title: 02: Intro into protein structure
short title: pp1_intro2_struc

lecture: Protein Prediction 1 (for Computational Biology) - Protein structure
TUM summer semester
Videos:  YouTube / www.rostlab.org

THANKS:

Special lectures:
- TBN

No lecture:
- 05/21 SVV (Student Representation)
- 05/30 Ascension Day
- 06/08-11 Whitsun holiday
- 06/11 TBA: likely no lecture
- 06/13 TBA: likely no lecture
- 06/20 Corpus Christi

LAST lecture: bef: Jul 18
             after: Jul 9

Examen:
- Makeup: NONE!

EXERCISES:

CONTACT: pp1ex@rostlab.org

OFFICIAL course web site rostlab.org

Let it go. Let it all unravel.
Let it free and it can be
A path on which to travel.
Recap
Next generation sequencing

Figure 1: Conceptual Overview of Whole-Genome Resequencing

A. Extracted gDNA.
B. gDNA is fragmented into a library of small segments that are each sequenced in parallel.
C. Individual sequence reads are reassembled by aligning to a reference genome.
D. The whole-genome sequence is derived from the consensus of aligned reads.

Illumina HiSeq 2016

Table 1: HiSeq X System Sequencing Capacity

<table>
<thead>
<tr>
<th></th>
<th>HiSeq X Ten System</th>
<th>HiSeq X Five System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum Number of Instruments</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>Annual Genome Capacity</td>
<td>&gt; 18,000</td>
<td>&gt; 9000</td>
</tr>
<tr>
<td>Price per 30x Genome</td>
<td>&lt; $1000</td>
<td>&lt; $1500</td>
</tr>
</tbody>
</table>

~2TB/run (<3 days)
Mimivirus


© Wikipedia
Reason for name

"We found a new species of ciliate during a marine field course in Rønbjerg and named it Cafeteria roenbergensis because of its voracious and indiscriminate appetite after many dinner discussions in the local cafeteria."

Tom Fenchel

© Wikipedia
Protein Prediction I - structure

1 Introduction
1.2 - Proteins/domains
1.3 - 3D comparisons
Marco Punta contributed the slides

- PhD in Trieste (MD for membrane proteins)
- Postdoc @ Columbia Univ in the City of New York (contact predictions)
- Senior scientist in NYCOMPS (Target selection for membrane proteins)
- IAS Fellow @ TUM
- Project manager @ Pfam @ Sanger Inst. Hinxtion (Cambridgeshire)
- now CNRS Paris
"Aquaria" website serves up 3D structures of proteins for researchers

Virtual journey to the innards of proteins

06.02.2015, Research news

In order to understand diseases and develop medications researchers require information on the three-dimensional structure and function of proteins. The "Aquaria" database provides quick and extensive insight into these vital compounds.

Proteins assume essential tasks in our bodies. Well-known examples of proteins are insulin, hemoglobin, fibrinogen and antibodies. The three-dimensional structure of the proteins is decisive for their functionality.

When planning experiments or producing antibodies scientists need all the information they can get on proteins. "There is a multitude of information available but "in the past, the search for protein structures was very tedious and required expert knowledge," explains Andrea Schafferhans from the Department of Bioinformatics at TUM. Together with Dr. Seán O’Donoghue from the Garvan Institute of Medical Research and CSIRO, she thus developed the "Aquaria" database. "In Aquaria all data are already processed," explains Schafferhans.

The scientists present their results in "Nature Methods." An international team of about a dozen programmers and bioinformatics experts contributed to the project, which commenced in 2009.

45 million computer models

Aquaria: simplifying discovery and insight from protein structures.

TOC today

- Previous lecture
  - Organisms, genes, central dogma

- TODAY: Protein introduction
  - Amino acids
  - Protein structure
  - Bonds & energies
  - domains

- NEXT lectures
  - 3D comparisons
  - sequence comparisons/alignments
Reality and images
Georges Braque - Houses at L'Estaque
Where is that?

