (schwer zu kristallisieren!)
High resolution possible but difficult to achieve; usually no atomic model

Mit welchen 2 Methoden kann man die Tertiärstruktur vorhersagen (in silico)? [L10]

1. Homology modeling:
Search for homologous proteins in de pdb
Align sequences
Determine structurally conserved regions / variable regions
Determine coordinates: backbone, SCR, SVR, sides chains
Restrictions:
Loops are not as structured as secondary structure elements
Identical sequences of up to 8 amino acids found in both alpha helix and beta-strand

2. Protein threading:
Usually a protein looks similar to something we know and can therefore be compared to known proteins
Put your sequence over all known protein folds
Calculate their energies and check for the best energy values.

Ab initio predictions: Statistical programs to predict secondary structure elements

Gegeben waren 2 Gruppen von Proteinen. Gruppe 1 (ein kurzes Peptid) hatte eine geringe Sequence identity (17%), aber einen sehr ähnlichen Fold. Gruppe 2 (bereits ein kleines Protein) hatte eine höhere Sequence identity (50%) aber einen komplett anderen Fold. Die Sequenz von jeweils einem der beiden Proteine in beiden Gruppen war synthetisch, das andere natürlich vorkommend. Wie steht diese Tatsache zu ihrer Antwort auf die vorige Frage? [L10]

Vorhersage sollte mit verschiedenen Methoden erfolgen (z.B. nicht Homology modeling alleine). Sonst noch was?

Wie kann man die Güte eines solchen Models evaluieren? [L10/L9]

Ramachandran plot (siehe nächste Frage)
Automated evaluations: Procheck, What-Check
Check if hydrophilic outside, hydrophobic inside
The resulting structure must obey the obvious structural features of protein structures

Warum ist eine Überprüfung im Ramachandran Plot sinnvoll? [L9]

Siehe vorherige Frage ;)
Ramachandran Plot is a diagram that shows the phi/psi angels of all bonds in the molecule and therefore shows if they are in regions with low or no steric hindrance. If they are in such regions the molecule model is assumed to be more favorable.

Gegeben war ein Protein mit Anker in einer Membran und einer heme-group. Warum ist es gerade bei diesem Protein schwierig, die Struktur aufzuklären? [L10]

Membrane bound proteins don't crystallize as without special tricks and are not in solution.

Die Aufklärung ist mit Hilfe von Mutationen gelungen, which Mautations would you propose? [L10]

Removal of the membrane anchor. [sonst noch was? Was ist mit der heme-group, stört die?]

Gegeben war ein kurzes Multiple Sequence Alignment von einigen Proteinsequenzen (mit Gaps!), das man in Form einer Regular Expression darstellen musste. [L3]

Regular expressions = Pattern notation to describe a motif

- 20 AA letters: ACDEFGHIKLMNPQRSTVWY
- X marks an nonspecified amino acid
- ( ) are used to give repetitive stretches of the same symbol, V(2) = 2 Valines, X(2,4) : a stretch of 2, 3 or 4 amino acids
- [ ] are used to give alternative amino acids, [R,K] : an Arg or a Lys
- { } are used to give disallowed amino acids, {F,Y,W} = anything but not Phenylalanine, Tyrosine or Tryptophan

Gegeben war ein Alignment einiger kurzer DNA Sequenzen und die dazugehörige Position Specific Scoring Matrix (PSSM): Was sagt sie aus? Man sollte die Consensus Sequence dieses Alignments anhand der PSSM angeben und mit den einzelnen Sequenzen vergleichen. [L3, S. 2ff]

Statistical representation of a multiple sequence alignment
Calculate observation frequencies for all residues at every position and in total (G=3 C=2 => G=0,6 C=0,4)

Normalize the frequencies with the occurrence of the residue (number/"all value", 27% G => 0,27 => 0,6/0,27 = 2,22)

Take the ^{2}log of the probabilities (Rechnung auf den Folien ist nicht ln() - warum?)

Calculate the likelihood of a new sequence:
Add up the log odd scores of the correct residues at every position

2^result = times more likely than by random change that this sequence fits to the PSSM

Was ist der Vorteil einer PSSM gegenüber einer Regular Expression? [L3]

PSSM gibt auch Wahrscheinlichkeiten für das Auftreten einer bestimmten AA an einer Stelle an. Die Darstellung als Regular Expression ist nicht gewichtet.

