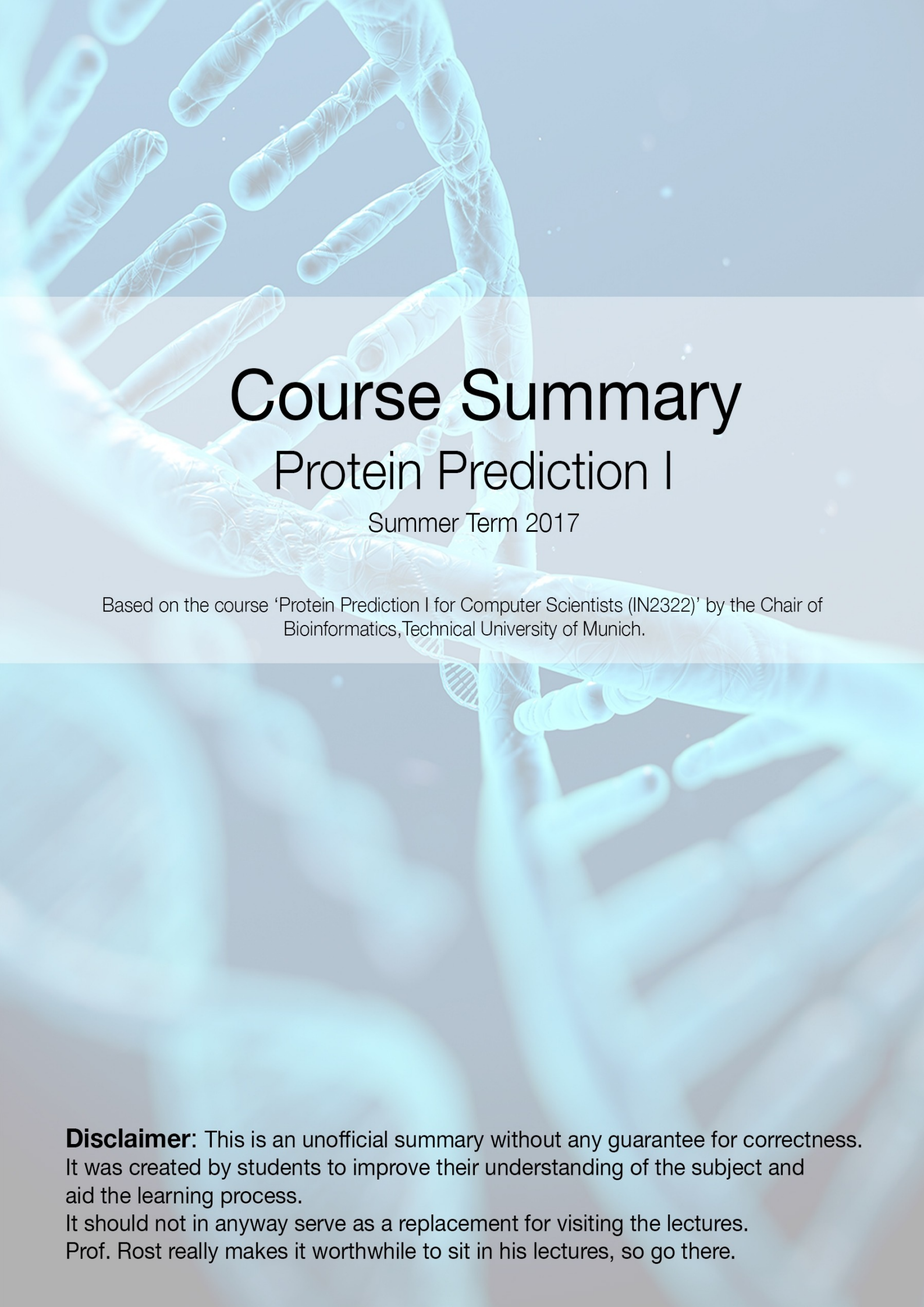


Course Summary

Protein Prediction I

Summer Term 2017

Based on the course 'Protein Prediction I for Computer Scientists (IN2322)' by the Chair of Bioinformatics, Technical University of Munich.

Disclaimer: This is an unofficial summary without any guarantee for correctness. It was created by students to improve their understanding of the subject and aid the learning process.
It should not in anyway serve as a replacement for visiting the lectures.
Prof. Rost really makes it worthwhile to sit in his lectures, so go there.

Table of Contents

Introduction 1.1

1. Lectures 1.2

    1.1 Introduction: Bioinformatics 1.2.1

    1.2 Introduction: Structure 1.2.2

    1.3 Alignments 1 1.2.3

    1.4 Alignments 2 1.2.4

    1.5 Comparative Modeling 1.2.5

    1.6 Secondary Structure Prediction 1.2.6

    1.7 Secondary Structure Prediction 2 1.2.7

    1.8 Secondary Structure Prediction 3 1.2.8

    1.9 Membrane Structure Prediction 1.2.9

    1.10 TMSEG 1.2.10

    1.11 Beta Membrane and Accessibility 1.2.11

    1.12 Protein Disorder 1.2.12

2. Exercises 1.3

    2.1 Introduction 1.3.1

    2.2 Biological Background 1.3.2

    2.3 Protein Structures 1.3.3

    2.4 Alignments 1.3.4

    2.5 Resources for BioInformatics 1.3.5

    2.6 Machine Learning 1.3.6

    2.7 Homology Modeling 1.3.7

    2.8 Wrap Up 1.3.8

3. Exam Questions 1.4

    3.1 Lecture Questions 1.4.1

    3.2 Exercise Questions 1.4.2

    3.3 Question Catalogue 1.4.3

Protein Prediction I

Course Summary, Summer Term 2017

> **tl;dr:** This purpose of this document to collaborativeley create a both concise and detailed course summary of the *Protein Prediction I* Lecture from 2017 Summer Term at TUM.
>
> To learn as effective as possible, I would like to encourage everyone to engage in the discussion evolving around the content of this document. If you have questions or challenge what someone else wrote please do so in a **constructive way**. We are all new to the subject of Protein Prediction and mistakes happen. Let's learn from them together!

Official Lecture Resources

Lecture Homepage: https://www.rostlab.org/teaching/ss17/pp1cs

Lecture Wiki: https://i12r-studfilesrv.informatik.tu-muenchen.de/sose17/pp4cs1/index.php/Main_Page

Youtube Channel: https://www.youtube.com/channel/UCU6j8BG4RbEtTgyIZJ6Vpow

Getting Started

This document is set up a **Gitbook** and hosted on **Github**. When you read this, you were already granted access to the repository so the first step is done.

The easist way to start contributing is to download **Gitbook Editor** (available for Mac, Linux, Windows) from here.

Before you add / change anything, please read through the Contribution Guide.

Contribution Guide

> Tell others what you work on | Write meaningful commit messages | Push often | Use American English

Why is there a contribution guide? I think it is in everyone's best interest to keep this summary as easy to understand as possible for everyone. This guideline should help to maintain consistency across the entire document.

Each section may contain a short additional information on how to format things specific to that section. Please have a look there as well.

1. Adding new content

1.1 Adding minor updates

If you add minor updates, like the answer to a single question, you can do this on the `develop` branch directly. Make sure your commit has a meaningful message.

1.2 Adding major updates

If you add major updates, like several related changes (e.g. an entire lecture summary), go along as follows:

1. Add a new **issue** on Github, describing what you are working on
2. Create a `feature/<issue-name>` branch and add your changes
3. Open a pull-request to merge back into `develop` and add the other contributers as reviewers
4. Once the pull request is merged, delete your feature branch and close the issue by referencing the merge commit

Why so complicated? This way the issues reflect new changes and are transparent for all contributors.

2. Challenging existing content

If you find obvious mistakes (typos, clearly wrong statements) just change them directly.

If you are challenging statements, answers to questions etc. which might not be trivial to understand go along as follows:

1. Open a new **issue** on github.
2. Reference the the statement in question you consider to be wrong
3. Provide an explanation why you think it is wrong
4. Provide your correct solution.

3. Adding new contributors

The purpose of this document is to foster collaborative learning - hence to make this as inclusive as possible. This being said, too many collaborators would probably lead to chaos . If you know other students personally, you want to add to the project shoot me a message and we will figure it out.

1. Lectures

- 1. Lectures
  - 1.1 Introduction: Bioinformatics
  - 1.2 Introduction: Structure
  - 1.3 Alignments 1
  - 1.4 Alignments 2
  - 1.5 Comparative Modeling
  - 1.6 Secondary Structure Prediction
  - 1.7 Secondary Structure Prediction 2
  - 1.8 Secondary Structure Prediction 3
  - 1.9 Membrane Structure Prediction
  - 1.10 TMSEG
  - 1.11 Beta Membrane and Accessibility
  - 1.12 Protein Disorder

1.1 Introduction: Bioinformatics

02.05.2017 | **Slides** | **Lecture Recording**

1. Defintions

Computational Biology: Biology Replacing experiments by computers (including neurobiology, image processing)

Bioinformatics: anything that has to do with storing and using the information about bio-sequences

2. Biology Introduction

Central to biology is the question: *How does life work?*

Question: What is common to life?

> DNA, Protein, RNA

Question: How many bacteria do we care around?

> About 2 kilos. Humans carry around more bacterial DNA than human DNA.

Question: Which elements make up life?

> - 65.0 % - O, Oxygen
> - 18.6 % - C, Carbon
> - 9.7 % - H, Hydrogen
> - 3.2 % - N, Nitrogen
> - 1.8 % - Ca, Calcium
> - 1.0 % - P, Phosphorus

Question: What is life? Can you define it?

> Descriptive definitions of life:
>
> - Homeostatis (regulation of internal environment to maintain constant state)
> - Organization (Unit: Cells)
> - Metabolism
> - Growth
> - Adaptation
> - Response to stimuli
> - Reproduction

Question: Are viruses life?

> Strictly speaking NO. Viruses on their own cannot replicate and thus are not alive. However, one could say that viruses are alive / represent life once they infected a cell and replicate.

2.1 Organisms

Different Type of Cells:

Prokaryotic Cells: Mainly found in bacteria and archaea.

- no nucleus
- usually unicellular
- no cell organells

Eukaryotic Cells: Found in animals and plants

- nucleus
- usually mulicellular
- cell organelles

Note: *The density within cells can be described as almost solid.*

Note: Different organisms use the same amino acids for proteins. However, they differ in their codon usage (which RNA triplets are translated into which amino acid).

Questions

Question: What is the smallest building block of life that can replicate?

> cells

Question: How many different cells are in a typical human?

> 200

Question: What are the parts of the cell called?

> organelles

Question: Which part of the cell is called the "powerhouse"?

> mitochondria

Question: What part of a plant is involved with photosynthesis?

> chloroplast

Question: What is mitosis?

> cell division

Question: Who first used the term cell?

> Robert Hooke

Question: How many elements are found in amounts larger than trace amounts (0.01%) in our bodies?

> 11

Question: When communities of living things interact with non living things they are called ... ?

> ecosystem

Question: The most common molecule in the human body is ... ?

> Water: H_2O

Question: What do bacteria have in common?

> Single Cells

2.2 Genes

Question: What is DNA made out of?

> DNA is a linear polymer out of 4 bases / nucleotides. DNA exists in cells mainly as a two-stranded structure called the double helix. Each of the bases has a complementary base.
>
> - G: Guanine => Cytosine
> - A: Adenine => Thymine
> - T: Thymine => Adenine
> - C: Cytosine => Guanine

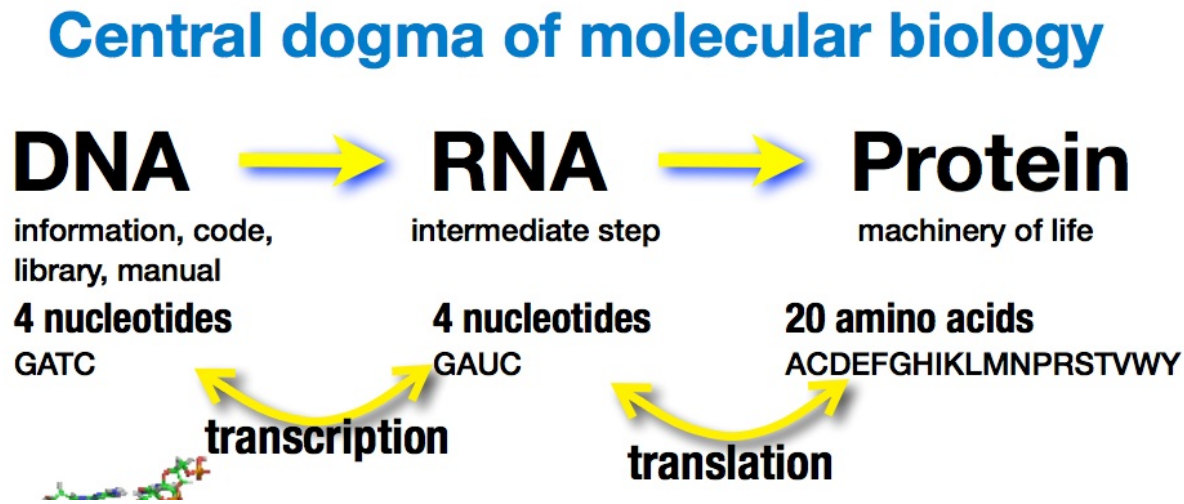
Question: What is RNA made out of?

> RNA is a single stranded linear polymer out of 4 bases / nucleotides.
>
> - G: Guanine
> - A: Adenine
> - U: Uracil
> - C: Cytosine

Question: What is a gene?

> A gene is a region of DNA, which contains all information for the creation of an entire RNA strand (= protein).

2.3 Central Dogma



DNA: Stores genetical information in a 4 letter alphabet. Double stranded helix.

RNA: Working copy of DNA, needed to produce a protein. (Oversimplification). Single stranded. Different types of RNA.

Protein: Composed of 20 letter alphabet (amino acids). Machinery of Life: Proteins do the work in our body.

Transcription: Process of turning a part of the DNA (a gene) into RNA.

Translation: Process of turning a RNA strand into a protein by a Ribosome. Each amino acid is encoded as a RNA nucleotides triplet.

In rare cases it is also possible that RNA translate to either RNA or DNA

```
# Note: SEQUENCE leads to STRUCTURE leads to FUNCTION. Always!
```

3. Protein Introduction

Question: How many proteins does a typical human have?

> Between 20.000 and 25.000 different kinds of proteins.

Question: What are functions of proteins?

> - Defense (e.g. antibodies)
> - Structure (e.g. collagen)
> - Enzymes (metabolism, catabolism)
> - Communication / Signaling (e.g. insulin)
> - Ligand binding / Transport (e.g. hemoglobin)
> - Storage (e.g. ferritin)

Question: How many residues long are typical proteins?

> Between 35 and 30.000 residues. The median is around 400.

Question: Do proteins consist of units?

> Proteins are built up of several domains. Most proteins have more than 2 domains.

Question: How many proteins are known?

> About 85 millions sequences are known. However, the 3D structure (experimentally determined) of only 120.000 proteins is known.

Question: Is this gap (known sequences vs known 3D structure) expected to increase?

> Yes, the gap is expected to increase. The amount of new sequences has increased drastically (far faster than Moore's Law) in the past. This is expected to continue. Advances in experimentally determining protein 3D structure could only improve marginally, but today experimentally determining the 3D structure of a proteins still costs about 100 000 EUR.

1.2 Introduction: Structure

04.05.2017 | **Slides** | **Lecture Recording**

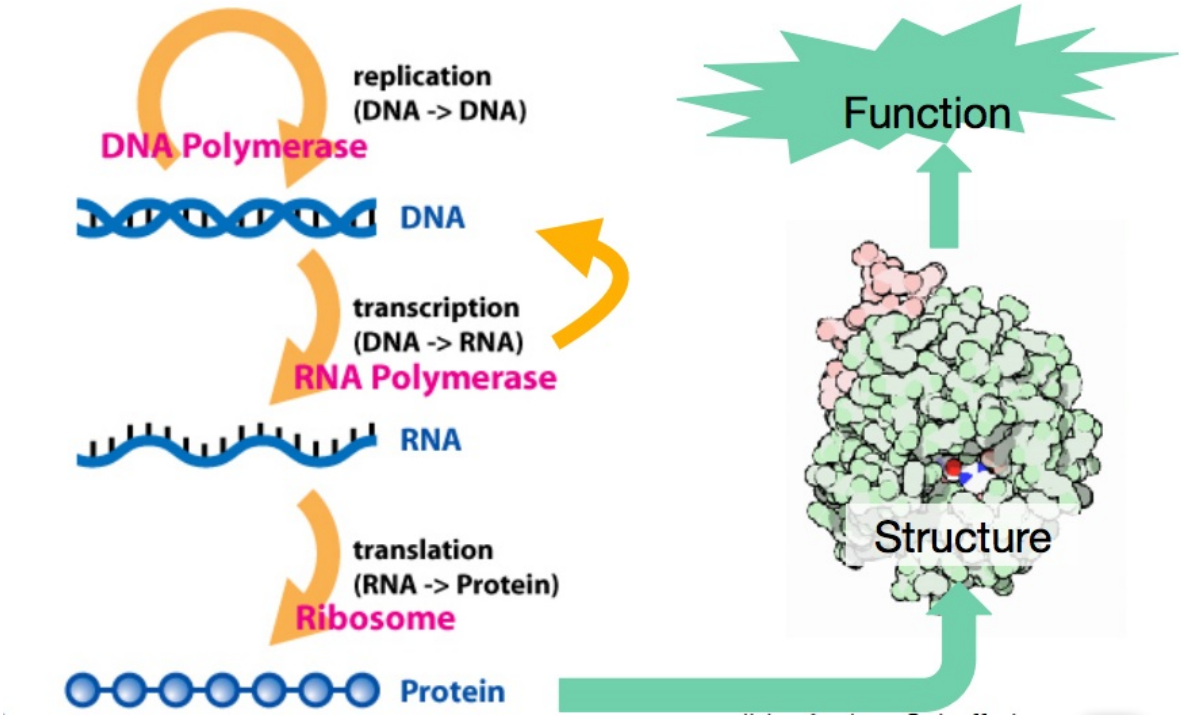
1. Recap

Common to Life: DNA / Cells

Proteins: Machinery of Life - they do everything that needs to be done

- about **85 Million** known protein sequences
- about 120 000 known 3D structures of proteins in PDB
- between 20.000 and 25.000 proteins in a typical human
- protein length (in amino acids): 35 - 30.000, with a median around 400

Central Dogma / Informationflow: DNA -> RNA -> Proteins



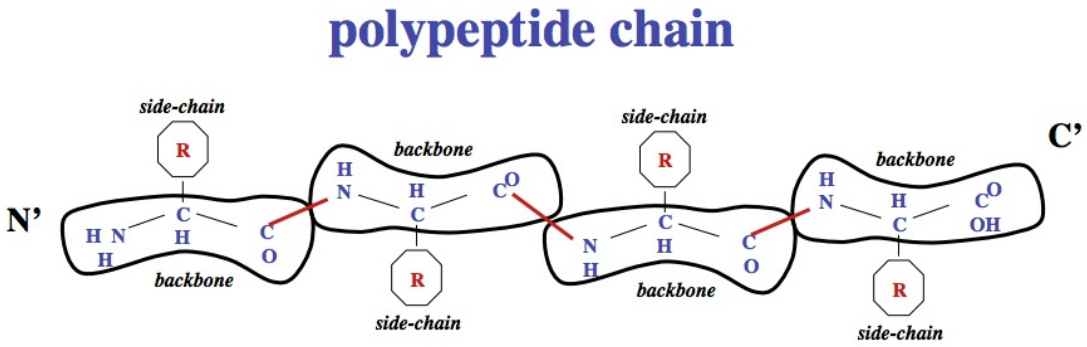
Translation: Proteins are made up of amino acids (20 different kinds). Each amino acids is encoded by a nucleotide triplet (codon) of DNA / RNA.

2. Proteins and Domains

```
# It is important to realize that every representation of a protein (sequence, image, ...
) is
# only a representation of reality.
```

2.1 Amino Acids

Proteins are built up out of a chain of amino acids. These amino acids are joined into a **linear polypeptide chain**, a protein. Each protein is therefore a combination of the **20 different types of amino acids.**



- Each **residue** (amino acid) in this chain has a backbone and a side chain
- Different amino acids have **different side-chains**
- Each amino acids has **the same backbone**, along which they are chained
- Proteins are always chained up from the **N-Terminus** to the **C-Terminus** in a condensation reaction (a H20 molecule is released)

Side Chains

Amino acids only differ in their side chains. These side chains determine the chemical properties of the respective amino acid. There are the following *features* an amino acid can have:

- polar (hydrophilic, likes water)
- non-polar (hydrophobic, avoids water)
- acidic (negatively charged)
- basic (positively charged)

Question: How many different amino acids are there?

> 20

Question: How do amino acids differ? What do they have in common?

> Different amino acids have different side-chains, which influence the chemical features of the respective amino acid. All of them share the same backbone.

Question: In which different feature groups can you categorize amino acids?

> polar, non-polar, acidic (negatively charged), basic (positively charged)

Question: Draw the basic chemical structure of an amino acid.

```
H   R   O
 \  |  //
  N-C-C
 /  |  \
H   H   O-H
```

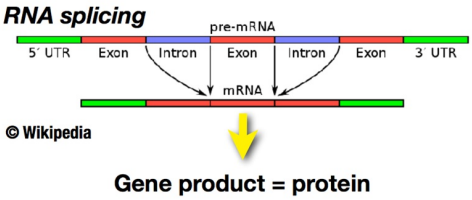
Question: How are amino acids linked together to form a protein?

> In the translation process, a **Ribosome** translates a **mRNA** strand to a protein, by decoding the RNA triplets into amino acids and then linking the amino acids by peptide bonds. They chaining ALWAYS happens from the **N-Terminus** to the **C-Terminus** releasing an H20 molecule as part of the reaction.

2.2 Protein Structure

> **What is a gene?**

A gene is a region of DNA, which contains all information for the creation of an entire RNA strand. (= protein)



UTR: Untranslated region (leader sequence, header sequence)

Exon: Part of a gene that will encode a part of the final mature RNA (and thus protein)

Intron: Part of a gene that will be removed by **RNA Splicing** before the protein is translated

2.3 Domains

Definition: *If I took a sequence out of a protein, string it up and put it into solvent, it adopts a unique 3D structure on its own.*

Proteins are built out of several such substructures. The question is **Can we guess domains from sequence?** By aligning and comparing proteins with know 3D structure, it is possible to find common, overlapping domains that adopt the same 3D structure across different proteins.

Question: What is the definition of a 'domain'?

> A domain is a protein sequence, which when put into solvent adopts a unique 3D structure on its own.

Question: How many domains does a protein have?

> - 61% of proteins in the PDB are single domain
> - 28% of proteins in the PDB are in 62 proteomes
>
> **Problem:** This is a biased view on proteins. The 3D structure of Single-Domain-Proteins is easier to experimentally determine, so more Single-Domain-Proteins have been analyzed.

Question: Can domains overlap?

> Yes, it can happen. However, it is not what is typically observed .

3. 3D Comparisons

There are many different methods for 3D alignments. The points is that comparing 3D structures is highly non trivial and ultimately comes back to the intuition about comparing 3D objects.

Question: How can we compare 3D structures?

One solution would be to align the corresponding residues of both sequences / 3D structures and take the **Root Mean Square Deviation**. (If one pair lies very far apart, it will result in an extremely low score)

> $RMSD(A,B) = \sum_{i=0}^{n}(r_i{}^a - r_i{}^b)^2$
>
> If the score is below a certain threshold, it is a match, otherwise it is not.

Question: How can align and compare the structure of 2 proteins?

> 1) Find the corresponding points (residues that match in 3D)
> 2) Find Superposition independent of domain movements and calculate score (e.g. RMSD)

Question: Why is global protein comparison most of the time impossible?

> The definition of protein enforces a per residue comparison (no scaling). Hence only proteins of the (almost) the same length can be compared globally. Since proteins are between 35 and 30.000 residues long, global comparison does not make sense in most of the cases.

Question: What is the difference between global and local alignment?

> In **global alignment** two structures / sequences are compared from beginning to end (compare the whole thing).
> In **local alignment** however, subunits (domains) of the proteins are aligned. (Problem: What is a valid unit? Where to cut?)

Question: How to decide what is a valid unit for local comparison of 2 proteins?

> (I couldn't identfiy a valid answer in the lecture recording)

Question: Which comparison not using cartesian RMSD could be used for comparison?

> 2D distance map: difference of differences. Only information about the chirality (mirror image) is lost.

1.3 Alignments 1

11.05.2017 | **Slides** | **Lecture Recording**

1. Recap

> **Sequence** leads to **Structure** leads to **Function**

Question: Why compare 3D shapes, when we are after function? Why not compare function?

> Because ...
>
> - we cannot compare function directly
> - structure is related to function
> - we CAN compare 3D structures
> - sometimes: similar structure -> similar function

Question: How do we get protein 3D shapes?

> - primarily by experiment (most accurate)
> - computational biology (most inferences)

Question: How much does it cost to experimentally determine the 3D shape of a protein?

> Today it costs on average about 100 000 $ per protein.

2. Tree of Life

- ### All life is related (common ancestor)

- 3 sections of tree of life
  - prokaryotes
    - (unicellular) bacteria
    - archaea
  - eukaryotes (plants, animals, ...)

Homology: Here (in the context of genes), it describes proteins originating from a common ancestor.

Definition of Species: We are talking about two different species, once they cannot produce fertile offspring together. (Example Bono an Chimpanzee)

Question: What are the 3 sections found in the tree of life?

> bacteria, archaea, eukaryotes

Question: What does Homology stand for?

> Here (in the context of genes), it describes proteins originating from a common ancestor. It is also frequently used to describe 'similar structure' for genes / proteins.

3. Pairwise Sequence Comparison

Correct alignment: We need an objective function

- simplest objective function: percentage of letters which are identical
- more complicated functions describing a match

BUT: the match score itself ignores, what we are after - biological similarity in function

Alignment

To find the optimal superposition of two sequence, it is first necessary to define what 'optimal' means.

Global Alignment:

- Align all residues from the beginning to the end
- Needleman-Wunsch

Local Alignment:

- Best match for locally aligned regions
- Smith-Waterman

How do we align 2 sequences?

Basically brute force: Visually (moving around), computationally (dynamic programming)

Dynamic Programming Algorithm: See Exercise 2.4 Alignments

Gap insertion penalty: Each wildcard (gap) used when aligning 2 sequences has a certain cost.

- Linear gap penalty: N gaps cost N*x
- Affine gap penalty: opening gaps become more expensive
  - Gap open: cost 10x
  - Gap extension (elongation): costs x

Local vs Global Alignment: What is better?

Question: What is better? High sequence identity of a short (local) sequence, or worse sequence identity when matching a longer sequence? How can we decide?

> Compile the probability of randomly matching a sequence considering the background distribution. The result of this would be a substitution matrix such as BLOSUM62.

Question: Is identity the best way to match two sequences?

> Not necessarily: What we really find is similar biological function. Some amino acids might have similar biophysical features and could be swapped without any significant influence on the structure of the protein. Such matches should also be considered 'postive'.

> Building a scoring matrix based on evolutionary conserved residues does optimize the algorithm. (e.g. BLOSUM62)

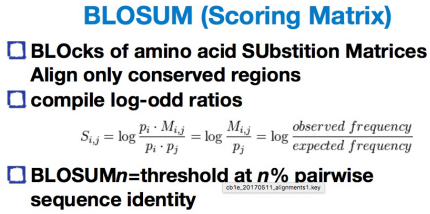
Question: What is the biological assumption behind an insertion when comparing sequences?

> Through evolutionary changes in the DNA (e.g. a point mutation) a new bump (= amino acid(s)) was introduced. Implicitly it is also assumed similar structure -> similar function.

Question: Why do linear gap penalties not model the reality of related genes / proteins well?

> With a linear gap penalty (N gaps cost N*x) equally distributed gaps would be as expensive as clustered gaps. Biologically, gaps clustered to blocks, are however far more likely to occur, while the protein maintains similar structure / function.
> It is more realistic to use an **Affine gap penalty** with higher costs for opening a new gap.

BLOSUM62



Today many more substitution matrices exist.

Interactive Tool to practice dynamic programming: http://melolab.org/sat

Question: Does dynamic programming give the best solution?

> Yes, dynamic programming produces one optimal solution. (There could be others, though)

Question: What are issues with dynamic programming?

> - Time used: O(n^2)
>   - Especially a problem, when comparing one protein agains the entire database.
> - How to choose parameters?
>   - Gap penalties
>   - substitution matrix

Question: How can we speed up the alignment of sequences?

> 1) Hashing (fast and dirty). e.g. BLAST

Question: How does BLAST (Basic Local Alignment Search Tool) work?

> 1. Start with indexed (hashed) seeds (words of size = 3) and find matching proteins
> 2. Extend matching 'words' into both directions
> 3. Begin dynamic programming from these strong local hits

4. Multiple Sequence Comparison

1.4 Alignments 2

16.05.2017 | **Slides** | **Lecture Recording**

1. Recap

3D Comparison: There is a way to look at proteins in 3 dimensions

- 1D - secondary structure prediction
- 2D - distance map
- 3D - real 3D coordinates

Dynamic Programming

- **Global** (Needleman-Wunsch) or **Local** (Smith-Waterman)
- With or without **Gaps**
  - Linear Gap Penalty: Opening and extending a gap have the same cost
  - Affine Gap Penalty: Opening a new gap is more expensive than extending an existing one
- **Scoring Matrices**
  - For each pair of residues you can read off, how much you gain by aligning these 2 residues

BLAST: *Basic Local Alignment Search Tool*

- Dynamic Programming is slow for large scale comparisons
- Speed search up by hashing words (seed = 3 amino acids / residues)
- After matching a word, try to extend the match by dynamic programming
- **Major Challenge:** Get the statistics right
  - *How significant is a match* against the background probability entire database

Question: What is the major challenge of BLAST?

> Getting the statistics right: BLAST needs to know, h*ow significant a match is*, by comparing it against the background probability of the entire database.

2. Pairwise Alignment Accuracy

PSI: Percentage Sequence Identity

Zones:

- **Daylight-Zone:** PSI, where it can be assumed that from a similar sequence follows similar structure
- **Twilight-Zone:** PSI, where it is not possible to infer similar structure from similar sequence (the signal fades)
- **Midnight-Zone:** PSI, where sequence similarity does not tell anything about structure similarity

(down to random changes)

```
# Note: The Midnight-Zone is, where most proteins of similar structure sit
```

Going deeper into the Twilight-Zone, the following results are to be expected:

1. True Positives go up in absolute numbers
2. False positive increase (drastically) in absolute numbers

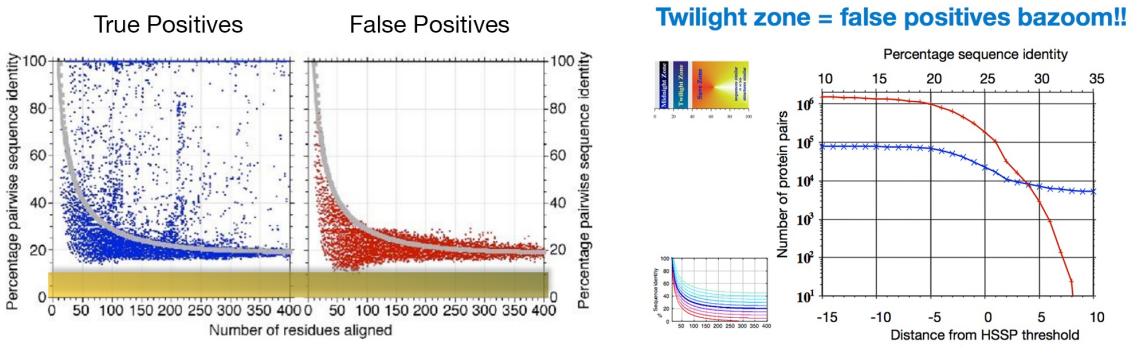
2.1 HSSP Curve

HSSP: Homology-derived Secondary Structure of Proteins

How to get the curve?

1. Get all 3D structures from PDB
2. Remove bias (sequence unique subset)
3. And compare 'all vs redundancy reduced set'
    i. compare 3D structure (e.g. RSMD)
    ii. compare sequence

Result: Sequence Conservation of protein structur



Observations

- Over the curve (daylight-zone), many true-positives (sequence identity -> similar 3D structure)
- Under the curve (upon entering the twilight zone)
    - Explosion of **false positives**
    - but also **significant increase** (x10) of **true positives** (This is why we want to go into the Twilight Zone!!)

Question: Why is it interesting to find similar proteins out of the Twilight / Midnight Zone?

> The Midnight-Zone is, where most proteins of similar structure sit.

Question: Why is it that even with only 40% PSI, we can still assume similar structure? Could we randomly change 60% of the residues in the lab and get a new protein with similar structure?

> - These 60% of changed residues happened under evolutionary pressure and are not

> random
>   - mutations that did not change structure & function survived (we can observe them today)
>   - mutations that did change structure & function most likely did not survive
> - Thus randomly changing 60% of residues in a proteins, would not result in a similar protein

Question: Why are certain proteins / structure multiple times in the PDB?

> - different resolution of 3D structure
> - different goals of publication produced (new) 3D structures
>   - folding sites
>   - binding partners
>   - etc ...

3. Multiple Sequence Alignment

Dynamic Programming cannot be done efficiently with multiple sequences.

3.1 Multiple Alignment Hack 1: Iterative Pairwise Dynamic Programming

Progressive 1

1. Align sequences pairwise int consensus sequences
2. Align resulting consensus

Progressive 2

1. Align all sequences pairwise
2. Start with highest matching alignment
3. Iteratively align sequences into consensus sequence

How to find consensus? Different methods, e.g. the first, the more meaningful aa, ...

3.2 Multiple Alignment Hack 1: Map to Tree / Pairwise

ClustalW/ClustalX

- all against all (pairs) by dynamic programming (varying substitution matrices)
- build **phylogenetic tree**
- slow, dynamic programming, for experts

4. Profiles

Profiles profit from relation of 'families'.

Building up a profile, we can see certain amino acids that are more conserved than others. Computationally we can identify **sequence motifs** that describe such a profile as a regular expression.

PSSM (Position Specific Scoring Matrix)

You could also write a **profile** into a substitution matrix: A matrix of numbers with scores for each residue or nucleotide at each position.

Building a PSSM

1. Absolute Frequencies
2. Add pseudo-counts if necessary
3. relative frequency
4. log likelihoods

Question: How are profiles built up? How are the normal noted down? Do we have to know a specific algorithm?

> **Build up algorithm:**
>
> - Take all proteins of PSI over a certain threshold ...
> -
>
> **Profile Formats:**
>
> - Regular Expression
> - PSSM (Position Specific Scoring Matrix)

Question: What is a PSSM (Position Specific Scoring Matrix)?

> A matrix of numbers with scores for each residue or nucleotide at each position. Built, e.g. by PSI-BLAST.

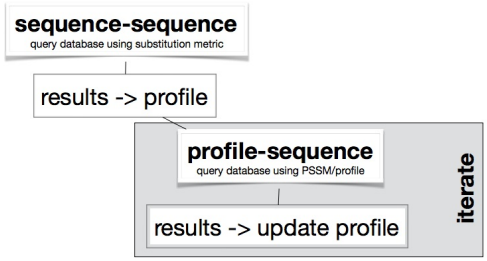
5. PSI-BLAST

Position-Specific Iterative Basic Local Alignment Tool

PSI-BLAST Steps

1) Fast Hashing: Like BLAST, match 'word'
2) Dynamic Programming Extension between matches: BLAST + Smith-Waterman
3) Compile Statistics: EVAL - Expectation Values
4) Collect all pairs and build profile
5) ... compare sequences (profile-sequence) and iterate

Steps involved for profile-based alignments



Question: Which steps are involved in building up a profile with PSI-BLAST?

> **1) Fast Hashing**: Like BLAST, match 'word'
> **2) Dynamic Programming Extension between matches:** BLAST + Smith-Waterman
> **3) Compile Statistics:** EVAL - Expectation Values
> **4) Collect all pairs and build profile**
> **5) ... compare sequences (profile-sequence) and iterate**

Question: Why is PSI-BLAST so fast?

> Because it drastically reduces the length of the comparisons with dynamic programming.

6. Hidden Markov Models (HMM)

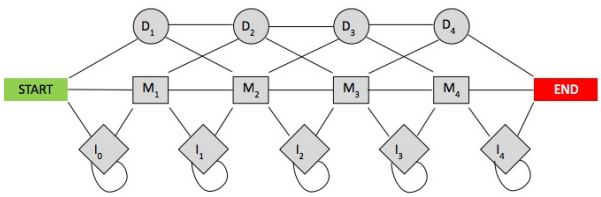
Hidden Markov Model are another method for creating Machine Learning Models. They are a good choice, if the structure of the problem is known beforehand.

- different states
- each state has transition probabilities to the neighboring states / itself

7. HMM for Alignment

Generic Profile HMM for alignment

- Captures matches, insertions, deletions
- Transition and emmission probabilities
- gap penalty handled by variation of transition probabilities
- calculation of probability by multiplying path variables



Entropy in alignment: Consider the residue at position **i**

- BEFORE any amino acid is aligned, we expect a particular amino acid to have some prior

background probability P_0 with entropy H_0

- AFTER the alignment we consider the same column with a *posterior probability* $P_i + priors \rightarrow H_i$.
  We expect $H_i = \begin{cases} 0, & if\ conserved \\ H_0, & if\ varied \end{cases}$
- $H_i - H_0$ reflects the "bits_saved" by the alignment

With small families (few members, little divergence) the entropy is dominated by priors (= the background noise dominates)

8. Genetic Algorithm for alignment

Independence Assumption NOT needed for genetic algorithm

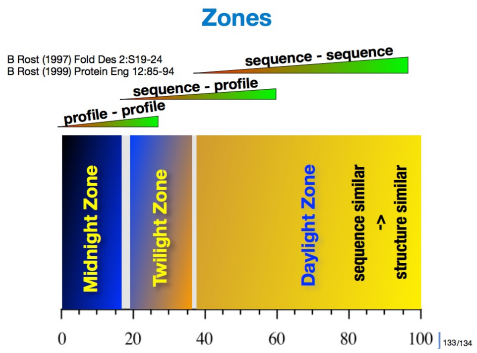
```
# All algorithms so far assumed *Independence between residues*:
# What happens at position i is independent of what happens at position i+x.

# The genetic algorithm does not make this indepence assumption!!
```

- The genetic algorithm works on segments
- through mutations it creates new alignments

T-Coffee: much slower, requires pre-processing, Genetic algorithm

9. Profile-Profile Alignments

Compare Profiles to go even deeper into the **Twilight- / Midnight-Zone**



1.3 Comparative Modelling

18.05.2017 | **Slides** | **Lecture Recording**

---

1. Recap

- Profile-Sequence comparisons are more accurate than sequence-sequence aligments
- Profile-Profile alignments gain even more accuracy

Question: How do you build up a family (profile) of sequences?

> 1. Find proteins of similar sequence with BLAST
> 2. Use the found proteins to build a PSSM (profile)
> 3. Use profile-sequence alignment with the calculated PSSM to retrieve more distant family members
> 4. Add the newly found proteins to the family by recalculating the PSSM

- When building up a profile, start with a high threshold (only very similar sequences are taken), so the profile is not wrong from the beginnig

2. Goal of structure prediction

Sequence uniquely determines structure! ➡ Thus, from a sequence it should be possible to predict 3D structure and function

How would you assess prediction performance?

CASP: Critical Assessment of Structure Prediction

- Yearly event
- Submit predictions for structures, which will be experimentally predicted before a deadline
- Compare (after release of experimental structures) how the methods performed

Current State

- Only Homology Modeling is good
- No general prediction of 3D structure from sequence yet
- BUT: Important improvement in many fields

3. Structure by Experiment

Different Methods to determine 3D structure

- 90% - X-Ray Crystallography
- 09% - Nuclear Magnetic Resonance Spectroscopy (NMR)

- 01 % - Cryo Electron Microscope (Cryo-EM)

X-Ray Crystallography

1. **Grow Crystal:** Force the protein to grow a crystal
2. **Observe Diffraction Pattern**: Shoot x-rays onto crystal and observe the diffraction pattern
3. **Compute Electron Density Map**
4. **Fit observations to atomic model**

NMR

1. Protein has to be in similar solution as naturally
2. Massive Magnets required

Cryo-EM

- worse resolution than other methods
- cheaper than other methods
- *Pushing the boundaries of resolution of Cryo-EM is the future*

Question: Which methods to experimentally determine the structure of a protein exist? How much are they used?

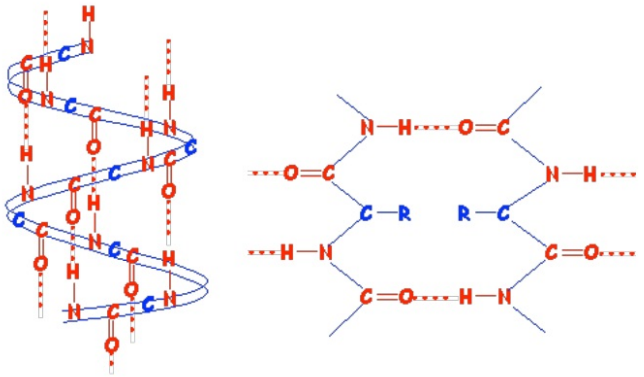
> Fraction of proteins in the PDB by experimental method:
>
> - 90% - X-Ray Crystallography
> - 09% - Nuclear Magnetic Resonance Spectroscopy (NMR)
> - 01 % - Electron Microscope (EM)

Question: How does X-Ray Crystallography Work

> 1. **Grow Crystal:** Force the protein to grow a crystal
> 2. **Observe Diffraction Pattern**: Shoot x-rays onto crystal and observe the diffraction pattern
> 3. **Compute Electron Density Map**
> 4. **Fit observations to atomic model**

Hydrogen Bond Formation

Idea: Secondary structure is completely explained by hydrogen bond formation.



Helix: Hydrogen-Bond between residue **i** and residue **i+4**, which stabilize the helix.
Sheet: Two strands come together to form a sheet by forming hydrogen bonds between them

Question: How to get 1D secondary structure from 3D coordinates?

> Two methods where used to annotate 3D coordinates:
>
> 1) DEFINE, based on geometry (not used anymore)
> 2) DSSP, based on hydrogen bond pattern (coulomb energy)

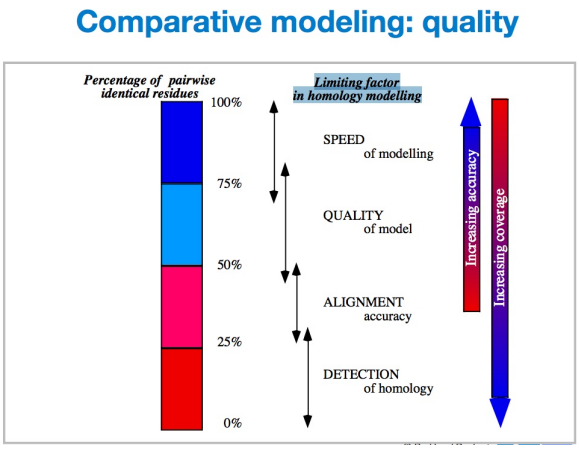
4. Comparative Modeling (=Homology Modeling)

Assumption: Sequence uniquely determines structure and therefore, from similar sequence follows similar structure.

How can we use this to predict 3D structure?

Target: Protein to model
Template: Protein to model from

1. **Identify Template:** Query the PDB for similar sequences to your **Target**
2. **Align Target / Template:** Select the best match as **Template** and assume the **Target** has the same structure
3. **Build Model**
4. **Assess Model**
5. **Refine Model**



Question: How does Homology Modeling (Comparative Modeling) work?

> **Target:** Protein to model
> **Template**: Protein to model from
>
> 1. **Identify Template:** Query the PDB for similar sequences to your **Target**
> 2. **Align Target / Template:** Select the best match as **Template** and assume the **Target** has the same structure
> 3. **Build Model**
> 4. **Assess Model**
> 5. **Refine Model**

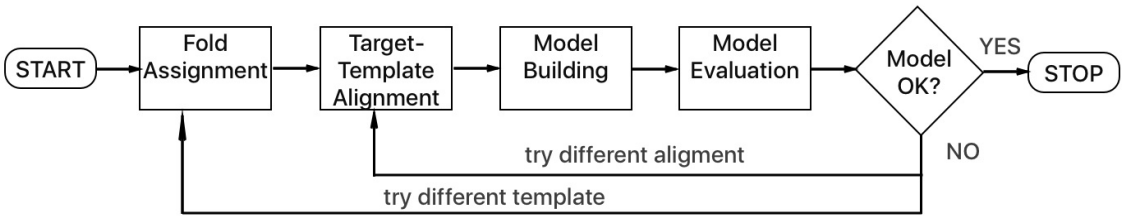
Question: Which tradeoff does comparative modeling face? What are the limiting factors based on PSI (Percentage Sequence Identity)?

> **Tradeoff:** Accuracy vs Coverage
>
> Limiting factor in homology modeling:
> 75% - 100% - Speed of Modeling
> 50% - 75% - Quality of Model
> 25% - 50% - Alignment Accuracy
> 0% - 25% - Detection of Homology

5. Comparative Modeling Methods

5.1 MODELLER

Summary: lots of whistles and bells, downloadable, very accurate



Constraint Satisfaction: use a set of objective functions to check whether the model is plausible

- $C_\alpha - C_\alpha$ distance
- Molecular dynamics
- Langevin dynamics
- Rigid bodies
- Rigid molecular dynamics
- ...

Optimization Steps (run repeatedly)

- explore different local minima

Typical Errors

- side chain packing
- misalignment
- wrong template

Pick the right solution:

- DOPE score (Discrete Optimized Protein Energy)
- based on knowledge based pair potentials

Question: How to handle a missing loop in comparative modeling?

> - One way would be to find similar loops and compute the average over them.
> - Another solution would be to apply molecular dynamics on the loop sequence. (only for shot loops)

5.2 SWISS-Model

Summary: automated, increasingly comprehensive and flexible

Underlying 'Philosophy'

- fully automated
- for non-expert users / experimental biologists
- do less, make less mistakes

Original

1. alignment by BLAST / PSI-BLAST
2. copy to coordinates
3. end

Today: More complicated ...

1.6 Secondary Structure Prediction

??.05.2017 | ??? | ???

```
# Note: One lecture was not recorded. Maybe this one?
#       The content of the slides can be found in the other summaries.
```

1.7 Secondary Structure Prediction 2

01.06.2017 | **Slides** | **Lecture Recording**

1. Recap

Goal of Structure Prediction: Predict the 3D structure and function from an input sequence.