PAM Matrix berechnen(?) und erklären wie sie funktioniert. [L2, S. 14 ff]

An empirical scoring matrix calculated from closely related proteins.
Positive score: substitution more often observed than random
Zero: substitution as expected from random
Negative score: substitution less often observed than random
PAM: Point Accepted Mutation
PAM1 matrix: derived from direct "genetic neighbours" in phylogenetic trees (1% observed mutation rate)
PAM250 matrix: likelihood for 80% observed mutation rate

Was für eine Berechnung wurde verlangt? Wo steht die?

Welche Methoden zur Erstellung von phylogenetischen Bäumen gibt es? [L4]

Two main categories:

1. based on the distance (the amount of dissimilarity) between pairs of sequences, computed by sequence alignment; the distance-based methods assume that all involved sequences are homologous

Clustering methods: Starting from the most similar sequence pairs in the distance matrix one tree is computed

UPGMA, The simplest method of tree construction, stepwise, sequential clustering algorithm [L4, S.47ff, sehr genau]

Neighbor Joining (NJ) builds a tree like UPGMA by a stepwise reduction of the distance matrix but it does not assume equidistance between root and taxa.

Optimality based algorithms:
Many alternative tree topologies are compared and one that has the best fit between the estimated distances in the tree and the actual evolutionary distances is chosen

Fitch-Margoliash (FM) selects the best tree among all possible trees by searching for the lowest squared deviation between of actual distances and calculated tree branch lengths

Minimum Evolution (ME)

Least Squares

2. Based on discrete characters (sequences), the basic assumption is that characters at corresponding positions (in a multiple sequence alignment) are homologous among the sequences. Therefore the character states of the common ancestor can be traced from this dataset.

Maximum Parsimony (MP) chooses a tree that has the fewest evolutionary changes or shortest overall branch lengths, based on Occam's razor

Maximum Likelihood (ML) uses probabilistic models to choose the best tree (with the highest likelihood of reproducing the observed data)

Bayesian Analysis (BA)

Vergleich Clustering methods und Optimality based algorithms [L4]

s. letzte Frage (1.)

Gene prediction in Eukaryotes Prokaryotes [L3]

RNA polymerase is a multisubunit enzyme
one subunit recognizes sequence elements 10 and 35 bp before start site
Transcription factors enhance or inhibit RNA polymerase
Bind specific DNA sequences (regulatory elements), can be 100s of bp away
One operon may encompass multiple genes

Start codon ATG (Methionine) or sometimes [TG]TG
Ribosomal Binding Site (RBS), Shine-Delgarno sequence in bacteria often AGGAGGT (helps to locate the start codon)
End codon TAA, TAG, TGA
End of the transcription operon (several proteins)

Determine open reading frame from 6 possibilities
Find long sequences without end codon
Find start and RBS signals
Translate gene into protein and match against homologs

Gene prediction in Eukaryotes [L3]

More complex than in Procaryotes: RNA polymerase I, II, III
RNA polymerase II is responsible for mRNA synthesis
Every gene has its own promoter
Many more transcription factors; RNA polymerase II does not bind the promoter itself, but through the transcription factors

Eukaryotic genomes are much larger, with a low gene density
Only 3% human genome actually codes -> difficult to find genes
Splicing (exons and introns) make it difficult
Start codon within a Kozak sequence: CCGCCATG

Ist die Gene prediction bei Eukaryonten oder Prokaryonten schwieriger? Warum? [L3]

It's much harder in Eukaryotes:
Eukaryotic genomes are much larger, with a low gene density
Introns/exons
Many more transcription factors
Regulatory elements can be far away from the initiator sequence

Warum ist es schwierig genregulatorische Sequenzen und anderes so kurze Sequenzen vorherzusagen? [L3]

Every gene seems to has its own set of regulatory motifs
Not translated into proteins to check sensible areas
Short sites: 6 - 8 nucleotides, which often appear in random stretches as well

Es gibt ab Initio und homology based methods zur Gene prediction, welche sind besser und warum?

Homology ist natürlich zu bevorzugen weil Vorinformationen vorliegen.

Ab initio based predictions:
Based on the sequence data only: detect gene signals
Start and stop codon, intron splice signals, transcription factors, ribosomal binding sites, poly-adenylation sites
Coding parts and non-coding parts have different occurrence statistics

Homology based predictions:
Compare the sequences to known genes (or proteins coded by them)

Zwei Sequenzen waren gegeben in der Form n-(X-Gly-Y)-(X-Y)- Dann Text über Collagen und Fibrinogen und dass eines aus Helices und eines aus Sheets besteht und eines hauptsächlich Serin und noch eine AA hat und das andere Glycin und noch eine andere AA. Welches Protein würden sie welcher Sequenz zuordnen? [L3]

s. Regular Expression