Proteins:

- are formed by stringing up amino acids in a chain
- amino acids are between 35 to 30.000 residues long
- amino acids form substructures, called domains in order to fold
- in principle: a domain put into solvent (water) folds on its own and adopts a unique 3D structure

Zones: There are different 'zones' by percentage of sequence identity, in which we can identify similar structures

- **Daylight-Zone** (100 % - 40%)
  - Sequence - Sequence Alignment
  - Assumption: sequence similar -> structure similar
- **Twilight-Zone** (40% - 20%)
  - Sequence - Profile Alignment
  - more distant relationships
- **Midnight-Zone** (20% - 0%)
  - Profile - Profile alignment
  - even more distant relationships

Global and Local Alignment:

Question: Relate the terms **Local** and **Global alignment** to the terms **Sequence-Sequence** and **Sequence-Profile**.

> Global alignments refers to aligning sequences (proteins) from start to end. Local alignments refers to only aligning parts of the sequences (e.g. 50 residues).
> Throughout Sequence-Sequence, Sequence-Profile and Profile-Profile methods both global and local alignment can be used. I practice mostly local alignment is done.

Comparative Modeling:

- Idea:
  - Find a similar sequence with known structure in the PDB (in the daylight-zone)
  - Try to use the known 3D structure of the similar protein to model the structure of the unknown protein
  - fix physical / chemical errors of the predicted 3D structure and find most plausible 3D

structure

- Reliably predicts **over 40 million proteins**
- However, for **most residues comparative modeling cannot be applied**

2. Secondary Structure Prediction

Secondary Structure Prediction happens in 1D, 2D and 3D. The following chapter will mainly be about 1D Secondary Structure Prediction.

DSSP (Define Secondary Structure of Proteins) algorithm: Has 8 states

- **H** = Helix
- **G** = 3_{10} Helix
- **I** = Pi Helix
- **E** = Extended
- **B** = Beta-bridge, single-strand residue
- **S** = bent
- **""** = loop

Local Sequence determines secondary structure!

- Certain local sequence always form the same secondary structure (α-helix, β-strand, loop).
- Others (penta-peptides) are found in 2 different state, **dependent on their environment**

Question: What would be a simple method to predict secondary structure?

> 1) Take known structure
> 2) Find longest consecutive run of motifs that **ONLY** occur in one of the 3 states: H (Helix), E (Strand), O (Other)
> 3) Check unknown sequence against found motifs

Question: What was the first secondary structure prediction method?

> Assuming that a **Proline** would break a helix, the occurrences of proline in a sequence was used to predict helices.

3. 1st Generation Secondary Structure Prediction

Idea: Build a frequency table over all amino acids, how often they occur in the secondary structure states based on the proteins where the structure is known.
Important: Bias reduction, to make the set table representative for future. Remove all proteins in comparative modeling range.

1. find a unique subset of proteins with known 3D structure (PDB)
2. convert 3D to 1D (secondary structure) with DSSP

Question: Where do we get the secondary structures from?

> From the DSSP, which defines 8 states in total based on H-bond patterns.

Question: What is the 1st generation of secondary structure prediction based on? What was the accuracy? Was is successful?

> - Based on single residues
> - Between 50% and 55% accuracy (Q3)
> - Clearly better than random - so it can be considered a success

How can we measure the performance of secondary structure prediction?

Q3: three-state per residue accuracy

$$Q3 = \frac{number\ of\ correctly\ predicted\ residues\ in\ states\ helix,\ strand,\ other}{number\ of\ residues\ in\ protein}$$

Question: How can the performance of secondary structure prediction be measured?

> One way to do it, would be to calculate the **Q3** accuracy of a method against a test set. The Q3 accuracy is the **number of correctly categorized residues into one of the categories helix, strand, other divided by the total amount of residues.**

4. 2nd Generation Secondary Structure Prediction

Question: How did the second generation of secondary structure prediction improve? Name one algorithm.

> Instead of using only single amino acids, it would consider a sliding window of the residues around a center amino acid.
> **Example:** GORIII, with a Q3 accuracy of 55% - 60%

Question: What were problem of secondary structure prediction until 1994?

> - the maximum accuracy of predictions was expected to be 65%
> - β-sheet prediction was below 40%
> - many predicted segments where too short to appear in nature

5. Introduction: Neural Networks

Goal: Use the representation of a set of examples (training set) for which the mapping *input -> output* is known to iteratively refine the weights of the connections betweens output and input units so that the error is minimized.

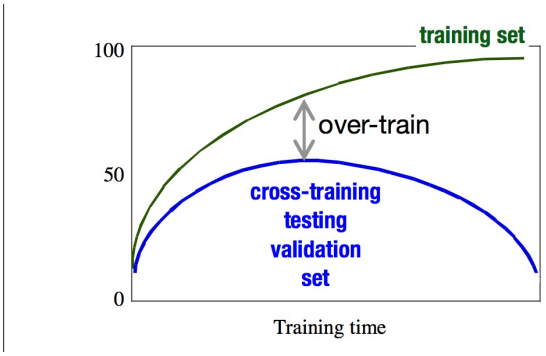
Principles of neural networks

- Free Variables: $Connections\ \{J\}$
- Output: $out_i = \sum_{i=1}^{N^{in+1}} J_{ij} in_j$
  - in_j value of input unit j
  - out_i value of output unit i

    - J_{ij} connection between input unit j and output unit i
- Error: $E = \sum_{i=1}^{N^{out}} (out_i - des_i)^2$
    - out_i value of output unit i
    - des_i secondary structure state observed for central amino acid for output unit j

Training: Change of connections $\{J\}$ such that E decreases (e.g. gradient descent)

Problem: Overtraining - happens if the network becomes too specific to the actual training set and looses accuracy for predicting unknown input. The point when to stop training can be found by using **cross-training, testing, validation sets**.



Cross-Validation: Split your available dataset into 3 sections

1. **Training** (50%): used to train ML algorithm
2. **Cross-Train** (25%): used to find threshold when to stop training and tweak parameters
3. **Testing** (25%): used ONLY to assess performance / accuracy of final ML algorithm

Question: How can the introduction of a new hidden layer in a neural network be described by means of a simple graph?

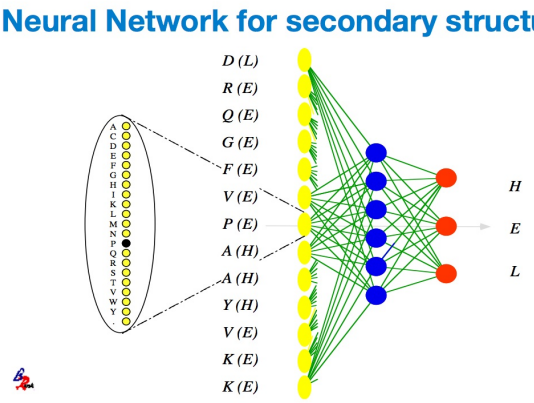
> Each new hidden layer basically introduces a new 'decision line' which can separate datapoints into different categories.

Question: What is cross-validation in the context of Machine Learning and why do we need it?

> Cross Validation in a method for estimating the performance of a predictive model (e.g. a neural network). To use it, the available dataset is split in 3 categories, 1) a training set, 2) a cross-training set and 3) a test set.
> 1) The training set is used to train the model
> 2) The cross-training set is used to estimate the performance of the model after x training steps
> 3) The test set is used to assess the final performance of the model after training is finished
>
> The cross-training set is needed to decide, when to stop training (when overtraining sets in) and to tweak certain parameters before running against the test-set.

6. Neural Networks for Secondary Structure

Goal: Solve the 3 problems at the time [1] accuracy, [2] strand performance, [3] short segments



Input: 13 * 21 input units

- 13 ??
- 20 amino acids + 1 spacer

  However, the final accuracy was only about **62%**

Balanced Training:

- Helices are overrepresented in the training data
- Choose the training data, so all 3 states (helix, strand, other) are equally represented

Result

- overall accuracy dropped to **60%**
- β-sheet prediction improved from **40%** to around **60%**

Question: Did balanced training improve the Q3 prediction accuracy? Which assumption did it prove wrong?

> Balanced training actually decreased the Q3 accuracy. However, it did improve the prediction accuracy for strands significantly, falsifying the hypothesis that strands could not be predicted with local information.

7. PHDSec: Structure to Structure Prediction

We still have the problem of bad segment prediction (too short segments). This is due to the fact that samples from the database are selected at random, loosing information about local correlations.

How can we get information about the local correlation (e.g. length of a helix) into the prediction model?

Solution: Add a second Neural Network, which takes the predicted sequences from the first network as input.

BUT: Accuracy was still only 60% + ε

Question: Which problem did PHDSec solve? How did it accomplish it?

> By introducing a **Structure-to-Structure** prediction model, PHDSec improved the prediction of too short segment. The Structure-to-Structure network would take structure (helix, strand, other) prediction of a sequence as input and predict segments based on them.

1.8 Secondary Structure Prediction 3

08.06.2017 | **Slides** | **Lecture Recording**

1. Recap

Goal of Structure Prediction: Predict the 3D structure and function from an input sequence.

Cell: The density of a cell is like solid state, but proteins are still surounded by water

Relation of Proteins: We can find out about the relation of proteins by comparing their sequence

- direct sequence-sequence comparison only in the daylight-zone
- building up profiles means to 'pick up the implicit evolutionary signal'

Question: Which ways of comparing proteins are there? Why do we need

> - Dynamic Programming (Brute Force)
> - Hashing (e.g. BLAST)

Question: Why are fast search algorithms such as BLAST needed?

> Comparing sequences of length n residues is in $O(n^2)$. For comparing a single pair this is still fine, but comparing one (newly found) protein against all known proteins in the PDB (about 120 000) is impossible. Thus we need 'shortcuts' such as BLAST to speed up the search.

Question: What is the normal approach when you find / analyse a newly found protein?

> 1) Sequence the new protein (if not done yet)
> 2) Run BLAST against the PDB
> 3) Run Dynamic Programming against the results from BLAST

Question: In terms of CPU, is sequence-sequence as fast as sequence-profile?

>

Question: How can it be that even with only 40% sequence identity we assume / observe similar structure?

> The changes in sequences we observe are not random, but follow underlying evolutionary rules. Changes, which affected the structure and thus the function of a protein are simply not likely to survive and thus we do not observe them. Changes, which did not influence the structure / function however, did survive.

Question: Why is protein sequence changing? Why are we mutating?

> - Replication Errors (point mutations)

> - Radiation
> - Viruses

Question: How much do any two unrelated typical humans differ on average?

> On average every pair of humans would differ in one amino acid per protein. (Though, changes cluster)

Question: In a structure to structure network, which additional information could be used to improve the prediction?

> - E.g. redundant information about the sequence, e.g. parts of it.
> - length of protein
> - Is the sliding window at the end / mid / start of the protein?

Question: When training a neural network, how do you choose the next training sample from your test set? Why so?

> Randomly, to avoid correlations

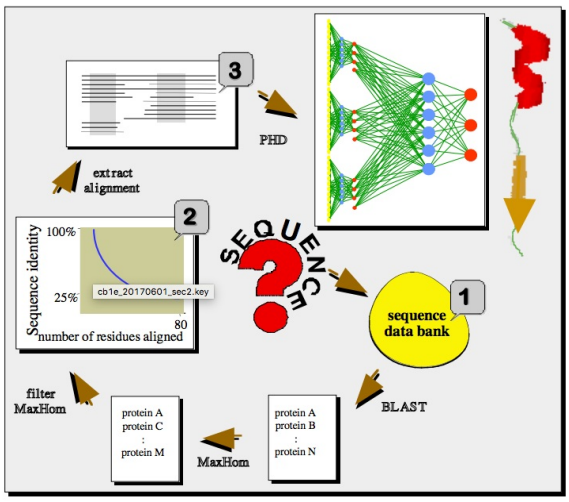
2. 3rd Generation Secondary Structure Prediction

Critical Question: How to improve beyond $60\% + \epsilon$ accuracy?

Evolution improves prediction: An **evolutionary profile** averaged built up over several species implicitly captures the history of an individual protein.

PHD: Neural Network and Evolutionary Information

- Build up the family (profile) for the protein and add it to the input of the network
- Each amino acid in the input now has a probability on how often it occurs in the family



Additional Input for PHDSec network

- family (profile)
- percentage of each amino acid in protein
- length of protein

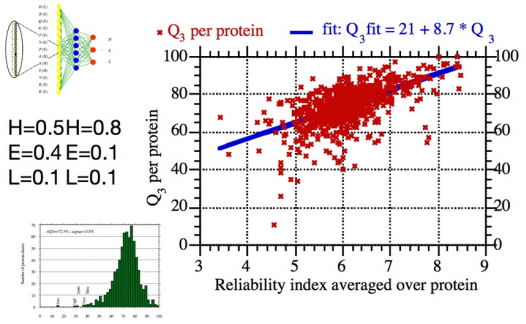
- distance: center, N-term
- distance: center, C-term

Jury decision improve accuracy: All of these input features are fed to different Networks, resulting in many independent predictors (**Jury**). All of these networks add their own 'white noise' to the prediction. The average over all the predictors is better than every single one.

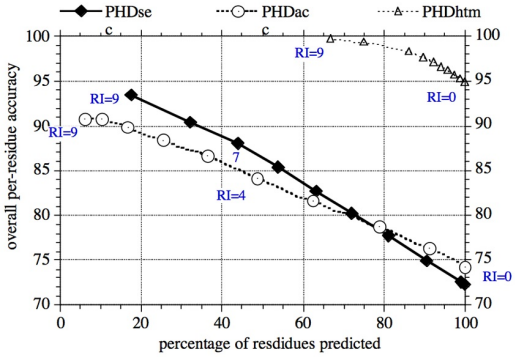
The final accuracy (on average) of ProfSec is about **72%**.

Prediction of correctly predicted residues: In additional, **ProfSec can give an estimation on the strength of the prediction** for each protein. (By counting the 'stars')

Stronger predictions more accurate!



Correct prediction of correctly predicted residues



Global Information improves ProfSec's per protein prediction.

	Q3 (per residue)	Q4 (per protein)
Only Sliding window (local)	72%	70%
Local & Global	72%	75%

Question: How does ProfSec overcome the 60% accuracy hurdle in secondary structure prediction?

> ProfSec uses evolutionary information - the family of the protein - as additional input. Furthermore, other relevant input data (e.g length of protein, distribution of amino acids, ...) are used to build up several different networks that independently predict the secondary structure. Together this jury of networks achieves a more accurate prediction than they would on their own.

Question: How would you build up a family for a protein?

> 1. Search the PDB for proteins in comparative modeling range. (Assumption: same sequence, same 3D structure, same secondary structure)
> 2. Use profile to search in twilight-zone for potential proteins of that family (possibly verify whether the found protein is plausible to have similar 3D structure) and add to family (recompute profile)

Question: How do you get from a sequence to a secondary structure prediction with PHD?

> 1. Use BLAST to find potentially similar proteins in sequence data bank

> 2. For the resulting proteins calculate the sequence identity (homology) with dynamic programming
> 3. Filter all proteins, which are below a threshold of sequence identity (only take those "over the curve")
> 4. Extract the profile by aligning the remaining proteins
> 5. Predict the secondary structure with the sequence and its family as input

Question: Which accuracy does ProfSec achieve on average? What are additional advantages of other secondary structure prediction methods?

> ProfSec achieves a Q3 accuracy of about 72% on average. Additionally it can also predict the strength of the prediction.

Question: Does adding global information improve ProfSec prediction?

> Yes it does. While the Q3 accuracy (per residue) is not improved, the Q4 accuracy (per protein) does improve.

3. Proper comparison of methods

For a meaningful comparison the methods should

- use the same (meaningful) measure (e.g. Q3)
- use the same dataset
- split training / testing
  - there must not be an overlap between sets
- is the difference (in accuracy) significant (= difference > standard error)
- was the test set not used for making decisions?

1.9 Membrane Structure Prediction

13.06.2017 | **Slides** | **Lecture Recording**

1. Introduction Membrane

Requirements for Cell Membrane

- separate the content of the cell from its surroundings
- control traffic into and out of the cell
  -
    - keep malicious things out
    - let good things in
- must be a dynamic structure

Main components of the cell membrane

1. Carbohydrates
2. Cholesterol
3. Phospholipids
4. Proteins

Phospholipids

- form the barrier that separates the inside of a cell from the outside
- pospholipids are arranged in a **lipid bilayer**
  - **Inside:** fatty acid tails (non-polar, hydrophobic)
  - **Outside**: phosphate group (polar, hydrophilic)

Membrane Proteins

- Provide several functions to the cell
  - help to be recognized by immune cells
  - transport proteins control substance flow in and out of the cell
  - receptor proteins bind hormons, which can change cell function
  - provide structural stability
- Membrane proteins can (easily) shift around laterally

Membrane Proteins are especially important for drug targets

```
# Note
Similar to the membrane, proteins also tend to have a **hydrophobic core**.
Membrane proteins however, tend to have a **hydrophobic outside** and an **hydrophilic co
re**.
```

Question: What are the requirements of a cell membrane?

> - separate the content of the cell from its surroundings
> - control traffic into and out of the cell
>   - keep malicious things out
>   - let good things in
> - must be a dynamic structure

Question: What are the 4 main structural components of the cell membrane?

> Carbohydrates, Cholesterol, Phospholipids, Proteins

Question: What is the cell membrane mainly made out of?

> The cell membrane is a so called **lipid bilayer** of **phospholipids**. Phospholipids have a non-polar, hydrophobic tail (membrane center) and a polar, hydrophilic head (outside of membrane).

Question: What are functions of membrane proteins?

> - help to be recognized by immune cells
> - transport proteins control substance flow in and out of the cell
> - receptor proteins bind hormons, which can change cell function
> - provide structural stability

Question: Can membrane proteins easily move around?

> It depends:
>
> - Membrane proteins can easily move laterally
> - But it is hard to move into / out of the lipid bilayer

2. Introduction Transmembrane Helix (TMH)

Although membrane proteins are especially interesting for drug targets, there are only limited 3D structures to be found in the PDB. Mostly, because it is extremely difficult to put them into a crystal in their 'natural' membrane environment.

There are essentially 2 important questions when it comes to TMH prediction:

- How many helices go through the membrane?
- In which direction do they go through the membrane? (topology)

Question: Why are there so few membrane proteins in the PDB?

> It is particularly difficult to experimentally determine the structure of membrane proteins due to the special environment they naturally occur (the membrane).

Question: What are they key questions TMH prediction tries to answer?

> - How many helices go through the membrane?

> - In which direction do they go through the membrane? (topology)

3. TMH Prediction

Question: Why could be a plausible reason why PHDSec failed for predicting transmembrane helices?

> Unlike 'normal' proteins, transmembrane proteins have an hydrophobic outside and a hydrophilic inside.

What could be a strategy to come around that? **Build up a hydrophobicity index.**

- There are different hydrophobicity scales, optimized for different problems

Identifying hydrophobic regions

- Whenever the hydrophobicity is over a certain threshold, consider it a membrane helices
- except the hydrophobic residues over (the lower) threshold are not long enough for a TMH (20 residues)

Identifying Topology: What is inside/outside of a TMH?

- **Positive Inside Rule:** Looking at the parts which connect TMHs within on protein, they look different depending on which side of the membrane they are: **There is a excess of positively charged residues on the inside.**

Question: How should we choose the threshold for the hydrophobic regions?

> 1. Predict the hydrophobicity for the protein
> 2. Assign a positive inside-out
> 3. choose the threshold to **optimize the inside out difference**

Question: What is the Positive Inside Rule and what is it used for?

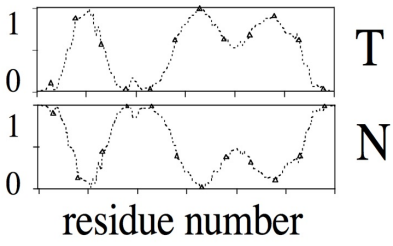
> The positive Inside Rule is used to find the topology of transmembrane proteins: The loops connecting TMHs on the inside of the cell membrane have an **excess of positively charged residues**.

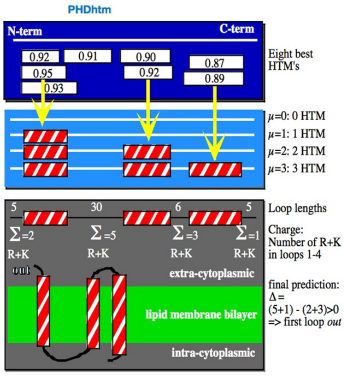
PHDhtm: Membrane Helix Prediction

Predict: **Membrane Helix** or **Not Membrane Helix**

Using the Sequence-to-Structure -> Structure-Structure approach now led to helices twice the length of observed TMH. Based on the number of TMH to expect based on the hydrophobicity a dynamic programming approach can be applied on the *NN energy.*

```
# Note: I didn't really get how this helps in reducing the problem of too long helices??
```





TMHMM: Membrane Helix Prediction

- Based on a Hidden Markov Model

4. When are TMHs correctly predicted?

Today: at maximum ±5 residues overlap

1.10 TMSEG

20.06.2017 | **Slides** | **Lecture Recording**

1. Introduction TMSEG

1.1 Rational: Why another predictor

- More Data available
- Less expensive machine learning (more computing power available)
- Improve runtime

1.2 Dataset

- 166 membrane protein sequences after redundancy reduction
- Data curated and linked from several databases (PDB, OPM, ...)
- 1441 proteins from the SignalP Training Set
    - 1142 soluble (after RR)
    - 199 membrane (after RR)
- **Split Dataset into 4 subsets**
    - each set maintaining distribution of TMPs, SPs and sequence length
    - use 3 sets for **cross-validation**
    - use 1 set for final independent evaluation (**blind set**)

2. TMSEG Prediction

Intro: Classification Trees and Random Forest

Classification trees

- Given N training samples and M inout features find the best recursive partitioning to predict the class labels in the leaf nodes
- Approaches differentiate algorithm: splitting, pruning, balancing, ...

Random Forest

How does it work?

- ensemble method: grow T trees for a forest
- for M input features choose $m < M$
- for each $t \in T$
    - Select N training samples with replacement from all N samples
    - At every split, choose m random features. Use the best split among those to build the tree

- The final prediction uses the prediction of all trees

Advantages

- fast
- robust against overtraining
- no black box
- intuitive to interpret
- good performance

2.1 Step 1 - Feature Sets

Initial Prediction

- Random Forest ($T = 100, m = 9$)
- Sliding Window of 19 residues ($w = 19$)
- 3 scores for each residue (0 - 1000)
  - Signal Peptide (often mistaken for TMHs)
  - TMH
  - Soluble

Feature Set

- **Global Features**
  - Global amino acid composition
  - Protein length
- **Local Features**
  - PSSM Score (Position Specific Scoring Matrix)
  - Distance to N- / C- Terminus
  - Average hydrophobicity (Kyte-Doolittle)
  - percentage of hydrophobic residues (in window size $w = 9$)
  - percentage of negative / positive charged residues (in window size $w = 9$)
  - percentage of polar residues (in window size $w = 9$)

2.2 Step 2 - Empirical Filter

- smooth score with median filter ($x = y$)
- Adjust scores to avoid overprediction
  - soluble += -185
  - TMH += -60
- Assign each residue the state with the highest score
- Remove signal peptides with <4 residues
- Remove TMHs with <7 residues

2.3 Step 3 - Refine TMH prediction

- **Neural Network** (25 hidden nodes)

    - Input: TMH segments of variable segments of variable length
    - Features:
        - Amino acid composition
        - Average hydrophobicity
        - percentage of hydrophobic residues
        - percentage of charged residues
        - segment length
- Split long TMHs (≥35 residues) into 2 shorter TMHs (≥17 residues)
- Adjust TMH endpoints by up to ±3 residues

2.4 Step 4 - Topology Prediction

- Random Forest ($T = 100, m = 7$)
- Assign soluble segments to side 1 or 2
- Features
    - Amino acid compositon
    - percentage of positive charged residues
    - percentage of absolute difference of positive charged residues on side 1 vs side 2
- **Only consider residues close to TMHs**
    - 15 residues nest to TMHs and 8 residues into TMHs
- Predict topology of N-Terminus and extrapolate
- if a SP is predicted, the residues after the SP are always 'outside' (SP = Signal Peptide)

Question: What are advantages of using a Random Forest?

> - Fast
> - robust against overtraining
> - no black box
> - Intuitive to interpret
> - good performance

3. TMSEG Performance measures

```
# Note: Per residue measures are often misleading!
#       => better score TMH segments
```

Whole Protein Scores: Q_{ok} and Q_{top} **=> What is a correctly predicted TMH?**

- **Strict Criteria**
    - Endpoint deviation ≤ 5 residues
    - Overlap at (observed / predicted) at least 50%

How can we measure the performance on predicting Transmembrane Helices?

Recall:

$$r_i = \frac{correctly\ predicted\ TMHs}{observed\ TMHs}$$

Precision:

$$p_i = \frac{correctly\ predicted\ TMHs}{predicted\ TMHs}$$

Q_{ok}:

$$Q_{ok} = \frac{100}{N}\sum_{i=1}^{N} x_i; x_i = \begin{cases} 1, & if\ p_i = r_i = 100\% \\ 0, & else \end{cases}$$

t_i:

$$t_i = \begin{cases} 100\%, & \text{if topology correct} \\ 0, & else \end{cases}$$

Q_{top}:

$$Q_{top} = \frac{100}{N}\sum_{i=1}^{N} y_i; y_i = \begin{cases} 1, & if\ t_i = p_i = r_i = 100\% \\ 0, & else \end{cases}$$

How can we measure the performance on distinguishing soluble proteins from transmembrane proteins?

FPR:

$$FPR = 100 * \frac{incorrectly\ predicted\ TMPs}{soluble\ proteins}$$

Sensitivity:

$$Sensitivity = 100 * \frac{correctly\ predicted\ TMPs}{observed\ TMPs}$$

```
# Result: TMSEG has exceptionally low misclassification rates compared to other methods.
#         Additionally, it is strong on topology predictions.
```

4. Future Work

How to get more data?

Check against data published after the release of the method. The data is then unknown by any method.

4.1 Applying TMSEG to other methods

- High modularity (step 1 - 4) of TMSEG allows it to be applied to other methods
- Apparently it can

4.2 Potential extensions

- Re-entrant regions not modeled (too little data)

1.11 Beta Membrane and Accessibility

22.06.2017 | **Slides** | **Lecture Recording**

Recap

Lipid bilayer (membranes)

- hydrophilic outside,
- hydrophobic inside

Normal surroundings of proteins are **solvent** (hydrophilic, water). Generally, the core of a protein is **hydrophobic**.

Trans Membrane Helices (TMH)

- really small fraction of experimentally known proteins (3D structure)
- but 15% to 25% of all proteins
- 60% of drug targets
- only about 2% of all *unique* structures have membrane helices
- **1D prediction very successful**

Beta Barrels

TMB = Trans Membrane Barrel

- "barrels" formed out of β-sheets connected by hydrogen bonds,which go through the membrane
- looking from the tops they have a hole
- they are pores, letting anything pass that is small enough

Beta Barrel Prediction: PROFtmb

Model Design:

- Hidden Markov Model
- structure based labels (states)
  - inside loop
  - outside loop
  - strand up
  - strand down

How to assess whether this model makes sense?

- Count the different states in the set of proteins, where you know (from experiments where the barrels are)

- Put the observation into the **priors** for the **HMM** (Hidden Markov Model) and train for all the others
- Check the results (per residue) predicted vs observed

Conclusion: Remarkable performance

BUT: Can we distiguish proteins with / without TMB?

Challenges:

- Where do barrel domains start / end?
- Sometimes barrels are built out of several peptide chains (proteins)
- Per Protein Performance: **Accuracy vs Coverage**
  - Where to put the threshold when analysing a new protein?
    - Intuitive / Literature: **Intersection of Accuracy and Coverage**
    - Optimized per Case: E.g. for Master Thesis high accuracy if more important than coverage, as experimental biologists will follow up on only a few of the found proteins in further research

Accessibility

What is it about? Why is this relevant?

- accessibility of residues to water
- outside vs inside

1) Absolute Accessibility: ASA (square Ångstrøm, 1 Å = 0.1 nm)

Long side chains may appear more accessible: *Different amino acids have a different length of their side chain and thus the absolute accessibility per amino acid differs.*

Using absolute accessibility may lead to wrong conclusions.

2) Relative Accessibility: ASA / max ASA

3) "States":

- buried, exposed
- buried, intermediate, exposed

> Note: It doesn't matter whether something is 80% or 100% exposed, but it does matter whether something is 0% or 20% exposed. Also, drawing the line where to set the "best" threshold between the states is discussed in academia.
>
> RostLab Approach: Square Root -> Switch from percentage to predicting 10 states

Solvent Accessibility

Accessibility helps in predicting protein function.

- sub cellular localization
- protein-protein interactions
- flexibility / motion from structure

Historically: Prediction by hydrophobicity

- hydrophobic: inside
- hydrophilic: outside

PHDacc: Machine Learning Approach

- 10 output units
- Advantage: No need to decide on treshold beforehand. Threshold can be chosen for future needs.
- Advantage: Mapped to a 2 state system (burried / exposed) each prediction also carries the confidence in the prediction

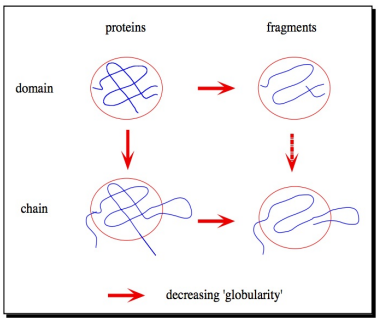
Detailed Prediction Problematic:

ConSurf: Significant gain by evolutionary information (in/out with > 75% accuracy)

More Globular - More likely expressed

> Note: I really don't get this slide / content. Anyone an idea, what is meant by that?

- **Domains** are compact structures on their own (= they fold on their own)
- **Question:** How can we see (by a sequence) what we are related to?(Related to what?)
  - Answer: Predict the residues on the surface. (Why???)
  - 1) Take a 2 state model (burried / exposed)
  - 2) Predict the residues which are exposed
  - 3) Check to which of these (see image) the sequence fits best
  - Assumption: Protein are spheres. (Which is apparently the case in an overwhelming fraction of proteins)



1.12 Protein Disorder

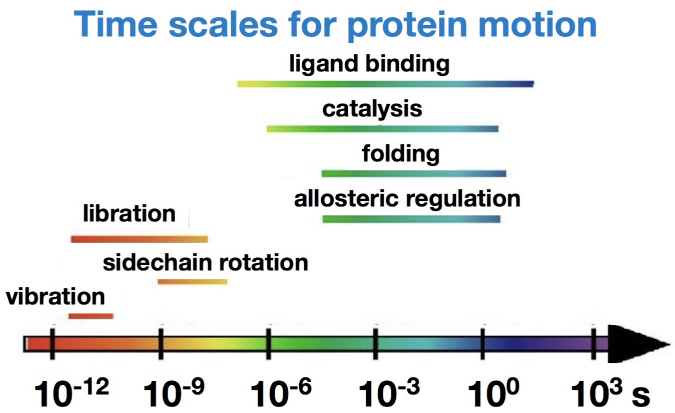
27.06.2017 | **Slides** | **Lecture Recording**

1. Recap

2. Intro

Proteins are dynamic structures and thus they move.

- Ligand Binding
- Catalysis
- Folding
- Allosteric Regulation (?)
- Libration (?)
- Sidechain rotation
- Vibration



Natively unstructured regions: a protein sequence that in solvent is unstructured (2 time points will give different images), but upon binding adopts a secondary structure (induced fit)

Features of Disorder

- Efficient binders (larger interface / outreach)
- Regulated through post-translation modifications
- Increasing complexity by structural plasticity
  - having a "key" for different locks (bindings sites) -> bind to different subtrates
- Active only in **disordered version** (large difference between on and off)
- **Disorder is used for**
  - buffering
  - extending reach (binding)
  - extending reach (signaling)

Coupled binding an folding

- Fly casting: increase surface to 'reach out'
- Initial contacts weak and non-specific
- Folding upon approach of target

Types of natively unstructured regions

- unstructured
- molten globule
- linked folded domains
- mostly folded local disorder

3. Disordered Data

Database: DisProt - a database for disordered proteins
NORS: no regulary structure
Loopy Disorder: less than 5% helices / strand and more than 70 residues

- not found in PDB
- with a (little) relaxation long 'connecting' loops in protein were found
- about **10% of biomass** genome has such (weird) structures (most of them in Eukaryotes)
- on average these NORS regions were **170 residues long**

Problem: Just looking for NORS does not predict shorter (< 70 residues) disordered regions well

4. Methods

4.1 B-Value Prediction

B-Value: Backbone flexibility (experimentally determined)

Can we predict them from sequence?

- B-Values across PDB have to first be normalized (different depending on family)
  - B-Values in principle are conserved
- For ML the flexibility has to be projected on a simpler space (flexible / not flexible)
- Where to put the threshold?
  - 2 thresholds on the sides: Throw away 90% of your data. Not a good idea.
-
  - *Offtopic*: Putting the threshold around the peak, where the experimental error is the highest will consequently influence the model. (never pick a peak!)

PROFbval: Predict residue flexibility

- Classical PROF, with 2 output nodes: FLEXIBLE, RIGID
- **captures aspects of protein dynamics, NOT disorder directly**
- How are B-Values related to disorder?

    - No clear proportional correlation =(

4.2 NORS Region Prediction

= distinguish unstructured from well structured loops

How can we detect shorter NORS regions, without lowering the threshold of 70 residues?

Idea: Machine Learning

- Positive dataset = all NORS predictions (<70 residues) in the entire proteomes
- Negative dataset = the whole PDB
- Problem: Dataset is flawed
    - many considered 'false' are actually 'true'
    - many considered 'true' are actually 'false'

How can a dataset with many mistakes be machine learned?

- If the error is random (white noise ≠ systematic error), a **consistent signal** is still strong enough to be picked up if the data set is big enough

4.3 Contact Deprived Region Prediction

Look at every pair of residues and look whether they are in contact -> predict contacts.

***Note:** 3D structure predictions from 1D structure is (today) still not solved. However, contact map prediction allows to distinguish between regions that are very constraint (e.g. binding sites) and those that are not.*

Until now we did not assume to know 'what' disorder is. How do we then handle disorder 'experimentally'?

Dunker Hypothesis: Residues NOT visible in 3D structure share disorder. (You don't see it, because it moves to much)

4.4 Meta-Disorder (MD)

Use a 'Meta-Predictor' which combines many methods.

- different methods focus on different aspects of protein disorder
- combining predictors substantially improves prediction

5. Findings and Applications

Main Findings

- Specific contacts are important for disorder prediction
- Hub proteins (?) are abundant with unstructured loops
- different methods focus on different aspects of protein disorder
- combining predictors substantially improves prediction

Eukaryotes dominate disorder (x4 - x10): One reason could be that disorder is one stepping stone of complexity.

- Disorder allows proteins to have more (different) interaction partners and this intrinsically increases complexity.
- Bacteria in extreme conditions (heat, salt, ...) are much more similar to other bacteria in the same habitat regarding disorder than to their evolutionary closest homologs.

2. Exercises

- 2. Exercises
  - 2.1 Introduction
  - 2.2 Biological Background
  - 2.3 Protein Structures
  - 2.4 Alignments
  - 2.5 Resources for BioInformatics
  - 2.6 Machine Learning
  - 2.7 Homology Modeling
  - 2.8 Wrap Up

2.1 Introduction

11.05.2017 | Slides | Wiki

```
# Note: No actual content was thaught in the first exercise lesson.
```

2.2 Biological Background

11.05.2017 | **Slides** | **Wiki**

1. Keywords

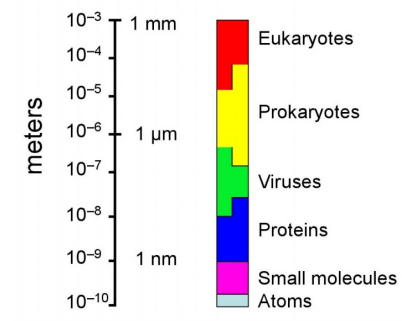
1.1 Unit and Dimension

Molecular Mass

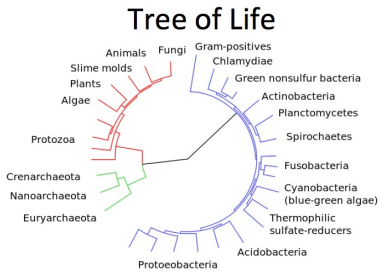
- **Unit:** 1 **u** or **Da** is defined as $\frac{1}{12}$ of the mass of ^{12}C isotope
- **Mole:** Amount of Substance, which contains as many elementary entites as atoms in $0,012kg\ of\ ^{12}C$ isotope
- **Avogadro Constant:** N_A or L, entities per mole and is defined as $6,022 * 10^{23}\ mol^{-1}$

Spatial dimensions:

- **Ångström:** 1 Å is equivalent to $0,1\ nm$ or $100\ pm$
- Radius of atoms: $0,3 - 3$ Å
- Distance of different chemical bonds:
  - **C-C:** $154\ pm$
  - **C=C:** $134\ pm$
  - **C-H:** $109\ pm$
  - **C-O:** $143\ pm$
  - **C=O:** $120\ pm$
  - **C-N:** $147\ pm$
  - **N-H:** $101\ pm$
- Average weight of amino acid: $100 - 110\ Da$
- Typical length for soluble proteins: $3 - 6\ nm$
  - ~ 300 amino acids for prokaryotic proteins
  - ~ 400 amino acids for eukaryotic proteins



"Relative scale" by TimVickers - Wikimedia



1.2 Morphology - Cellular Structure

Terms:

1. Cell **Membrane**: A selective barrier separating the inside of a cell / organelle from the outside, consisting of a **phospholipid bilayer**.
2. **Cytoplasm:** semi-liquid medium inside the cytoplasm membrane
3. **Cytoplasmatic Membrane:** the membrane around the cell
4. **Cell Compartment:** a region within the cell mostly enclosed by a membrane
5. **Cell Organelle:** a special type of compartment that has a certain function within the cell
6. **Nucleus:** a membrane enclosed volume of the cell, which contains most of the genomic material
7. **Cell Wall:** Some cells have an extra cellular rigid layer

Cell Organelles: Ribosome, Mitochondrium, Chloroplasts

Prokaryotic vs Eukaryotic Cells:

Eukaryotic Cell	Prokaryotic Cell
membrane enclosed compartments	No membrane enclosed compartments
genomic material in nucleus	Genomic Material located in nucleoid region
organelles: mitochondria, chloroplasts	No organelles
compartments: endoplasmatic reticulum, Golgi-apparatus	

1.3 Biomolecule classes, building blocks and physiological function

Hydrophilic: likes to interact with water because of partial electrical charges (mostly polar)
Hydrophobic: avoids interaction with water, because of distinct charging points (mostly non-polar)

Lipophilic: likes to interact with fatty / unpolar molecules (-> hydrophobic)
Lipophobic: avoids interaction with fatty / unpolar molecules (-> hydrophilic)

Biomolecules

- typically polymers (= constructed from many identical or similar residues)
- can form complexes
- can contain both hydrophilic and lipophilic parts

Carbohydrates: (sugars)

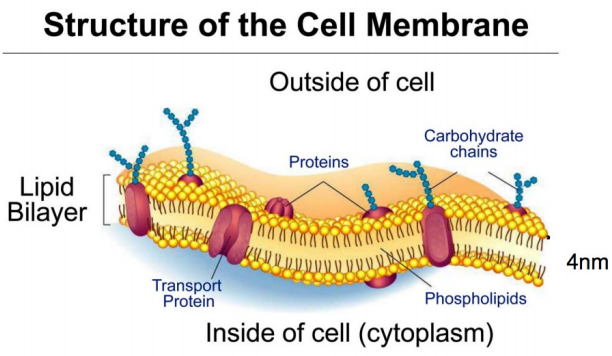
- molecule consisting of **C, H, O** atom
- $C_m(H_2O)_n$ - usually hydrogen-oxygen ratio of 2:1
- hydrophilic
- purpose: fuel, energy storage, structural components, ...

Lipids:

- loosely defined as **being soluble in non-polar solvents**
- fats, waxes, vitamins, ...
- purpose: energy storage, signaling, structural components

Glycerolipids: (Phospholipids)

- Glycerol with 3 docking slots (hydrophobic)
- 3 fatty acids: triglycerids (energy storage)
- 2 fatty acids + 1 phosphate group: **Phospholipid**, major building block of membranes



Nucleic Acids:

- Store and transmit genetic information
- A polymer built out of **Nucelotides** (base+sugar+phosphate)
  - hydrophobic nucleobases on the inside
  - hydrophilic backbone (phosphate + sugar)
- **RNA:** single stranded, but can adopt complex secondary structure with itself or other RNA
- **DNA:** typically double stranded (**double helix**) formed by a reverse-compliment strand
  - Has a **5' end** and a **3' end**.
  - always annotated from **5' to 3'**
- Bonds: non-covalent **hydrogen bonds** (rather weak)

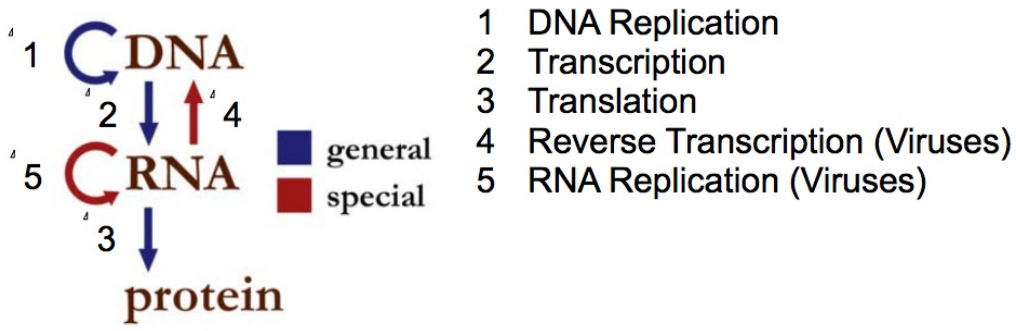
Proteins:

- polymerized amino acids
- *work horse* or *machinery of life*
- functions
  - catalysis of chemical reactions (enzymes)
  - structural elements (collagen fibers)
  - sensing
  - immune system
  - etc ...

1.4 Genetics

Gene: Any discrete locus of heritable, genomic sequence which affect an organism's traits by being expressed as a functional product or by regulation of gene expression
= every subsequence of DNA that encodes a functional protein

Central Dogma of Molecular Biology: (!!!) Describes the flow of information from **DNA** to **RNA** to **Protein**



1) Replication: Duplication of DNA

- highly controlled, carried out by multi-protein complex
- for each strand in the double helix, a complementary strand is synthesized
- in **5' -> 3'** direction

2) Transcription: Creation of working copy of genes

- synthesis of a single stranded RNA from a template sequence in the DNA (after a promoter region)
- carried out by a multi protein complex
- the resulting **mRNA** undergoes several maturing steps before Translation
  - mRNA: translated into protein
  - xRNA: function
  - rRNA: component of the Ribosome
  - tRNA: amino acid carrier for translation

3) Translation: Conversion of **mRNA** into protein

- carried out by **ribosomen**
- starts with **AUG** codon
- by default: Sequence of RNA and Protein are noted in the same direction

Inheritance / Mutation: Copy errors lead to rise of evolution

- Mutation Types
  - (longe range) rearrangements
  - point mutations
- **Rearrangements**
  - rearrangements of DNA segments due to error in recombination process
  - induced by
    - DNA damage (chemical, radiation)
    - virus infections

    - consequences: loss of gene functions or loss of controls
- **Point Mutations**
    - change of a single DNA residue
    - **Frame-Shift** (loss / gain of a single residue), effect rather unpredictable
    - **Substitution**
        - Transition: A <--> G or C <--> T
        - Transversion: A,G <--> C, T
        - Consequences:
            - silent
            - missense (= amino acid change)
            - nonsense (= stop codon introduced)
            - affects splicing site
    - Induced by: replication errors

Gene Regulation: Takes place on several layers

- transcription rate
- stability of mRNA
- translation

1.5 Metabolism / Physiology

Anabolism: Reactions that aim to the synthesis of a new substance
Catabolism: Reactions that aim to the degradation of a substance (energy generation, removal of damaged structures)

steady state / equilibrium: degradation and synthesis rate are balance (System is dynamically stable)

1.6 Proteins

More about proteins in future lectures.

2. Exercises

∅

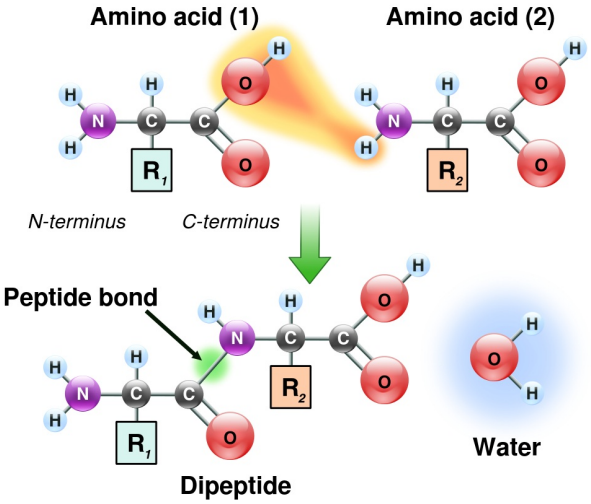
2.3 Protein Structure

18.05.2017 | **Slides** | **Wiki**

1. Keywords

Amino acids, side chains, residues

- 20 different amino acids
- **Backbone:** all of them have the same backbone (see following picture)
- **Side Chain:** the differ in their side-chains, which also give them unique features.
  - electrically charged side chains (positive or negative)
  - polar uncharged side chains
  - hydrophobic side chains
  - special cases
- Amino Acids are the building blocks of proteins. Once it is part of a protein, the amino acids are referred to as a residues



Protein sequence

- linear sequence of amino acids (residues), connected by their backbone
- formed by a condensation reaction (formation of a peptide bond, under creation of a water molecule)
- oriented from **N-Terminus** to **C-Terminus**
- typically starts with **Methionine (AUG codon)**

Secondary, tertiary, quaternary protein structure

- **Secondary Structure**
  - local structure elements: α**-helix, β-sheet, loops**

    - building blocks for higher order structures
    - stabilized by **hydrogen bonds**
    - amino acids have 'preferences' for certain secondary structure elements
- **Tertiary Structure**
    - spatial arrangement of secondary structure elements of a protein
    - alternative arrangements can exist
    - can be used to **hierachicaly organize** found proteins
- **Quarternary Structure**
    - formation of **multi-protein complexes**
    - difficult to determine precisely

Coding / Representation	Protein Aspect
1D Information: sequence of amino acids as a string	**Primary Structure**: amino acid sequence
2D Information: 2D-Array, contact map	**Secondary Structure:** helices, sheets, ...
3D Information: coordinates or atom couplings	**Tertiary Structure:** Spatial arrangement of secondary structure elements

Hydrogen bonds

- **weak** bond between H-atom (donor) and **O** or **N** atom
- Distance: $160 - 200\ pm$ (relatively large distance)
- stabilize secondary structure formations

Alpha helix, beta sheet, loop, random coil, disordered region

- **Alpha Helix**
    - **3,6 amino acids** per turn, spiral forming
    - typically **4-40** residues long
    - stabilized by hydrogen bonds between backbone atoms
- **Beta Strand / Sheet**
    - Several **parallel** or **anti-parallel** strands form a sheet
    - 'long-range' hydrogen bonds (in terms of residues involved)
    - 'flat'
- **Loop / Turn / Coil**
    - connector between undefined secondary structure elements
    - end / start of polypeptide chain
- **Disordered Region / Random Coil**
    - no clear secondary structure elements identifiable
    - like 'statistical distribution' of shapes
    - biologically: 'adapter' to different target shapes, which stabilize upon contact with partner

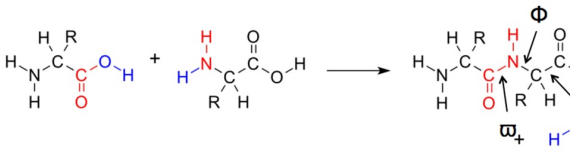
Protein features

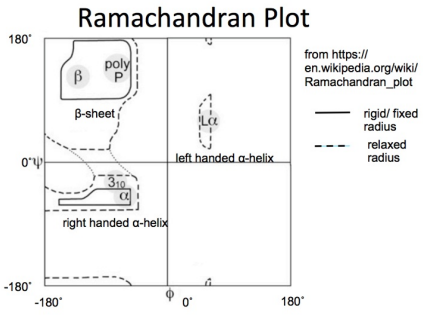
- Size
- Amino Acid Composition
- Surface Area
- Hydrophobicity
- Solvent accessibility
- Binding Sites / Active Binding Sites
- Iso-Electric Point

Ramachandran plot

When chaining up the amino acids on the polypeptide bond, the main bond $\omega_+$is rather rigid and cannot rotate. The two bonds around the center atom (ϕ, ψ) of each amino acid, however are free to rotate (-180, $+180$). However, these rotation angles are subject to certain constraints (due to e.g. the side chain of the amino acid)

A **Ramachandran Plot** plots the typical / allowed regions for α**-helices** and **β-sheets**.





Classification of Structures

Two similar methods (CATH / SCOP) both aiming to **organize** the protein structures available in the PDB based on **single domains.**

- Hierarchical System
  - Secondary Structure Content
  - Fold
  - Super Families
  - Families

SCOP: Structural Classification Of Proteins

- fully manually curated, driven by expert analysis

CATH: (Class, Architecture, Topology, Homologous Superfamily)

- semi-automatic procedure for deriving a novel hierarchical classification of protein domain structures
- 4 main levels

    - **Protein Class:** Mainly secondary structure composition of each domain
    - **Architecture:** Summarizes shapes based on orientation of secondary structure elements
    - **Topology:** Sequential connection considered
    - **Homologous Superfamily:** High similarity with similar functions, evolutionary relationship assumed
- **Numbering Scheme**
    - C: (Alpha, Beta, Alpha/Beta) +1
    - A: Same Architecture, different Topology (31) + 10
    - T: Topology (connection of 2ndary structure elements) (505) +10
    - H: Homology (families)(645) +10

Protein domain, PFAM

- **PFAM - Protein Families Database**
- focus on **single domains**
- 559 clans, 16295 families
- **PFAM-A**
    - manually curated
    - HMM profiles for seed alignment
    - HMM profiles for full alignment
- **PFAM-B:** Automatically created

Family: Collection of related protein regions
Domain: Structural Unit
Repeat: short unit, unstable in isolation, but forms stable structure when found in multiple copies
Motif: short unit found outside globular domains
Clans: related group of PFAM family entries

Protein Data Bank (PDB)

- collection of high resolution protein structures (120 000 entries)
- **X-Ray Crystallography, NMR, Cryo-EM**
- slowly growing
- different quality of data

Root mean square deviation (RMSD)

Measure to determine **3D similarity** between structures

1. Superimpose
2. Align sequences to 'guess' corresponding residues
3. Calculate the distances (mostly C_α)

Protein Function

- **Gene Ontology (GO)**
    - controlled vocabulary

    - hierarchically structured
    - 3 Main Sections
        - cellular component
        - molecular function
        - biological process
- **Enzyme Commission Number**
    - Numerical Classification of enzymes
    - Based on chemical reactions the catalyse
    - recommends enzyme name
    - does not imply any evolutionary relation

2. Exercises

2.1 Questions

Question: What are the building blocks of proteins?

> - Proteins are polypeptides, which consist of amino acids as building blocks

Question: Define protein backbone and amino acid side chain in 1 or 2 sentences for each term.

> - **Backbone:** All amino acids have the same with an N-Terminus on one and a C-Terminus on the other side by which they are chained up into a protein. (Always read / write from N- to C-Terminus)
> - **Side Chain:** Each Amino Acid has a different side chain, which equips it with unique features (length, polar, non-polar, electrically charged)

Question: How many amino acids appear in proteins? How can they be classified?

> All proteins are built from **20 different amino acids**. The can be classified into the following 5 groups /categories
>
> - positively electrically charged side chain
> - negatively electrically charged side chain
> - polar uncharged side chains
> - hydrophobic side chains
> - special cases

Question: Name atom types involved in a hydrogen bond. Do S-H groups form hydrogen bonds? Why (not)?

> Hydrogen bonds normally involve
>
> - an **H** atom (obviously) as donor
> - either an **O** or **N** atom as acceptor
>
> Due to its lower electronegativity Sulfur does not engage in classical Hydrogen-Bonds. (However nonclassical hydrogen bonds with Sulfur seem to exist, according to literature)

Question: How are alpha helices held together?

> Alpha-Helices are stabilized by **hydrogen bonds** along the backbone. Every turn takes about **3,6** residues.

Question: What is similar and what is different in the hydrogen bonding of the alpha helix and the beta sheet?

> - Hydrogen bonds in **Alpha Helices** are 'ultra-local' and occur along the backbone of the helix (each 3.6 residues)
> - Hydrogen bonds between **Beta Sheets** happen between the beta-strands (running parallel or antiparallel) which can be rather far apart in sequence.

Question: Why do we find "forbidden" areas in a Ramachandran plot?

> The Ramachandran plot shows the angles (ϕ, ψ) of amino acids in which alpha-helices and beta sheet are typically observed. Forbidden areas are those that are not possible due to physical constraints due to e.g. the side chain of the amino acid.

Question: What is a protein domain?

> A domain is a conserved (sub-)sequences of a protein, which adopts a unqiue 3D structure when put into solvent.

Question: How many amino acids are typically found a in a domain? Why is there a minimum/maximum size?

> Proteins domains are found between 36 and 690 residues long. Most protein domains have a length of around 100 residues.
>
> Minimum Size: ? Maximum Size: ?

2.2 PDB

Question: How many structures are stored in the PDB? How many of those are protein structures?

> About 130 000 entries can be found in the PDB, of which around 120 000 are protein structures.

Question: Which experimental methods are (mainly) used to determine the structures? How long does it on average take to find the 3D structure for one protein for each method?

> The largest part (90%) of proteins structures are determined by X-Ray crystallography, followed by NMR (9%) and Cryo-EM (1%). However, the number of new Cryo-EM structures is projected to overtake the number of new NMR entries in 2017. Current sciene pushing the limits of Cryo-EM resolution, makes it a promising technology for future 3D structure determination.
>
> **Costs / Time per Method**
>
> - X-Ray: (100 000, ?)

> - NMR: (?, ?)
> - Cryo-Em: (?, ?)

Question: How many human ("Homo sapiens") protein structures are in the PDB?

> About 37 000 protein structures of homo sapiens. (Not, sequence unique entries though)

Question: Why does the number of protein structures already decreases when reducing at 100% sequence identity? Why does it decrease when reducing at even lower sequence identity further?

> - The PDB has redundant 3D structures for certain proteins from different experiments. This has different reasons such as competing groups, different research goals, better resolution.
> - Since most proteins developed under evolutionary pressure, large parts of proteins of the same family share a high sequence overlap.

2.3 Molecular Visualization

```
# Note: Neither mentioned in exercises nor lectures. Let's hope there is no question rega
rding this.
```

2.4 Alignments

01.06.2017 | **Slides** | **Wiki**

1. Keywords

Pairwise alignments: global, local (Needlman-Wunsch / Smith-Waterman)

Alignment is needed to

- compare sequences
- find the best possible alignment to callculate distance measure

Goal: Find the arrangement of residues that minimizes / maximizes the scoring function
Biology: Try to maximize the overlap between the sequences

Parameters:

- Substitution matrix (for each pair)
- Gap Penalties (linear, affine)
- global, free-shift (global, but gaps at the start and end of the alignment are not counted) , local

Multiple equally scoring alignments are possible.

Global Alignment: Needleman-Wunsch

- Full Length Alignment (comprises both sequences)
- Makes sense for sequences of nearly equal length

$$Naive\ Recursive\ Formular:\ S_{m,n} = max \begin{cases} S_{m-1,n-1} + d(m,n) \\ S_{m-1,n} + Gap \\ S_{m,n-1} + Gap \end{cases}$$

Local Alignment: Smith-Waterman

- Find the best matching subsequence(s) between two sequences
- Length differences do not matter
- Sequences may be quite dissimilar

$$Naive\ Recursive\ Formular:\ S_{m,n} = max \begin{cases} S_{m-1,n-1} + d(m,n) \\ S_{m-1,n} + Gap \\ S_{m,n-1} + Gap \\ 0 \end{cases}$$

Dynamic programming as solution for cascading recursion, Backtracking

Dynamic Programming: (for examples, look at bottom)

1. **Setup matrix:**
    i. Template t horizontal, Query q vertical
    ii. Fill the 1^{st} row / column
        i. Needleman-Wunsch: with increasing gap penalties
        ii. Needleman-Wunsch (with free-shift): with zeros
        iii. Smith-Waterman: with zeros
2. **For each row** m_i (top -> down)
    i. **For each position** $m_{i,j}$ (left -> right)
        i. $s_1 = m_{i-1,j-1} + SubstitutionMatrix(q_i, t_j)$
        ii. $g_1 = m_{i-1,j} - GapPenalty$
        iii. $g_2 = m_{i,j-1} - GapPenalty$
        iv. **Enter** $max(s1, g1, g2)$ **into the box and remember which score was selected**
3. **Backtrace**
    i. Find the maximum value
        i. Needleman-Wunsch: in the last row
        ii. Smith-Waterman: in the matrix
    ii. Follow the path back to the origin and note down the aligment
        i. **Diagonal Step:** Match the corresponding residues
        ii. **Step up**: Only take the **Query-Value** (vertical), add a gap to the template sequence
        iii. **Step Left:** Only take the **Template-Value** (horizontal), add a gap to the Query sequence

Substitution matrix

What is better: A short sequence matched perfectly or a long, but only partly matched sequence? We can't tell yet.

Substitution Matrices are **observation derived weight matrices**

- the score reflects the likelihood of an exchange
- depending on the underlying dataset
- thus, choose your matrix according to the proteins of interest
- Several matrices available: **BLOSUM62, PAM250, PAM100, PAM50**

Sequence identity, similarity, conservation

[Source: http://homepages.ulb.ac.be/~dgonze/TEACHING/stat_scores.pdf]

Score: A number used to asses the biological relevance of a finding. (Here it describes the quality of the alignment)

$$S = \sum_{i=1}^{L} S_{r1,i,r2,i} \; with \; respective \; gap \; penalties, \; if \; r_{1,i} \; or \; r_{2,i} \; is \; a \; gap$$

Bit-Score: A log-scaled version of the score

- In context of sequence alignments (BLAST), the **bit-score s'** is a normalized score, which lets you **estimate the magnitude of the search space** to find a score greater or equal to the one you got by chance.

$$S' = \frac{\lambda S - ln(K)}{ln(2)} \; \lambda,\; K \; depend \; on \; the \; substitution \; matrix \; and \; gap \; penalties$$

- If $S' = 30$ this would imply that $2^{30} = 1 \; billion$ random independent pairwise alignments are needed on average to find a similar score by chance
- Size of the search space can be calculate by $K * sequenceLength * DBEntries$

P-Value: Probability that an event occurs by chance

- The **p-value** associated to a score **S** is the probability to obtain by chance a score **x** at greater or equal to **S.**

$$PVal(S) = P(x \geq S) : PVal_S^{MSP} = Ke^{-\lambda S} = Ke^{-ln(2)S' + ln(K)} = 2^{-S}$$

E-Value: (Expectation Value) Correction of the p-value for multiple testing

- he **e-value** associated to a score **S** is the number of distinct alignments, with a score greater or equal **S**, which are expected to occur in a database search by chance.

$$E = m * n * Pval = Kmn * e^{-\lambda S} = \frac{m * n}{2^{S'}} \; with \; n = length \; of \; query \; sequence, \; m \; = length \; of \; database$$

PSI: Percent Sequence Identity (Percentage of perfectly aligned residues)

Similarity: Percent of residues matched with a positive score in the Substitution Matrix

FASTA format (basic principle)

???

One letter code for amino acids (you do not have to learn the code, just know what it means)

Yes. Find a table with abbreviations here https://en.wikipedia.org/wiki/Proteinogenic_amino_acid

Homology, homologues/homologs

- **Homology**: Assumption of a common ancestor (in this context)
  - Sequence-Similarity is used as evidence for homology
  - not direct observable, because of missing fossils
- **Homolog:** A gene inherited by two species from a common ancestor

Multiple Sequence Alignments (MSAs)

- **Why?** Search strategies are needed to avoid a combinatorial explosion when running one against all alignments
- many different algorithms available

- Search Tree
- Combine Pairwise Alignments
- Profiles

Sequence profile, iterative profile creation

- **Idea:** Build a PSSM (Position Specific Matrix)
  - = A probability vector for every amino acid at every sequence position
- several sequences needed to build a profile
- therefore, first search the database for similar sequences
- then compile these sequences into profil
- search database with the profile to retrieve more candidates
- rebuild profile

BLAST

- indexing of database for typically **3-aa-word-seeds**
- list of high-scoring words is used to find similar candidates
- word extension to **HSP (High Scoring Pairs)**
- Use Local Alignment of the full sequences
- MANY optimization heuristics

PSI-BLAST

- Run BLAST
- Use retrieved proteins to build a profile (PSSM)
- Use profile (PSSM) to query database for more candidates
- Rebuild Profile
- Repeat search for candidates with new profile until convergence or another stop criterion is reached

2. Exercises

2.1 Questions

Question: How can you define similarity between two protein sequences?

> Two ways to define similarity are to compute the percentage of residues that were aligned
>
> - and matched on the same amino acid (PSI, Percent Sequence Identity)
> - and matched with a positive score in the Substitution Matrix

Question: What does "conservation" mean in the context of sequence alignments?

> - Conserved sequences are similar or identical (sub)sequences that occur within protein sequences.
> - By compiling homolgues proteins (a family) into a profile, it become clear, which subsequences are more conserved and thus more important for the function of the protein

> family

Question: Why are sequence alignments useful?

> - They are useful because the allow use to find the 'best' possible match between two sequences. Only this allows use to
>   - compare their 3D structure
>   - create predictions about their similarity in 3D structure
>   - make assumptions about their evolutionary relatedness

Question: What are the main differences in the algorithms of Global and Local alignment? Why does it make sense to not always perform a global alignment.

> - Global alignment methods always align 2 sequences from beginning to end.
> - Local alignment methods only align subsequences.
> - For Global alignment to be meaningful, the sequences should have similar length. Since proteins are between 35 and 30000 residues long, it does not make sense to always use global alignment.
> - Also when e.g. looking for proteins with a certain domain it does not make sense to use global alignment, since we are explicitly looking for subsequences.

Question: Which amino acids can (with high likelihood) be substituted for Leucin without having an effect on protein function?

> Methionine (2). Isoleucin (2), Valine (2), Phenylalanine (0), because the have a positive score in the BLOSUM62 matrix

Question: Which substitution is more probable according to PAM250 and according to BLOSUM62: [a] W <> F [b] H <> R

> - W (Tryptophan) <--> F (Phenylalanine)
>   - BLOSUM: 1
>   - PAM250: 0
> - H (Histidine) <--> R (Arginine)
>   - BLOSUM: 0
>   - PAM250: 2
>
> => For BLOSUM W <--> F is more likely to be observed
> => For PAM250 H <--> R is more likely to be observed

Question: What is a multiple sequence alignment?

> A method to align multiple sequences against each others.
>
> - building a consensus sequence and aligning new sequences against it
> - building a search tree
> - building a profile (like PSI-BLAST)

Question: What kind of sequences are likely to be used for an MSA? In which relationship are they to each other?

> - Most likely, sequences which we assume a evolutionary relationship (homology) will be used. Following the homology assumption, we expect sequences with a high PSI to have a similar structure because they have a common ancestor.
> - Building up a profile with them, would then allow us to discover conserved regions and use the profile to find more candidates in the Twilight Zone with Profile-Sequence comparison.

Question: Why would you want to align multiple sequences? What kind of information is contained in MSAs but not directly in e.g. all-against-all pairwise alignments?

> Building up a profile with similar sequences, would allow us to discover conserved regions which developed under evolutionary pressure.
> By compiling such a family of proteins (we assume that they have a common ancestor -> homology) into a profile, we can find more candidates in the Twilight Zone with Profile-Sequence comparison.

Question: Given your knowledge of the algorithms for pairwise alignments, how could you calculate an MSA? Is that a feasible approach? Why?

> One approach would be to sequentially align sequences against a consensus sequence. This approach, however, comes with problems such as the need for a strategy how to find the consensus for a certain amino acid at a certain position, the fact that the order of alignment might matter, etc.
>
> A better approach is the creation of a PSSM (position specific scoring matrix), which contains for each position of the profile the likelihood that a certain amino acid occurs there.

Question: You have a sequence which you would like to find in a database. Which search method and which E-value cutoff do you use, [a] if you know your sequence is in the database and only want to find that entry [b] if you would like to find homologs.

> [a] Pairwise alignment (??? is this right???)
> [b] BLAST

Question: What is the difference between BLAST and PSI-BLAST?

> BLAST uses the BLOSUM matrix to retrieve homologs. It runs only once and returns the found sequences.
>
> PSI-BLAST uses BLAST in the first run to find homologs and build a profile. By using, and iteratively rebuilding, the profile it can find more distant (in terms of sequence identity) homologs with Profile-Sequence alignment.

2.2 Needleman-Wunsch

Find the best alignment between the sequences “WHAT” and “WHY”, using the Needleman-Wunsch algorithm, with +1 for a match, -1 for a mismatch, and -2 for a gap.

Note: *In each row all value are written in brackets in the following format (fromDiagonal, FromLeft, FromTop). The chosen value is entered in the box*

		W	**H**	**A**	**T**
	0	-2	-4	-6	-8
W	-2	1 (**1**, -4, -4)	-1 (-3, **-1**, -6)	-3 (-5, **-3**, -8)	-5 (-7, **-5**, -10)
H	-4	-1 (-3, -6, **-1**)	2 (**2**, -3, -3)	0 (-2, **0**, -5)	-2 (-4, **-2**, -7)
Y	-6	-3, (-5, -8, **-3**)	0 (-2, -5, **0**)	1 (1, -2, -2)	-1 (**-1**, **-1**, -5)

Note: *Two alignments with the same score are possible here.*

Backtrace 1

		W	**H**	**A**	**T**
	0	-2	-4	-6	-8
W	-2	**1**			
H	-4		**2**	**0**	
Y	-6				**-1**

Alignment:

```
WHAT
||
WH.Y
```

Backtrace 2

		W	**H**	**A**	**T**
	0	-2	-4	-6	-8
W	-2	**1**			
H	-4		**2**		
Y	-6			**1**	**-1**

Alignment:

```
WHAT
||
```

```
WHY.
```

2.3 Smith-Waterman

Find the best alignment between the sequences “WHAT” and “WHY”, using the Smith-Waterman algorithm, with +1 for a match, -1 for a mismatch, and -2 for a gap.

Note: *In each row all value are written in brackets in the following format (fromDiagonal, FromLeft, FromTop). The chosen value is entered in the box*

		W	**H**	**A**	**T**
	0	0	0	0	0
W	0	**1** (**1**, -2, -2)	**-1** (**-1**, **-1**, -2)	-1 (**-1**, -3, -2)	**-1** (**-1**, -5, -2)
H	0	-1 (**-1**, -2, **-1**)	2 (**2**, -3, -3)	0 (-2, **0**, -3)	-2 (**-2**, **-2**, -3)
Y	0	-1 (**-1**, -2, -3)	0 (-2, -3, **0**)	1 (**1**, -2, -2)	-1 (**-1, -1**, -4)

Note: *The same alignment as before are possible to align the full sequence. However, since we only want to align a local sequence, we can just start with the highest score we find in the matrix and backtrace only this* ***subsequence****.*

Possible Subsequences Alignment: Since, we did not have a substitution matrix, which compiles the likelihood of randomly aligning sequences into the the alignment algorithm, we cannot say for sure what is better: Aligning 2/4 residues or aligning a subsequence of 2/2.

```
WH
||
WH
```

2.5 Resources for Bioinformatics

08.06.2017 | **Slides** | **Wiki**

1. Keywords

Resources / Databases

- Represent a shared fund of knowledge
- "*data collections*" often more precise than database
- have evolved over decades (legacy issues)
- **primary** and **secondary** databases
- **Primary Databases**
  - Genbank / EBI / DDBJ (nucleic acid sequences)
  - UniprotKB (protein sequences)
  - PDB (3D structures)
- **Secondary Databases** (contains refined of topic selected / processed informa+on derived from primary databases)
  - STRING
  - PROSITE
  - PFAM

Genbank, EMBL,EBI

The entries in the Genbank is growing exponantially!

Genbank: sequence database maintained by the NCBI (National Center for Biotechnology Information, USA)
EMBL: The European Molecular Biology Laboratory maintained by EBI (European Bioninformatics Institute)

Flat File Format: Entries come compressed as **text files** with an uncompressed size of over **700 GB**

- Records consist of 2 parts
  - Columns (1-10), Entry Field name
  - Remaining line with the content (sequence)
- Version / Accession Format
  - Unique identifier: Value of the VERSION field
  - VERSION: Accession number + "." + integer
  - Update of version ONLY when sequence changes
  - Other field values could be changed without notice

Uniprot

Universal Protein Database

- combines information from Swissprot, TrEMBL, PIR

Swissprot/Trembl

Manually annotated and revied section of UniProt.

- **Annotation Process**
-
    1. Sequence Curation
    2. Sequence Analysis
    3. Literature curation
    4. Family based curation
    5. Evidence attribution (every annotation is attributed to its original source)
    6. Quality assurance, integration and update

ExPasy

Expert Protein Analysis System

- **Categories**
    - Proteomics
    - Genomics
    - Structural Bioinformatics
    - System Biology
    - Phylogeny / Evolution
    - Populations genetics
    - Transcriptomics
    - Biophysics
    - Imaging
    - Drug Design

PDB

- **PDB file format**
    - **Header**
        - Protein information
        - citation
        - details of structure resolution
    - **Coordinates and Connectivity**
- **mmCIF**
    - Standard format of PDB since 2014
    - originated from the crystallopgraphic community
    - **Advantages**
        - Extensible
        - Flexible ordering

        - Few syntax rules
        - Facilitates automatic validation
    - **Concepts**
        - Entity: (polymer, non-polymer, water)
        - Chemical Component (blocks that build entities -> non-standard residue)
        - Structural Component (structural features, e.g. helix)
        - Asymmetric Unit Component (chain, two components can refer to same entity)
        - Biological Component (sub- and super-components of structure)

Accessing these databases: BioPython

Use existing parsers/ libraries wherever you can! One of them is e.g. **BioPython**
https://github.com/biopython/biopython.github.io/

PFAM

Protein Families Database: (Especially useful **PFAM-A** with Profile HMMs, seed alignments, full alignments)
(*Secondary Database)*

STRING

Proteins Interaction Network: Known and predicted protein interactions
(*Secondary Database)*

PROSITE

Documentation entries describing proteins domains, families, functional sites, associated patterns and profiles
(*Secondary Database)*

- Domain specific descriptions
- provide reference sequence and alignments
- provide consensus pattern

2. Exercises

2.1 Questions

Question: How many structures in PDB have a resolution with =< 2 Angstrom

> ??? High-resolution structures, with resolution values of 1 Å or so, are highly ordered and it is easy to see every atom in the electron density map. Lower resolution structures, with resolution of 3 Å or higher, show only the basic contours of the protein chain, and the atomic structure must be inferred. Most crystallographic-defined structures of proteins fall in between these two extremes. As a general rule of thumb, we have more confidence in the location of

> atoms in structures with resolution values that are small, called "high-resolution structures". from https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/resolution# But no word on percentage.

Question: Which term from computer science you would use to describe PROSITE patterns (e.g. PDOC00022).

>

> -> I would argue it reminds me of regular expressions, but this should be verified. -> I agree, but the syntax is differing from bash regex
> eg see here: http://www.hpa-bioinfotools.org.uk/ps_scan/PS_SCAN_PATTERN_SYNTAX.html compare to: http://tldp.org/LDP/abs/html/x17129.html

Question: Which type of information does STRING provide?

> STRING is a secondary database which provides information about known and predicted protein interactions.

Question: What does PFAM-A contain?

> PFAM-A contains manually curated information about proteins families. It is especially useful, because it contains
>
> - Profile HMMs
> - Seed alignments
> - Full alignments with all hits

2.6 Machine Learning

22.06.2017 | **Slides** | **Wiki**

Machine Learning

1.1 **Ideas and Vocabulary**

- a machine learning device can generalize from real world observations into a “formal” model
- the model should reflect a concept or commonalities and not individual characteristics
- every learning scheme discards some aspects of reality to construct a model (inductive bias), this might also already happens on the level of feature extraction
- feature is a variable describing a specific aspect of real world observations
- training: phase of analyzing real world observations in a formalized representation to derive parameters and/or internal structure
- test phase: phase of model application to determine the reliability of statements (predictions) on instances not used for training
- unsupervised learning: concept learning, frequent item sets, clustering
- supervised learning: everything with labeled data which allows to make a prediction

1.2 **Data Preprocessing**

- Feature Extraction: Conversion of observation records into a formalized, computer-readable representation. Requires background knowledge from expert domains.
- Feature Selection: Removing values from instances, i.e. discard some features of a data set because these are: irrelevant, redundant, noisy/faulty. Possible benefits: improve efficiency and accuracy, prevent overfiing, save space. Strategies: unsupervised (based on domain knowledge, random sampling) and supervised (Gini-index, information gain, use a learning scheme’s performance)

1.3 **Machine Learning and Bioinformatics**

- the gain in speed to generate sequence data (nucleotide sequences) has clearly outpaced the speed of analysis and knowledge discovery
- current lab technology even cannot fill the gap between sequence and structure

1.4 **Prevalence of ANN and SVM**

- they are capable to handle a huge number of attributes
- they are quite robust against uninformative features
- they implicitly adjust feature weights during the training phase

- you do not need to have an idea about the meaning of an input i.e. no background knowledge or understanding for feature selection or even stronger for feature generation necessary
- disadvantage: these methods are “black box” models, so inspec+ng the model does not really increase your knowledge/understanding
- probable disadvantage: performance depends on number of assumptions in the various processing steps
- probable disadvantage: consider the number of free parameters in respect to the number of available training instances

1.5 **Redundancy Reduction**

Due to “experimental” reasons the sample represents only a special subset of the entities. The sampling of the “global” distribution is not fair. Possible solution is applying redundancy reduction to make the data a “fair” sample.

- CD-HIT: clusters sequences according to a user given threshold
- UniqueProt: creates representative, unbiased set of protein sequences based on HSSP values

1.6 **Performance Estimation**

- LOOCV: Leave on out cross validation: always one example is hold out for testing, the remaining for training
- N-fold cross validation, typically n=10:
  - partition the data in n partition
  - use n-1 partitions for training
  - use 1 partition for performance assessment
  - repeat with a different hold-out partition
  - average performance

1.7 **Drawbacks**

- Too many free parameters (edges) as well as overtraining leads to **overfiing** (prediction model is biased towards the training examples)
- Class imbalances. Solution: oversample minority class, downsample the majority class

Questions

Question 1: Can we predict secondary structure from protein sequence?

> Yes, we can

Question 2: What information do we obtain when predicting protein secondary structure? What features are predicted?

> Helix, sheet, coil formations

Question 3: How can we estimate the performance of secondary structure prediction methods?

> - Accuracy

> - Qx-measure (Q3)
> - Significance
> - Cross-validation

Question 4: Most often secondary structure predictions refers to the prediction of alpha helices, beta sheets and random coils. What other features of protein structure can be considered as secondary structure and be predicted?

> I asked Lothar about it and he said that this question doesn't have any meaning, even he could not answer it: only something like different types of turnes, but it's not very relevant

Question 5: List two secondary structure prediction methods.

> Prediction methods: Chou-Fasman, GORIII, ANN (PHD)

Question 6: Initially, prediction methods often focused on alpha-helices or underpredicted beta-sheets. What is the difficulty in recognizing beta-sheets from a window-based prediction method?

> The general opinion was, that the distance between the single strands, that form a beta-sheet, are too distant to be catpured by just local information. This was disproved after balancing the dataset (equal representation of helix, strand, other). The overall Q3 accuracy dropped a bit, but the accuracy of sheets increased significantly.

2.7 Homology Modeling

29.05.2017 | **Slides** | **Wiki**

1. Ideas and Keywords

Ideas

- Sequence Determination (short: sequencing) of DNA is highly automated today and very cheap
- Computer programs can help identify genes and coding regions
- From coding regions you can infer the protein sequence (1D information)
- **The entire sequencing process is cheap and quick, everything after that isn't.**

Available types of Data

- sequences (1D information)
- Annotations of already investigated proteins
- (few) protein structures (3D information)
- **Goal:** clever combination to infer more knowledge about yet unknown protein

Additional Tools / Databases

- UniProtKB, PDB
- Blast, Smith-Waterman
- PSI-Blast, ClustalW/CluastlX, MaxHom, SAM / HMMer, T-Coffee
- HHblits: SSearch, PSI-Search

Homology / Comparative Modeling

- **Goal**: Direct link from 1D to 3D structure (this would be the ultimate jackpot, but it does not work so far)
- Work around: Borrow structure from already know, sequence-similar proteins
- Tools: Modeller, Swiss-Model
- **Modeller**
  - uses a set of **spatial restraints** applied as PDFs (probability density functions)
    - $C_\alpha - C_\alpha$ distances
    - main chain $N - O$ distances
    - main-chain and side-chain dihedral angles
  - Which PDFs? Derived from analysis of 17 homologous protein families
  - needs a related template with a known 3D structure
  - Features

        - models non-hydrogen molecules
        - de-novo (?) prediction of loops
        - local installation
    - **Typical steps**
        - 1) identify templates / fold recognition
        - 2) align
        - 3) model
        - 4) assess
        - 5) refine
- **Swiss-Model**
    - originally: fully automated, little user interaction
        - 1) selection of templates
        - 2) modeling (copying coordinates)
        - assessment
    - now: more interactive and sophisticated model assessment
    - sever based service
    - convergent evolution with Modeller

Secondary Structure Prediction

- actual secondary structure of amino acid depends on the local sequence (context)
- even identical stretches (up to 5 aa) can occur in different secondary structures
- which structure is preferred depends on the available **hydrogen bond** opportunities
- more hydrogen bonds => more stable
- **Chou-Fasman**
    - simply look a the frequency an amino acid occurs in each secondary structure
    - search for **nucleation regions**
        - for helix: 4 out of 6
        - for sheet: 3 out of 5
    - extend until a window of 4 amino acids drops below 1
    - turns also check for Proline and Glycine
    - More info, in case this was not enough to understand: https://en.wikipedia.org/wiki/Chou%E2%80%93Fasman_method
- **GOR I**
    - 17 amino acid window
    - considers the state of 8 aa neighbors on each side (bayesian)
    - builds on three matrices (17X20) for helix, sheet and coil
    - (the original 'turn-matrix' was removed since it showed too high variability for a window of 17 aa)
    - thresholds:
        - 4 amino acids for helix
        - 2 amino acids for sheet
- **GOR III**
    - in addition for GOR I it considers all pairs with on the sliding window (= segment)

    - still not good for sheets, since the could be formed by non-local interaction
- **PHDxxx**
    - usage of **local evolutionary information** in the form of sequence profiles generated from multiple alignments
    - usage of global features (length, aa composition, ...)
    - the use of **redundancy-reduced**, **balanced data set** for training can be useful
- **PHD(-acc, -sec, -htm)**
    - add a second layer of networks (PHDsec)
        - L1: sequence residue -> secondary structure of that resisude
        - L2: secondary structure state -> secondary structure state for consolidated (smoothened) predictions
    - create a **jury** between balanced and unbalanced trained networks and different output states

Performance Assessment

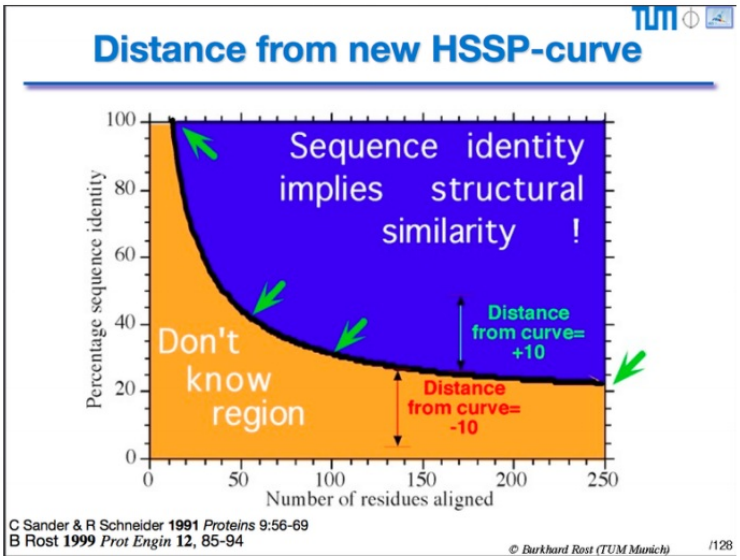
- Precision, Recall, Accuracy
- Qx-Measure: For x states, fraction of correct predictions (TrueNegative + TruePositives) of all predictions
- Significance?
    - determine the average Q on you dataset
    - calculate the standard deviation (sigma)
    - calculate the standard error (sigma / sqrt(N))
    - N is size of test set
- Compare Methods
    - compare always on the same instances
    - test / training split have to be the same for both tools
    - not overlap allowed between test and training set (structures in comparative modeling range violate this)
    - alternative: compare on fresh data published after publication of methods

Membrane Proteins

- hydrophobic stretches, typically 17-21 amino acids long
- Positive Inside Rule (connecting sequences on inside of cell are positively charged) to determine topology
- Signal Peptides (often confused with TMHs)
- in 3D: **hydrophobic parts on the outside**
- many hydrophobicity indices out there
- optimized scoring matrices
- Nowadays:
    - interrupted TMHs
    - reentrant parts (leave membrane on entry side)
    - coil regions inside membrane

HSSP Curve

Homology-derived Secondary Structure of Proteins



2. Exercises

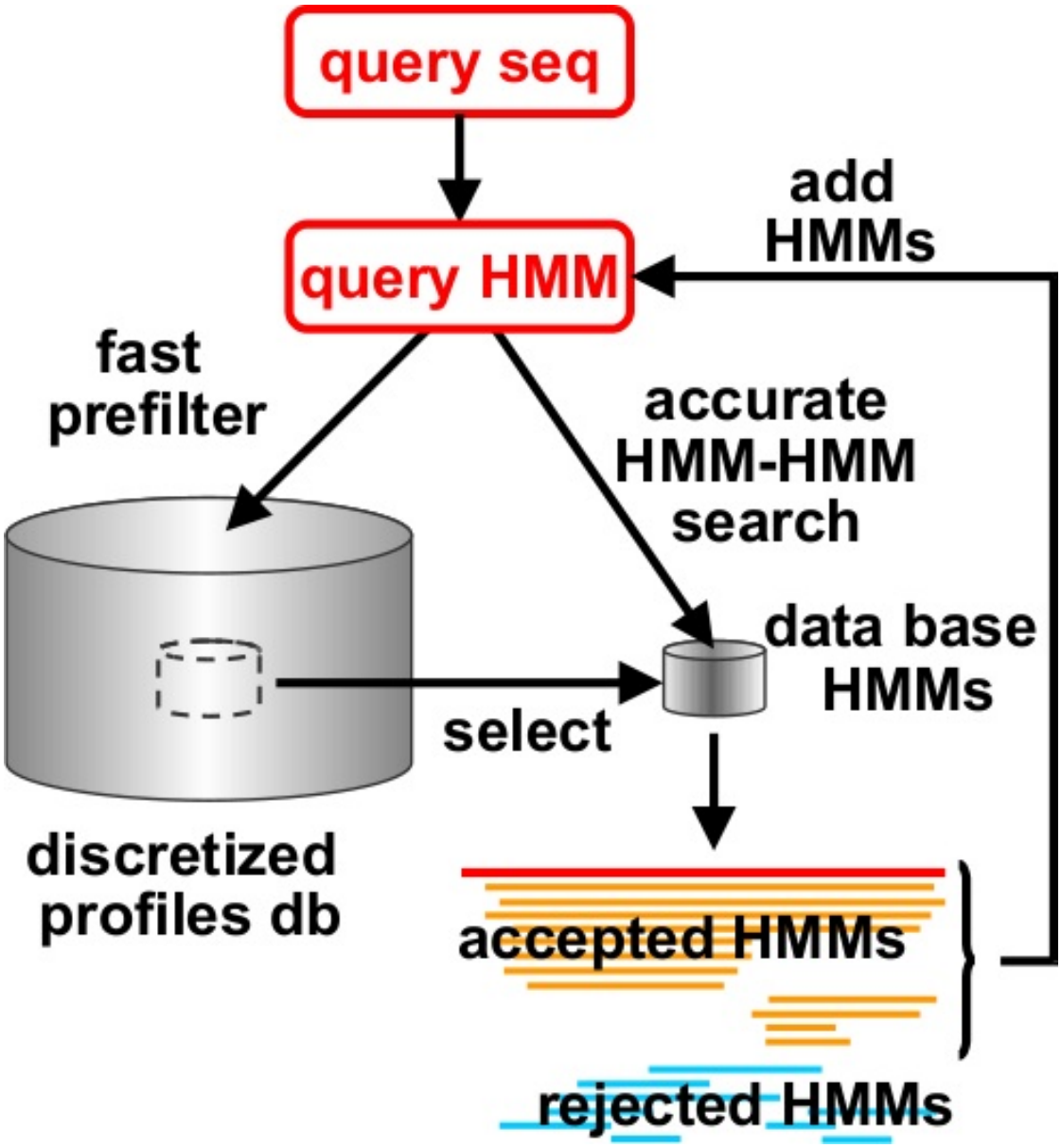
Questions

Question: How does ClustalW work? How does it differ from BLAST?

> Clustal is a series of widely used computer programs used in Bioinformatics for multiple sequence alignment. from Wikipedia All variants of Clustal align sequences by three main steps:
>
> - Do a pairwise alignment
> - Create a guide tree (or use a user-defined tree)
> - Use the guide tree to carry out a multiple alignment
>
> ? BLAST itself is not a multiple sequence alignment tool

Question: What is HHblits?

> A multiple sequence allignment tool using HMMs *HHblits-Schematic*



Question: What is the definition of accuracy? Is it the same as Qx?

> ?

2.8 Wrap Up

06.07.2017 | Slides | Wiki

3. Exam Questions

> This section contains possible exam questions compiled from different sources.

- 3. Exam Questions
  - 3.1 Lecture Questions
  - 3.2 Exercise Questions
  - 3.3 Question Catalogue

Contribution Guide: Exam Questions

Since all questions here are answered by students, there might be some mistakes in them. Hence a few more words on how to best handle this section.

1. Adding a new question

Just add the question in the respective file. Optimally, you can already provide an answer.

2. Answering a question

To clearly distinguish questions from answers, please put answers in **blockquotes** right under the respective question.

Example:

- How can 1D secondary structure information be used to derive a 3D model?
- > It is not possible to derive a 3D model from 1D information. (Trick Question)

3. Updating an answer

If you think an answer does not properly answer a question (e.g. it is wrong or the answer is not sufficient), mark the answer and open a new **issue** on Github to discuss the question and share your improved answer.

(Use the emoji to mark the possibly wrong answer inline)

Example:

- How can 1D secondary structure information be used to derive a 3D model?
- > It is not possible to derive a 3D model from 1D information. (Trick Question)

3.1 Lecture Questions

> This section contains possible exam questions asked Professor Rost in the lectures he dedicated to answering student questions. They are **highly relevant** because he will sample exam questions from this pool.

Questions (Thursday, 22nd June)

Question: How can you choose the **e-value** for PSI-BLAST depending on the size of the dataset?

> - The E-value indicates significance of alignment/ hits returned by chance when searhing through DB.
> - It depends on the size of dataset and length of query.
>
> So higher e-values from large DB aren't always bad (and opposite: smaller e-values from small sample space isn't always good). (?)

Question: You want to develop a new method to predict e-values, how do you prepare your data?

> You need to look at how the e-value changes through iterations, width of background distribution, height of score

Question: What is the regular process when you want to analyse a new sequence?

> [Note: A question such as this would need more information I think. What do I want to know about the new sequence?]
>
> 1. Search UniProt whether the sequence is known. If so, also check the PDB - maybe there is already a 3D structure.
> 2. Run BLAST against the PDB to find homologs. If there are suitable hits, homology modeling might be possible. In such a case 'Modeller' or 'Swiss-Model' could be used to predict a 3D structure for the protein. (In addition, informed estimations about the function of the protein can be made based on the homologs found.)
> 3. If Homology Modeling is not possible, there is still more to uncover about the protein. First, one could search for motifs and patterns. Second, PSI-BLAST could be used to builda profile of the family, thereby uncovering more distant relatives and conserved regions.
> 4. To further analyse the sequence secondary structure and/or membrane predictions (is it a membrane protein?) can be of used.
>
> Which analysis method we choose, ultimately depends on, what we want to find out about the discovered sequence.

Question: What is a structural domain? What is a functional domain and How can we deal with the fact that they can be in different places?

> - structural domain: part of sequence with unique 3D structure

> - functional domain: part of sequence with unique function
> - use local alignment

Questions (Tuesday, 27th June)

Question: How do we predict proteins?

> [Note: Too general question for the exam...]
>
> [1] Predict Function:
>
> - We cannot predict function directly. The general assumption is that from same 3D structure we can assume same function.
>
> [2] Predict 3D Structure
>
> - The only reliable way to predict 3D structure is **homology modeling**.
> - Find homologs with known 3D structure by sequence identity (assumption: high sequence identity -> similar 3D structure)
> - Use homology modeling (Modeller, Swiss-Model): Apply known 3D structure of homolog to query sequence and refine constraints
>
> [3] Predict Secondary Structure
>
> - Use algorithms (GOR III, ProfSec) to predict secondary structure. Possibly as input for further analysis.
>
> [4] Predict Membrane Proteins
>
> - Predict different features about Trans-Membrane Proteins.
> - Is a protein a membrane protein (check e.g. hydrophobicity)
> - Are there Trans-Membrane Helices (TMH) or Beta-Barrels? At which postion are they?
> - In which direction is the membrane crossed? (Toplogy, Inside-Positive Rule)

Question: Why would someone give you a sequence?

> The amount of newly discovered sequences constantly increases (sequencing is very cheap and fast) And even for known sequences we still can refine information that we have about them

Question: How do you run a sequence against the DB?

> - BLAST search (uses indexing technique, scoring matrix and dynamic programming to find short similar segments)
> - PSI-BLAST search (at first it uses Blast search, creates profile (PSSM) from highest scoring hits and uses it to replace substitution matrix in a subsequent searches, this process can be repeated many times to refine profile)

Question: How do you build a family?

> 1. Run BLAST against the entire database (e.g. UniProt) to find proteins with high sequence similarity. From these 'homologs' found, we can assume an evolutionary relationship based on their high PSI. (Important! Get the statistics right and only add proteins which are from the Daylight Zone)
> 2. Calculate a PSSM with the homologs found. (= profile) This reveals conserved residues / regions.
> 3. Use this PSSM to perform a Profile-Sequence search against the database to discover more distant family members.
> 4. Recalculate the PSSM to refine the profile with the newly found proteins. Perform another round of Profile-Sequence search against the Database, if no stop-criterion was reached.

Question: Pairwise/Multiple alignment: what can we achieve, what is the risk?

> Pairwise Alignment aligns exactly 2 sequences. This is necessary to e.g. further compute 3D similarity or make assumptions about evolutionary and functional relatedness. It is important to always consider the background distribution, length of sequence, size of database. (= How probable is it that such an alignment score happened by chance?
>
> Multiple Sequence Aligment tries to align more than 2 sequences. This is computationally infeasable for more than 6 proteins, so 'shortcuts' are needed. (e.g. CLUSTAL, BLAST). Multiple sequence alignments help to uncover conserved evolutionary information about a protein family, which in turn can help to find more distant relatives (in terms of sequence identity). (In case of Multiple alignment there is a risk of pollution, any errors in the initial alignments cannot be corrected later as new information from other sequences is added.)

Question: What is the difference between pairwise and multiple alignment?

> Pairwise alignment compares 2 sequences, multiple aligment- 3 and more. (Also look previous question)

Question: Can we predict something we have not observed?

> When we predict features we can try to find somthing that people "want" to see.

Question: Sliding windows introduce information from the sequence environment around a residue. Why do we need a second neural network on top of that? Why do we need anything else on top?

> The second (independently trained) structure-to-structure network helps to improve prediction performance for helices and sheet by solving the 'short segment' problem. It gives us ability to detect motifs whenever it's in sequence window. It exploits spatially local correlation.

Question: How do we prepare data to predict B-value? *(not sure about correctness of question)*

> 1. B-Values in principle are conserved. Thus the B-Values for proteins across the PDB have to first be normalised (differently depending on family)
> 2. Thresholding. In order to map the continuous space to outputs (rigid / flexible), thresholds have to be set (not at the peak of the distribution, since the experimental error is the biggest there.

Questions (Thursday, 29th June)

Question: With a matching *profile-profile* comparison what can you say about the two families?

> The assumption is that matching profiles share a similar / same structure and function. (?)
> Also evolutionary connection?

Question: When I build a profile of a famliy: Do they share the same structure? Should I verify that they do? How do I do that?

> - The very assumption is that the proteins of one family share the same structure and function.
> - When iteratively refining the profile with proteins retrieved by *profile-sequence* of *profile-profile* comparison (from the twilight- / midnight-zone), it can make sense to double check the new proteins with secondary structure prediction to avoid adding false-positives to the family.

Question: Cross-Validation: What is it? How does it work? Why do we need it?

> What is it? How does it work?
>
> - Partion trainingdata in 'n' folds. Use 'n-1' for training and 1 for performance assessment. Repeat with different hold out partitions. Average perfomance.
> - Leave-one-out-cross-validation: Use only one sample as validation set instead of 1/n-th. In practice, this is only used if extremely few samples are available.
>
> Why do we need it?
>
> - Validation is generally used to ASSESS the parameters and hyper parameters of an algorithm and then ITERATE over those parameters. Because we want our assessment to be as good (high generalization) as possible. We want to use all available data to do the assessement.

Question: What is the difference between a BLOSUM matrix and a PSSM (Position Specific Substitution Matrix)?

> - In case of a PSSM amino acide substitution scores are given separately for each position in a protein sequence.
> - In BLOSUM62 we have substitution scores for all possible substitution pairs (210 in total) of the 20 standard amino acids. (NOT specific to their position in a sequence)

Question: What is the most successful method to predict 3D structure?

> Homlogy (Comparative) Modeling.

Question: What is homology modeling (= comparative modeling) and how does it work? What are the limitations of it?

> 1. Align sequence with proteins in PDB and set a threshold. Introduce gaps as loops.

> Limitation - no similarities found (templates are unavailable or fragmentary), matching right residues (errors in sequence alignment produce errors in the homology model)

Question: How can you predict structure in the [a] daylight- [b] twilight- [c] midnight-zone?

> By using [a] sequence-sequence, [b] sequence-profile and [c] profile-profile alignment respectively.

Question: What is the assumption behind all alignment methods that is incorrect and nevertheless seems to work? Give a method that aligns 2 proteins without that assumption.

> The assumption is that the alignment of the residue at position **i** is independent of the residue **i+x**. (In short: alignment i and i+1 are independent).
> The only method that does not rely on this assumption is the **Genetic Algorithm** (e.g. T-Coffee)

Question: Why do we have so few experimentally confirmed structures in the PDB?

> ca. 85 million proteins are known (UniProt), but only the 3D structure of about 120 000 are in the PDB.
>
> - Sequencing technology has improved and become a lot cheaper and faster. Hence, many more proteins were discovered.
> - While there were improvements in 3D structure determination, it still costs at least 100 000 euros to experimentally determine the 3D structure of a protein.
> - It is expected that this divide will further increase.

Question: Why do we need to perform redundancy reduction on our dataset befor training our machine learning model?

> The training data for the ML model should be representative for the problem for which it should predict.
>
> - The known 3D structures in the PDB largely share a high sequence similarity (because of different resolutions, competing groups, different research goals, ...), are thus applicable for homology modeling and not representative for the dataset we want to predict for in the future.
> - Removing the sequences in homology modeling range (redundancy reduction) creates an unbiased dataset.

Question: Say the 3D structure for **N** thousand proteins were known and they serve as input for a method predicting 1D structure. How can you define the value for *sequence-unique* that you have to apply to create an unbiased data set? Why do you need an unbiased dataset?

> The threshold should be the border (or a little bit left) of daylight and twilight zone. This is because for proteins in the daylight zone we can do comparative modeling and thus there is no need to make predictions on them. Since those proteins will also never be used as inputs

> for our predictor, we NEED to remove them in order not to bias our model on them. (The model needs to be trained on the instance/sample space from which the to be predicted instances will be drawn at prediction time = not in homology modeling distance).
>
> If we would still need to include the proteins in homology distance, we should at least reduce the top X percent to remove duplicates (they might be in there due to competing groups doing research at the same time, difference in age and resolution,..).

Question: How do you compare proteins of different length?

> By using local alignment (Smith-Waterman)

Question: What is the significance in using information from protein families (inferred evolutionary information) as input to the ML device predicting 3D structure?

> - The profile is a record with information about the 3D reality of the protein, showing evolutionary conserved regions.
> - It is thus additional information for the ML device, which is clearly relevant for the structure.

Question: How can I use 1D information to get a 3D structure? What can you do with a 1D structure?

> It is impossible to reconstruct a full 3D structure from 1D information.
>
> 1D structure can be used for
>
> - optimizing a profile
> - predict whether a protein is soluble
> - predict whether a protein is a transmembrane protein
> - input for further secondary structure prediction

Question: What is a 2D contact map (distance map)? How can it be obtained?

> - A contact map shows the pairwise distance between all amino acids in a sequence (protein).
> - It contains the same information as the 3D structure, except for the chirality (mirror image)
> - It is obtained from the known 3D structure with the use of distance functions (i.e. Voronoi contacts)

Question: Explain the concept between the notation of 1D, 2D, 3D structure. What is in the PDB? What does the DSSP give?

> - 1D - secondary structure (HEL)
> - 2D - contact map
> - 3D - tertiary structure, 3D shape of protein
> - In PDB there is only 3D structure
> - DSSP assigns secondary structure according to hydrogen-bond pattern from 3D structure as input.

Questions (Tuesday, 4th July)

Question: Why do we need separate methods to predict secondary structure for membrane and water-soluble / non-membrane proteins? What is needed for membrane prediction beyond secondary structure?

> - Methods trained to predict secondary structure in soluble environment fail for membrane proteins (empirical observation).
> - Reason for this is that the environment (the membrane) is very different.
> - When predicting membrane proteins the following information is important:
>   - Number of TMHs
>   - Position of TMHs (+- 5 residues overlap)
>   - Topology of TransMembrane Protein (where is inside / outside)

Question: What is the principle difference between PSI-BLAST and CLUSTALW (or any other multiple sequence alignment method)?

> The Optimization Criteria is different:
>
> - BLAST: Builds up a PSSM
> - CLUSTALW: Optimizes the family alignment (in a dynamic programming fashion, wich is computationally slower, but may prodice a more 'optimal' result)
>
> The differ in the way the background statistics are compiled:
>
> - BLAST is so fast because it compiles the background statistics only once
> - CLUSTALW on the other hand uses random sampling.

Question: How do you measure the similarity between a profile (PSSM) and a sequence?

> [Note: This question was not sampled by Prof. Rost. I just think it is important to understand the workings of a profile.]
>
> - You simply calculate the PSSM score for the given sequence.

Question: In the 1st iteration PSI-BLAST finds the most low hanging fruits through pairwise comparison. [a] What does it do in the 2nd iteration? [b] Why can this work better? [c] What could happen that makes the 3rd iteration not find more more hits than the 2nd one (for all n=1=…N)? [d] Say n+1 finds many more hits than n: everything ok?

> [a] It uses the compiled PSSM to run a profile-sequence comparison against the database.
>
> [b] Because the PSSM contains information about the entire family, the initial query sequence belongs to. Within the PSSM there is information about the evolutionary preserved (and thus for structure / function important) segments/residues for this specific family.
>
> [c] This either happens if the profile converged. There are 2 reasons why this happens.
>
> - All the proteins belonging to the family where found

> - If this happens very early e.g n=2, it indicates that BLAST could not find enough proteins to build up a profile (family is too small).
>
> [d] It could have happened that the profile was messed up by proteins, which are not part of the family in an earlier run. The new PSSM does not discriminate well enough between proteins.

Question: How is the **e-value** for PSI-BLAST, FASTA calculated.

> The e-value is a property of a score (the score of an alignment consisting of two sequences and a scoring matrix). It reflects the expectation value of the number of alignments that have a score >= the Score S.
>
> Calculation: E(Score) = N / (2 ^ S'(Score)) S'(Score) = (lambda*Score - ln(K)) / ln(2)
>
> where lambda and K depend on the scoring matrix
>
> For PSI-BLAST the e-value needs to be recalculated after each run.

Question: Why might per-residue scores poorly reflect the performance of transmembrane prediction? Invent an alternative method to score TMH prediction methods.

> - For most membrane proteins, most residues are NOT in the membrane, thus a high per residue score could still missed most TMHs.
> - It is not really what I want to predict. I want to predict the specific TMH. How many TMH do I have? At which position are they? Which topology does the protein have?
>
> Alternative Scoring Method: All TMH must be predicted correctly (+- five residues overlap, mind. 50% overlap) Additionally: The topology has to be predicted correctly.

Question: Method A is published to predict solvent accessibility at Q2=61%, Method B in another publication claims to achieve Q2=63%. What do you have to check to ascertain that methods B is really a better methods? (Address the terms significance (statistical, scientific) in your response)

> - Do both methods use the same way of computing Q2?
> - Do the use the same dataset?
> - Which data was used to train the method? (Did the have a proper blind-test-set)
>   - Best, test both methods on proteins that where released after both methods were published.
> - Is there a difference between the methods beyond the standard error? Is the difference in accuracy statistically significant (= higher than the standard error)?
> - Does the +2% help me to get better scientific results / insights? Is there an advantage in scientific terms? Baseline: Is it better than random?

Question: What features can be used to predict secondary structure from sequence? Argue why?

> - The PSSM is the most important feature!!! (provides evolutionary information)
> - … (many more)
>   - Length of the Protein
>   - Position within the protein (distance to N- / C-Terminus)

> - Amino Acid Composition of protein
> - ...

Question: What is the most accurate way to predict protein 3D structure (explain the idea behind the method)? Why does this methods hardly work for membrane proteins and even less well for disorder proteins?

> Homology (Comparative) Modeling: **Idea:** If we have a very similar protein in sequence identity of which we know the 3D structure, we can use this structure as a template for our query sequence. **Assumption:** High Sequence Similarity -> Same (very similar) 3D structure
>
> - For disordered proteins it does not work, because of the definition of disorder proteins. (No unique structure)
> - For membrane proteins, the set of available 3D structures is just too small. (166 redundancy reduced according to TMSEG lecture)

Question: UniProt currently holds about 85 Million proteins sequences: Do we have any idea about the structure of any of those? Roughly for how many? Do we have any idea how many of the 85 million are membrane / disorder proteins?

> - Yes, we have a 3D structure for those in the PDB (about 120k)
> - By membrane prediction methods, we can predict which proteins are membrane proteins. So we can predict the number of membrane / disorder proteins under consideration of prediction accuracy.
> - For the human proteon, it is estimated that about 25% are membrane proteins, however the 3D structure of only 1000 - 2000 membrane proteins has been experimentally determined.

Question: You want to use a regular neural network (input/hidden/output) too solve a certain prediction model. (e.g. predict disordered regions). How can you find how many hidden units you need? What the best input is? How to best code the output (here disorder)?

> How can you find how many hidden units you need?
>
> - There is really only one way: Try it. Usually one starts with too many (and overfits) and then reduces it until the test-performance does not get better anymore.
>
> What the best input is?
>
> - This depends on the task. Hence expert domain knowledge is needed. In general, all features that influence the to be predicted property should be included. NN learns the priorisation of the features itself.
>
> How to best code the output (here disorder)?
>
> - Try the most straight forward first. Here probably binary: isDisorderRegion isNotDisorderRegion. If doesn't work, try to split the label classes up (see. solvent accessibility).

Question: You want to develop a method to predict secondary structure in 3 states (HEL). How can you use DSSP to convert the 3D structure in the PDB to HEL. What do you have to watch? Can you use the entire PDB? Once you have N proteins in your dataset: how can you assess prediction performance? (How to measure ‘right’, name a few score that are relevant, how to measure statistical significance, how to measure scientific significance)

> 1. Redundancy reduce the dataset: "Remove" all proteins in comparative modeling range
> 2. "Remove" fraction of dataset for later blind testing (no overlap with training set allowed)
> 3. Use remaing dataset to train model
> 4. Use blind test to assess perfomance
>   i. Q3 = number of correctly predicted residues / total number of residues
>   ii. Prediction performance for each H, E and L
> 5. Statistical significance: Is my method better than others?
>   i. Determine average Qx - value on your dataset
>   ii. Calculate the standard deviation (sigma)
>   iii. Calculate standard error (sigma/sqrt(N))
>   iv. Compare the new method's performance to other methods' (baseline) performances. If the performance increase is larger than the standard error, it is statistically significant
> 6. Scientific significance
>   i. Does the perfomance increase helps to push new scientific findings?

Question: TMH prediction: How can you predict the direction of a helix? What assumption does comparative modeling make? Why do proteins always have to adopt the same 3D structure? Do different organisms use different proteins? How much does it cost (time/money) to experimentally determine the 3D structure of an average protein?

> 1. Positive-inside rule
> 2. Similar sequence -> similar struction
> 3. Yes different species have different proteins. However there is a evolutionary history / relatedness between species, so that similar proteins (homologs) are found.
> 4. X-ray: ~1 year, 100 000 euro (can go up to million in certain cases).
> 5. NMR crystallography: more time consuming and more expensive (due to spectrometer costs and isotope labelling) than X-ray crystallography. A standard 600 MHz NMR costs roughly $800,000, but the 900 MHz sells for about $5 million.
> 6. Cyro-EM: ??

Question: How would you define life? How are proteins crucial to maintain it?

> Descriptive definitions of life:
>
> - Homeostatis (regulation of internal environment to maintain constant state)
> - Organization (Unit: Cells)
> - Metabolism
> - Growth

> - Adaptation
> - Response to stimuli
> - Reproduction
>
> Functions of proteins: *Machinery of Life*
>
> - Defense (e.g. antibodies)
> - Structure (e.g. collagen)
> - Enzymes (metabolism, catabolism)
> - Communication / Signaling (e.g. insulin)
> - Ligand binding / Transport (e.g. hemoglobin)
> - Storage (e.g. ferritin)

Question: Why do we need membranes around cells? What are they made of? Why do proteins pass through membranes?

> - It physically separates the intracellular components from the extracellular environment and provides shape of the cell. Also membrane is a dynamic structure (cell can grow and shrink).
> - The cell membrane is a bilayer of phospholipids.The phospholipid bilayer is hydrophilic on the outside and hydrophobic on the inside.
> - Transmembrane proteins are anchored into the bilayer by their nonpolar segments. TMP can move laterally (sideways) in the membrane.

Questions (Thursday, 6th July)

Question: What does the reliability of variants tell us about the severity of effect and why?

>

Questions (Tuesday, 11th July)

Question: State one example in which the replacement of graphs by hypergraphs reduces the information-loss when modeling cellular and/or physical systems.

> When modeling protein interactions, some functions are the result of protein complexes with more than 2 proteins. A hypergraph can have edges between an arbitray number of vertices, whereas a regular graph can only connect 2 vertices with and edge. If we modeled the interaction of 3 proteins with a regular graph, we would [1] either connect all nodes pairwise, losing the information whether they do (or do not) interact pairwise as well [2] don't model the 3-way interaction at all (loosing exactly this information)

3.2 Exercise Questions

> This section contains possible exam questions asked and answered as part of the exercises.

Exercise 2: Protein Structure

Question: What are the building blocks of proteins?

> - Proteins are polypeptides, which consist of amino acids as building blocks

Question: Define protein backbone and amino acid side chain in 1 or 2 sentences for each term.

> - **Backbone:** All amino acids have the same with an N-Terminus on one and a C-Terminus on the other side by which they are chained up into a protein. (Always read / write from N- to C-Terminus)
> - **Side Chain:** Each Amino Acid has a different side chain, which equips it with unique features (length, polar, non-polar, electrically charged)

Question: How many amino acids appear in proteins? How can they be classified?

> All proteins are built from **20 different amino acids**. The can be classified into the following 5 groups /categories
>
> - positively electrically charged side chain
> - negatively electrically charged side chain
> - polar uncharged side chains
> - hydrophobic side chains
> - special cases

Question: Name atom types involved in a hydrogen bond. Do S-H groups form hydrogen bonds? Why (not)?

> Hydrogen bonds normally involve
>
> - an **H** atom (obviously) as donor
> - either an **O** or **N** atom as acceptor
>
> Due to its lower electronegativity Sulfur does not engage in classical Hydrogen-Bonds. (However nonclassical hydrogen bonds with Sulfur seem to exist, according to literature)

Question: How are alpha helices held together?

> Alpha-Helices are stabilized by **hydrogen bonds** along the backbone. Every turn takes about **3,6** residues.

Question: What is similar and what is different in the hydrogen bonding of the alpha helix and the beta sheet?

> - Hydrogen bonds in **Alpha Helices** are 'ultra-local' and occur along the backbone of the

> helix (each 3.6 residues)
> - Hydrogen bonds between **Beta Sheets** happen between the beta-strands (running parallel or antiparallel) which can be rather far apart in sequence.

Question: Why do we find "forbidden" areas in a Ramachandran plot?

> The Ramachandran plot shows the angles (ϕ, ψ) of amino acids in which alpha-helices and beta sheet are typically observed. Forbidden areas are those that are not possible due to physical constraints due to e.g. the side chain of the amino acid.

Question: What is a protein domain?

> A domain is a conserved (sub-)sequences of a protein, which adopts a unqiue 3D structure when put into solvent.

Question: How many amino acids are typically found a in a domain? Why is there a minimum/maximum size?

> Proteins domains are found between 36 and 690 residues long. Most protein domains have a length of around 100 residues.
>
> Minimum Size: ? Maximum Size: ?

2.2 PDB

Question: How many structures are stored in the PDB? How many of those are protein structures?

> About 130 000 entries can be found in the PDB, of which around 120 000 are protein structures.

Question: Which experimental methods are (mainly) used to determine the structures? How long does it on average take to find the 3D structure for one protein for each method?

> The largest part (90%) of proteins structures are determined by X-Ray crystallography, followed by NMR (9%) and Cryo-EM (1%). However, the number of new Cryo-EM structures is projected to overtake the number of new NMR entries in 2017. Current sciene pushing the limits of Cryo-EM resolution, makes it a promising technology for future 3D structure determination.
>
> **Costs / Time per Method**
>
> - X-Ray: (100 000, ?)
> - NMR: (?, ?)
> - Cryo-Em: (?, ?)

Question: How many human ("Homo sapiens") protein structures are in the PDB?

> About 37 000 protein structures of homo sapiens. (Not, sequence unique entries though)

Question: Why does the number of protein structures already decreases when reducing at 100% sequence identity? Why does it decrease when reducing at even lower sequence identity further?

> - The PDB has redundant 3D structures for certain proteins from different experiments. This has different reasons such as competing groups, different research goals, better resolution.
> - Since most proteins developed under evolutionary pressure, large parts of proteins of the same family share a high sequence overlap.

Exercise 3: Alignments

Question: How can you define similarity between two protein sequences?

> Two ways to define similarity are to compute the percentage of residues that were aligned
>
> - and matched on the same amino acid (PSI, Percent Sequence Identity)
> - and matched with a positive score in the Substitution Matrix

Question: What does "conservation" mean in the context of sequence alignments?

> - Conserved sequences are similar or identical (sub)sequences that occur within protein sequences.
> - By compiling homolgues proteins (a family) into a profile, it become clear, which subsequences are more conserved and thus more important for the function of the protein family

Question: Why are sequence alignments useful?

> - They are useful because the allow use to find the 'best' possible match between two sequences. Only this allows use to
>   - compare their 3D structure
>   - create predictions about their similarity in 3D structure
>   - make assumptions about their evolutionary relatedness

Question: What are the main differences in the algorithms of Global and Local alignment? Why does it make sense to not always perform a global alignment.

> - Global alignment methods always align 2 sequences from beginning to end.
> - Local alignment methods only align subsequences.
> - For Global alignment to be meaningful, the sequences should have similar length. Since proteins are between 35 and 30000 residues long, it does not make sense to always use global alignment.
> - Also when e.g. looking for proteins with a certain domain it does not make sense to use global alignment, since we are explicitly looking for subsequences.

Question: Which amino acids can (with high likelihood) be substituted for Leucin without having an effect on protein function?

> Methionine (2). Isoleucin (2), Valine (2), Phenylalanine (0), because the have a positive score in the BLOSUM62 matrix

Question: Which substitution is more probable according to PAM250 and according to BLOSUM62: [a] W <> F [b] H <> R

> - W (Tryptophan) <--> F (Phenylalanine)
>   - BLOSUM: 1
>   - PAM250: 0
> - H (Histidine) <--> R (Arginine)
>   - BLOSUM: 0
>   - PAM250: 2
>
> => For BLOSUM W <--> F is more likely to be observed
> => For PAM250 H <--> R is more likely to be observed

Question: What is a multiple sequence alignment?

> A method to align multiple sequences against each others.
>
> - building a consensus sequence and aligning new sequences against it
> - building a search tree
> - building a profile (like PSI-BLAST)

Question: What kind of sequences are likely to be used for an MSA? In which relationship are they to each other?

> - Most likely, sequences which we assume a evolutionary relationship (homology) will be used. Following the homology assumption, we expect sequences with a high PSI to have a similar structure because they have a common ancestor.
> - Building up a profile with them, would then allow us to discover conserved regions and use the profile to find more candidates in the Twilight Zone with Profile-Sequence comparison.

Question: Why would you want to align multiple sequences? What kind of information is contained in MSAs but not directly in e.g. all-against-all pairwise alignments?

> Building up a profile with similar sequences, would allow us to discover conserved regions which developed under evolutionary pressure.
> By compiling such a family of proteins (we assume that they have a common ancestor -> homology) into a profile, we can find more candidates in the Twilight Zone with Profile-Sequence comparison.

Question: Given your knowledge of the algorithms for pairwise alignments, how could you calculate an MSA? Is that a feasible approach? Why?

> One approach would be to sequentially align sequences against a consensus sequence. This approach, however, comes with problems such as the need for a strategy how to find the consensus for a certain amino acid at a certain position, the fact that the order of alignment might matter, etc.

> A better approach is the creation of a PSSM (position specific scoring matrix), which contains for each position of the profile the likelihood that a certain amino acid occurs there.

Question: You have a sequence which you would like to find in a database. Which search method and which E-value cutoff do you use, [a] if you know your sequence is in the database and only want to find that entry [b] if you would like to find homologs.

> [a] Pairwise alignment (??? is this right???)
> [b] BLAST

Question: What is the difference between BLAST and PSI-BLAST?

> BLAST uses the BLOSUM matrix to retrieve homologs. It runs only once and returns the found sequences.
>
> PSI-BLAST uses BLAST in the first run to find homologs and build a profile. By using, and iteratively rebuilding, the profile it can find more distant (in terms of sequence identity) homologs with Profile-Sequence alignment.

Exercise 3: Resources for Bioinformatics

Question: How many structures in PDB have a resolution with =< 2 Angstrom

> ?

Question: Which term from computer science you would use to describe PROSITE patterns (e.g. PDOC00022).

> ???
>
> -> I would argue it reminds me of regular expressions, but this should be verified.
> eg see here: http://www.hpa-bioinfotools.org.uk/ps_scan/PS_SCAN_PATTERN_SYNTAX.html

Question: Which type of information does STRING provide?

> STRING is a secondary database which provides information about known and predicted protein interactions.

Question: What does PFAM-A contain?

> PFAM-A contains manually curated information about proteins families. It is especially useful, because it contains
>
> - Profile HMMs
> - Seed alignments
> - Full alignments with all hits

3.3 Question Catalogue

> This section contains possible exam questions sourced from students of previous Protein Prediction I lectures and the lecture recordings.

3.3.1 Exam Structure: 2016ST

> We where able to obtain last years exam structure. Let's try to answer all of the concrete questions :-)

Part 1 is mandatory, for the rest choose 3 out of 4.

1. Multiple Choice (5 questions, 10 points)

- Secondary Structure
- RMSD - Protein Similarity
- Hydrogen Bonds (α-helix, β-sheets, long/short bonds)
- 100% sequence identity => same structure? (PIDE)
- Can modern prediction methods correctly predict structure in the midnight zone?
- About "Cryo-Microscope"
- About "X-Rays"

2. Sequence Alignment (10 points)

- Explain each of the following alignment techniques and provide one method for each
    - Sequence - Sequence
    - Sequence - Profile
    - Profile - Profile
- General scoring BLOSUM62 matrix vs. PSSM
- Why is the sequence information valuable?
- How BLAST speed up pairwise alignments?
- Global vs Local alignment

3. Sequence Structure (10 points)

- What data is needed to predict the structure with ML?
- Which tools and db you will use?
- How to prepare data for ML
- Which 2-3 features will help to predict?
- Would you apply method to all protein (query)?
- Which measure would you use to evaluate your method?

4. Protein Structure (10 points)

- Why it is important to know 3D structure?
- Why is it so hard to compare 3D structure?
- Most successful ML algorithm for predicting structure, steps
- Method for experimental structure determination. Short explanation. How many structures are experimentally known?

5. Machine Learning (10 points)

- General definition of Machine Learning
- Cross validation
- What is ‘feature’?
- ETP explain, example
- Name and describe one ML method
- Name and describe "sequence" in context of PP
- Discuss how to predict Protein Structure from amino-acid sequence using ML
- Q2: which is better, how to prove your’s is better, which value you will publish? (What is Q2)

3.3.2 Lecture 1: Introduction Bioinformatics

Question: What is common to life?

> DNA, Protein, RNA

Question: How many bacteria do we care around?

> About 2 kilos. Humans carry around more bacterial DNA than human DNA.

Question: Which elements make up life?

> - 65.0 % - O, Oxygen
> - 18.6 % - C, Carbon
> - 9.7 % - H, Hydrogen
> - 3.2 % - N, Nitrogen
> - 1.8 % - Ca, Calcium
> - 1.0 % - P, Phosphorus

Question: What is life? Can you define it?

> There is no holistic definition of life: Descriptive definitions of life are
>
> - Homeostatis (regulation of internal environment to maintain constant state)
> - Organization (Unit: Cells)
> - Metabolism
> - Growth
> - Adaptation
> - Response to stimuli
> - Reproduction

Question: Are viruses life?

> Strictly speaking NO. Viruses on their own cannot replicate and thus are not alive. However, one could say that viruses are alive / represent life once they infected a cell and replicate.

Question: What do bacteria have in common?

> Single Cells

Question: What are the differences between prokaryotic and eukaryotic cells?

> **Prokaryotic Cells:** mainly found in bacteria and archaea, usually unicellular, no nucleus, no cell organelles
> **Eukaryotic Cells:** Found in animals and plants, usually multicellular, have nucleus, have cell organelles

Question: How can the density of a cell be described?

> The state inside a cell is almost solid. We can think of a cell similar to a Christmas day on Time Square: Everything is densely packed, but there is still movement.

Question: What is the smallest building block of life that can replicate?

> cells

Question: How many different cells are in a typical human?

> 200

Question: What are the parts of the cell called?

> organelles

Question: Which part of the cell is called the "powerhouse"?

> mitochondria

Question: What part of a plant is involved with photosynthesis?

> chloroplast

Question: What is mitosis?

> cell division

Question: Who first used the term cell?

> Robert Hooke

Question: How many elements are found in amounts larger than trace amounts (0.01%) in our bodies?

> 11

Question: When communities of living things interact with non living things they are called ... ?

> ecosystem

Question: The most common molecule in the human body is ... ?

> Water: H_2O

Question: What do bacteria have in common?

> Single Cells

Question: What is a gene?

> A gene is a region of DNA, which contains all information for the creation of an entire RNA strand.

Question: What is DNA made out of?

> DNA is a linear polymer out of 4 bases / nucleotides. DNA exists in cells mainly as a two-stranded structure called the double helix. Each of the bases has a complementary base.
>
> - G: Guanine => Cytosine
> - A: Adenine => Thymine
> - T: Thymine => Adenine
> - C: Cytosine => Guanine

Question: What is RNA made out of?

> RNA is a single stranded linear polymer out of 4 bases / nucleotides.
>
> - G: Guanine
> - A: Adenine
> - U: Uracil
> - C: Cytosine

Question: Do all organisms use the same amino acids / codons?

> Different organisms use the same amino acids for proteins. However, they differ in their codon usage (which RNA triplets are translated into which amino acid).

Question: How many proteins does a typical human have?

> Between 20.000 and 25.000 different kinds of proteins.

Question: What are functions of proteins?

> - Defense (e.g. antibodies)
> - Structure (e.g. collagen)
> - Enzymes (metabolism, catabolism)
> - Communication / Signaling (e.g. insulin)
> - Ligand binding / Transport (e.g. hemoglobin)
> - Storage (e.g. ferritin)

Question: How many residues long are typical proteins?

> Between 35 and 30.000 residues. The median is around 400.

Question: Do proteins consist of units?

> Proteins are built up of several domains. Most proteins have more than 2 domains.

Question: How many proteins are known?

> About 85 millions sequences are known. However, the 3D structure (experimentally determined) of only 120.000 proteins is known.

Question: Is this gap (known sequences vs known 3D structure) expected to increase?

> Yes, the gap is expected to increase. The amount of new sequences has increased drastically (far faster than Moore's Law) in the past. This is expected to continue. Advances in experimentally determining protein 3D structure could only improve marginally, but today experimentally determining the 3D structure of a proteins still costs about 100 000 EUR.

Question: How much data is produced by one sequencing machine per day?

> At full capacity about 5 - 10 terabytes of data per day.

3.3.3 Lecture 2: Introduction Protein Structure

Question: How many different amino acids are there?

> 20

Question: How do amino acids differ? What do they have in common?

> Different amino acids have different side-chains, which influence the chemical features of the respective amino acid. All of them share the same backbone.

Question: In which different feature groups can you categorize amino acids?

> polar, non-polar, acidic (negatively charged), basic (positively charged)

Question: Draw the basic chemical structure of an amino acid.

```
H   R   O
 \  |  //
  N-C-C
 /  |  \
H   H   O-H
```

Question: How are amino acids linked together to form a protein?

> In the translation process, a **Ribosome** translates a **mRNA** strand to a protein, by decoding the RNA triplets into amino acids and then linking the amino acids by peptide bonds. They chaining ALWAYS happens from the **N-Terminus** to the **C-Terminus** releasing an H20 molecule as part of the reaction.

Question: What is the definition of a 'domain'?

> A domain is a protein sequence, which when put into solvent adopts a unique 3D structure on its own.

Question: How many domains does a protein have?

> - 61% of proteins in the PDB are single domain
> - 28% of proteins in the PDB are in 62 proteomes
>
> **Problem:** This is a biased view on proteins. The 3D structure of Single-Domain-Proteins is easier to experimentally determine, so more Single-Domain-Proteins have been analyzed.

Question: Can domains overlap?

> Yes, it can happen. However, it is not what is typically observed **.**

Question: How can we compare 3D structures?

> One solution would be to align the corresponding residues of both sequences / 3D structures and take the **Root Mean Square Deviation**. (If one pair lies very far apart, it will result in an extremely low score)
>
> $RMSD(A,B) = \sum_{i=0}^{n}(r_i{}^a - r_i{}^b)^2$
>
> If the score is below a certain threshold, it is a match, otherwise it is not.

Question: How can align and compare the structure of 2 proteins?

> 1) Find the corresponding points (residues that match in 3D)
> 2) Find Superposition independent of domain movements and calculate score (e.g. RMSD)

Question: Why is global protein comparison most of the time impossible?

> The definition of protein enforces a per residue comparison (no scaling). Hence only proteins of the (almost) the same length can be compared globally. Since proteins are between 35 and 30.000 residues long, global comparison does not make sense in most of the cases.

Question: What is the difference between global and local alignment?

> In **global alignment** two structures / sequences are compared from beginning to end (compare the whole thing).
> In **local alignment** however, subunits (domains) of the proteins are aligned. (Problem: What is a valid unit? Where to cut?)

Question: How to decide what is a valid unit for local comparison of 2 proteins?

> (I couldn't identfiy a valid answer in the lecture recording)

Question: Which comparison not using cartesian RMSD could be used for comparison?

> 2D distance map: difference of differences. Only information about the chirality (mirror image) is lost.

3.3.4 Lecture 3: Alignments I

Question: Why compare 3D shapes, when we are after function? Why not compare function?

> Because ...
>
> - we cannot compare function directly
> - structure is related to function
> - we CAN compare 3D structures
> - sometimes: similar structure -> similar function

Question: How do we get protein 3D shapes?

> - primarily by experiment (most accurate)
> - computational biology (most inferences)

Question: How much does it cost to experimentally determine the 3D shape of a protein?

> Today it costs on average about 100 000 $ per protein.

Question: What are the 3 sections found in the tree of life?

> bacteria, archaea, eukaryotes

Question: What does Homology stand for?

> Here (in the context of genes), it describes proteins originating from a common ancestor. It is also frequently used to describe 'similar structure' for genes / proteins.

Question: Why do linear gap penalties not model the reality of related genes / proteins well?

> With a linear gap penalty (N gaps cost N*x) equally distributed gaps would be as expensive as clustered gaps. Biologically, gaps clustered to blocks, are however far more likely to occur, while the protein maintains similar structure / function. It is more realistic to use an **Affine gap penalty** with higher costs for opening a new gap.

Question: What is better? High sequence identity of a short (local) sequence, or worse sequence identity when matching a longer sequence? How can we decide?

> Compile the probability of randomly matching a sequence considering the background distribution. The result of this would be a substitution matrix such as BLOSUM62.

Question: Is identity the best way to match two sequences?

> Not necessarily: What we really find is similar biological function. Some amino acids might have similar biophysical features and could be swapped without any significant influence on the structure of the protein. Such matches should also be considered 'postive'.
>
> Building a scoring matrix based on evolutionary conserved residues does optimize the algorithm. (e.g. BLOSUM62)

Question: What is the biological assumption behind an insertion when comparing sequences?

> Through evolutionary changes in the DNA (e.g. a point mutation) a new bump (= amino acid(s)) was introduced. Implicitly it is also assumed similar structure -> similar function.

Question: Why do linear gap penalties not model the reality of related genes / proteins well?

> With a linear gap penalty (N gaps cost N*x) equally distributed gaps would be as expensive as clustered gaps. Biologically, gaps clustered to blocks, are however far more likely to occur, while the protein maintains similar structure / function. It is more realistic to use an**Affine gap penalty**with higher costs for opening a new gap.

Question: Does dynamic programming give the best solution?

> Yes, dynamic programming produces one optimal solution. (There could be others, though)

Question: What are issues with dynamic programming?

> - Time used: O(n^2)
>   - Especially a problem, when comparing one protein agains the entire database.
> - How to choose parameters?
>   - Gap penalties
>   - substitution matrix

Question: How can we speed up the alignment of sequences?

> 1) Hashing (fast and dirty). e.g. BLAST

Question: How does BLAST (Basic Local Alignment Search Tool) work?

> 1. Start with indexed (hashed) seeds (words of size = 3) and find matching proteins
> 2. Extend matching 'words' into both directions
> 3. Begin dynamic programming from these strong local hits

3.3.5 Lecture 4: Alignments II

Question: What is the major challenge of BLAST?

> Getting the statistics right: BLAST needs to know, h*ow significant a match is*, by comparing it against the background probability of the entire database.

Question: Why is it interesting to find similar proteins out of the Twilight / Midnight Zone?

> The Midnight-Zone is, where most proteins of similar structure sit.

Question: Why is it that even with only 40% PSI, we can still assume similar structure? Could we randomly change 60% of the residues in the lab and get a new protein with similar structure?

> - These 60% of changed residues happened under evolutionary pressure and are not random
>   - mutations that did not change structure and function survived (we can observe them

> today)
>   - mutations that did change structure and function most likely did not survive
> - Thus randomly changing 60% of residues in a proteins, would not result in a similar protein

Question: Why are certain proteins / structure multiple times in the PDB?

> - different resolution of 3D structure
> - different goals of publication produced (new) 3D structures
>   - folding sites
>   - binding partners
>   - etc ...

Question: How are profiles built up? How are the normal noted down? Do we have to know a specific algorithm?

> **Build up algorithm:**
>
> - Take all proteins of PSI over a certain threshold ...
> -
>
> **Profile Formats:**
>
> - Regular Expression
> - PSSM (Position Specific Scoring Matrix)

Question: What is a PSSM (Position Specific Scoring Matrix)?

> A matrix of numbers with scores for each residue or nucleotide at each position. Built, e.g. by PSI-BLAST.

Question: Which steps are involved in building up a profile with PSI-BLAST?

> **1) Fast Hashing**: Like BLAST, match 'word'
> **2) Dynamic Programming Extension between matches:** BLAST + Smith-Waterman
> **3) Compile Statistics:** EVAL - Expectation Values
> **4) Collect all pairs and build profile**
> **5) ... compare sequences (profile-sequence) and iterate**

Question: Why is PSI-BLAST so fast?

> Because it drastically reduces the length of the comparisons with dynamic programming.

3.3.6 Lecture 5: Comparative Modeling

Question: How do you build up a family (profile) of sequences?

> 1. Find proteins of similar sequence with BLAST
> 2. Use the found proteins to build a PSSM (profile)
> 3. Use profile-sequence alignment with the calculated PSSM to retrieve more distant family

> members
> 4. Add the newly found proteins to the family by recalculating the PSSM

Question: Which methods to experimentally determine the structure of a protein exist? How much are they used?

> Fraction of proteins in the PDB by experimental method:
>
> - 90% - X-Ray Crystallography
> - 09% - Nuclear Magnetic Resonance Spectroscopy (NMR)
> - 01 % - Electron Microscope (EM)

Question: How does X-Ray Crystallography Work

> 1. **Grow Crystal:** Force the protein to grow a crystal
> 2. **Observe Diffraction Pattern** : Shoot x-rays onto crystal and observe the diffraction pattern
> 3. **Compute Electron Density Map**
> 4. **Fit observations to atomic model**

Question: How to get 1D secondary structure from 3D coordinates?

> Two methods where used to annotate 3D coordinates:
>
> 1) DEFINE, based on geometry (not used anymore) 2) DSSP, based on hydrogen bond pattern (coulomb energy)

Question: How does Homology Modeling (Comparative Modeling) work?

> **Target:** Protein to model
> **Template**: Protein to model from
>
> 1. **Identify Template:** Query the PDB for similar sequences to your **Target**
> 2. **Align Target / Template:** Select the best match as **Template** and assume the **Target** has the same structure
> 3. **Build Model**
> 4. **Assess Model**
> 5. **Refine Model**

Question: Which tradeoff does comparative modeling face? What are the limiting factors based on PSI (Percentage Sequence Identity)?

> **Tradeoff:** Accuracy vs Coverage
>
> Limiting factor in homology modeling:
> 75% - 100% - Speed of Modeling
> 50% - 75% - Quality of Model
> 25% - 50% - Alignment Accuracy
> 0% - 25% - Detection of Homology

Question: How to handle a missing loop in comparative modeling?

> - One way would be to find similar loops and compute the average over them.
> - Another solution would be to apply molecular dynamics on the loop sequence. (only for shot loops)

3.3.7 Lecture 6: Secondary Structure Prediction 1

3.3.8 Lecture 7: Secondary Structure Prediction 2

Question: Relate the terms **Local** and **Global alignment** to the terms **Sequence-Sequence** and **Sequence-Profile**.

> Global alignments refers to aligning sequences (proteins) from start to end. Local alignments refers to only aligning parts of the sequences (e.g. 50 residues).
> Throughout Sequence-Sequence, Sequence-Profile and Profile-Profile methods both global and local alignment can be used. I practice mostly local alignment is done.

Question: What would be a simple method to predict secondary structure?

> 1) Take known structure
> 2) Find longest consecutive run of motifs that **ONLY** occur in one of the 3 states: H (Helix), E (Strand), O (Other)
> 3) Check unknown sequence against found motifs

Question: What was the first secondary structure prediction method?

> Assuming that a **Proline** would break a helix, the occurrences of proline in a sequence was used to predict helices.

Question: Where do we get the secondary structures from?

> From the DSSP, which defines 8 states in total based on H-bond patterns.

Question: What is the 1st generation of secondary structure prediction based on? What was the accuracy? Was is successful?

> - Based on single residues
> - Between 50% and 55% accuracy (Q3)
> - Clearly better than random - so it can be considered a success

Question: How did the second generation of secondary structure prediction improve? Name one algorithm.

> Instead of using only single amino acids, it would consider a sliding window of the residues around a center amino acid.**Example:**GORIII, with a Q3 accuracy of 55% - 60%

Question: What were problem of secondary structure prediction until 1994?

> - the maximum accuracy of predictions was expected to be 65%
> - β-sheet prediction was below 40%
> - many predicted segments where too short to appear in nature

Question: How can the performance of secondary structure prediction be measured?

> One way to do it, would be to calculate the **Q3** accuracy of a method against a test set. The Q3 accuracy is the **number of correctly categorized residues into one of the categories helix, strand, other divided by the total amount of residues.**

Question: How can the introduction of a new hidden layer in a neural network be described by means of a simple graph?

> Each new hidden layer basically introduces a new 'decision line' which can separate datapoints into different categories.

Question: What is cross-validation in the context of Machine Learning and why do we need it?

> Cross Validation in a method for estimating the performance of a predictive model (e.g. a neural network). To use it, the available dataset is split in 3 categories, 1) a training set, 2) a cross-training set and 3) a test set.
> 1) The training set is used to train the model
> 2) The cross-training set is used to estimate the performance of the model after x training steps
> 3) The test set is used to assess the final performance of the model after training is finished
>
> The cross-training set is needed to decide, when to stop training (when overtraining sets in) and to tweak certain parameters before running against the test-set.

Question: Did balanced training improve the Q3 prediction accuracy? Which assumption did it prove wrong?

> Balanced training actually decreased the Q3 accuracy. However, it did improve the prediction accuracy for strands significantly, falsifying the hypothesis that strands could not be predicted with local information.

Question: Which problem did PHDSec solve? How did it accomplish it?

> By introducing a **Structure-to-Structure** prediction model, PHDSec improved the prediction of too short segment. The Structure-to-Structure network would take structure (helix, strand, other) prediction of a sequence as input and predict segments based on them.

3.3.9 Lecture 8: Secondary Structure Prediction 3

Question: Which ways of comparing proteins are there? Why do we need

> - Dynamic Programming (Brute Force)
> - Hashing (e.g. BLAST)

Question: Why are fast search algorithms such as BLAST needed?

> Comparing sequences of length n residues is in $O(n^2)$. For comparing a single pair this is still fine, but comparing one (newly found) protein against all known proteins in the PDB (about 120 000) is impossible. Thus we need 'shortcuts' such as BLAST to speed up the search.

Question: What is the normal approach when you find / analyse a newly found protein?

> 1) Sequence the new protein (if not done yet)
> 2) Run BLAST against the PDB
> 3) Run Dynamic Programming against the results from BLAST

Question: In terms of CPU, is sequence-sequence as fast as sequence profile?

>

Question: How can it be that even with only 40% sequence identity we assume / observe similar structure?

> The changes in sequences we observe are not random, but follow underlying evolutionary rules. Changes, which affected the structure and thus the function of a protein are simply not likely to survive and thus we do not observe them. Changes, which did not influence the structure / function however, did survive.

Question: Why is protein sequence changing? Why are we mutating?

> - Replication Errors (point mutations)
> - Radiation
> - Viruses

Question: How much do any two unrelated typical humans differ on average?

> On average every pair of humans would differ in one amino acid per protein. (Though, changes cluster)

Question: In a structure to structure network, which additional information could be used to improve the prediction?

> - E.g. redundant information about the sequence, e.g. parts of it.

Question: When training a neural network, how do you choose the next training sample from your test set? Why so?

> Randomly, to avoid correlations

Question: How would you build up a family for a protein?

> 1. Search the PDB for proteins in comparative modeling range. (Assumption: same sequence, same 3D structure, same secondary structure)
> 2. Use profile to search in twilight-zone for potential proteins of that family (possibly verify whether the found protein is plausible to have similar 3D structure) and add to family (recompute profile)

Question: How do you get from a sequence to a secondary structure prediction with PHD?

> 1. Use BLAST to find potentially similar proteins in sequence data bank
> 2. For the resulting proteins calculate the sequence identity (homology) with dynamic programming
> 3. Filter all proteins, which are below a threshold of sequence identity (only take those "over the curve")
> 4. Extract the profile by aligning the remaining proteins
> 5. Predict the secondary structure with the sequence and its family as input

Question: Which accuracy does ProfSec achieve on average? What are additional advantages of other secondary structure prediction methods?

> ProfSec achieves a Q3 accuracy of about 72% on average. Additionally it can also predict the strength of the prediction.

Question: Does adding global information improve ProfSec prediction?

> Yes it does. While the Q3 accuracy (per residue) is not improved, the Q4 accuracy (per protein) does improve.

3.3.10 Lecture 9: Membrane Structure Prediction

Question: What are the requirements of a cell membrane?

> - separate the content of the cell from its surroundings
> - control traffic into and out of the cell
>   - keep malicious things out
>   - let good things in
> - must be a dynamic structure

Question: What are the 4 main structural components of the cell membrane?

> Carbohydrates, Cholesterol, Phospholipids, Proteins

Question: What is the cell membrane mainly made out of?

> The cell membrane is a so called **lipid bilayer** of **phospholipids**. Phospholipids have a non-polar, hydrophobic tail (membrane center) and a polar, hydrophilic head (outside of membrane).

Question: What are functions of membrane proteins?

> - help to be recognized by immune cells
> - transport proteins control substance flow in and out of the cell
> - receptor proteins bind hormons, which can change cell function
> - provide structural stability

Question: Can membrane proteins easily move around?

> It depends:
>
> - Membrane proteins can easily move laterally
> - But it is hard to move into / out of the lipid bilayer

Question: Why are there so few membrane proteins in the PDB?

> It is particularly difficult to experimentally determine the structure of membrane proteins due to the special environment they naturally occur (the membrane).

Question: What are they key questions TMH prediction tries to answer?

> - How many helices go through the membrane?
> - In which direction do they go through the membrane? (topology)

Question: Why could be a plausible reason why PHDSec failed for predicting transmembrane helices?

> Unlike 'normal' proteins, transmembrane proteins have an hydrophobic outside and a hydrophilic inside.

Question: How should we choose the threshold for the hydrophobic regions?

> 1. Predict the hydrophobicity for the protein
> 2. Assign a positive inside-out
> 3. choose the threshold to **optimize the inside out difference**

Question: What is the Positive Inside Rule and what is it used for?

> The positive Inside Rule is used to find the topology of transmembrane proteins: The loops connecting TMHs on the inside of the cell membrane have an **excess of positively charged residues**.

3.3.11 Lecture 10: TMSEG

Question: What are advantages of using a Random Forest?

> - Fast
> - robust against overtraining
> - no black box
> - Intuitive to interpret
> - good performance

3.3.12 Lecture 11: Beta Membrane and Accessibility

3.3.13 Lecture 12: Protein Disorder