Illustration by David S. Goodsell, the Scripps Research Institute, UCSD, USA
Mycoplasma genitalium

Illustration by David S. Goodsell, the Scripps Research Institute, UCSD, USA
Eukaryotic cell

Illustration by David S. Goodsell, the Scripps Research Institute, UCSD, USA
Doyle et al. (1998) Science 280:69-77 - The structure of the potassium channel: molecular basis of K+ conduction and selectivity
Alcohol dehydrogenase (ADH)

http://www.proteopedia.org/wiki/images/7/7b/1htb2.png

homodimer ADH5


http://upload.wikimedia.org/wikipedia/commons/thumb/a/a5/Protein_ADH5_PDB_1m6h.png/800px-Protein_ADH5_PDB_1m6h.png
Umberto Boccioni - Dynamism of a soccer player
Umberto Boccioni - Dynamism of a soccer player
Different levels of abstraction

(a) Diagram with labels Ser132 and Phe 34

(b) Unlabeled diagram

Wu et al. unpublished

Photograph: Filippo Monteforte/AFP/Getty Images

Umberto Boccioni - Dynamism of a soccer player

Slide: Marco Punta
Constituents of proteins: amino acids
Amino acid

side-chain

backbone

slide: Marco Punta
Joining amino acids into proteins

isolated amino acid

side-chain

backbone
Joining amino acids into proteins

a dipeptide

R

side-chain

H

N

C

H

O

backbone

side-chain

H

N

C

H

O

R

From Wikipedia
Joining amino acids into proteins

a dipeptide

side-chain

backbone

side-chain

From Wikipedia

www.webchem.net/notes/chemical_bonding/covalent_bonding.htm
Joining amino acids into proteins

polypeptide chain
Joining amino acids into proteins

- **backbone**
  - H
  - N
  - C
  - O
  - OH

- **side-chain**
  - R
Joining amino acids into proteins
Joining amino acids into proteins
Joining amino acids into proteins
Peptides

\[
\text{R}\ 	ext{COOH} + \text{H}_2\text{N} - \text{R}' \rightarrow \text{R} - \text{N} - \text{R}' + \text{H}_2\text{O}
\]

N-terminus

C-terminus

C–N double bond character in amide (peptide) bonds

planar peptide bond segments

Images: https://www2.chemistry.msu.edu/faculty/reusch/virttxtjml/protein2.htm

slide: Andrea Schafferhans
Rationalizing biophysical features of constituents
Excursion: Physical Chemistry
Covalent bond

Jacek FH@en.wikipedia
© http://www.webchem.net/notes/chemical_bonding/covalent_bonding.htm
Covalent bond – energy

[Diagram of covalent bonding]

© http://www.webchem.net/notes/chemical_bonding/covalent_bonding.htm
Double bond

Bond energy:
C-C  347 kJ/mol
C=C  611 kJ/mol

Energy barrier to rotation:
C-C  13 - 26 kJ/mol
C=C  264 kJ/mol

JoJan,Vladsinger@en.wikipedia
Polar bonds

Difference in electronegativity
(= potential to attract electrons)

→ Polar bond, dipole
Ionic interactions

\[ V_{21} = \frac{1}{4\pi \varepsilon_0} \frac{q_2}{r} \]

Coulomb potential

Non-bonded: Lenard Jones Potential

\[ V(r) = 4\epsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^{6} \right], \]

Non-covalent interactions in proteins

Covalent (electron sharing) < Hydrogen bond < Van der Waals
1 \quad 1/10 \quad 1/100-1/1000

From Wikipedia
Salt bridges

Arg, Lys, (His) -> Glu, Asp

From Wikipedia, Author: Chem5409grp6
\( \pi-\pi \) interactions


Trp, Tyr, Phe -> Trp, Tyr, Phe
Summary of typical interactions

- Hydrogen bonds
- Ionic interactions / salt bridges
- Metal complexes
- Cation - π

Back to proteins
## Amino acid structure

<table>
<thead>
<tr>
<th>Name</th>
<th>Formula</th>
<th>Abbreviations</th>
<th>Name</th>
<th>Formula</th>
<th>Abbreviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycine</td>
<td>(\text{H}_2\text{C}\text{COOH})</td>
<td>(\text{Gly} \text{ G})</td>
<td>Cysteine</td>
<td>(\text{HS}\text{C}\text{O}\text{OH})</td>
<td>(\text{Cys} \text{ C})</td>
</tr>
<tr>
<td>Alanine</td>
<td>(\text{H}_3\text{C}\text{C}\text{O}\text{OH})</td>
<td>(\text{Ala} \text{ A})</td>
<td>Methionine</td>
<td>(\text{H}_3\text{C}\text{S}\text{C}\text{O}\text{OH})</td>
<td>(\text{Met} \text{ M})</td>
</tr>
<tr>
<td>Valine</td>
<td>(\text{H}_3\text{C}\text{C}\text{C}\text{O}\text{OH})</td>
<td>(\text{Val} \text{ V})</td>
<td>Lysine</td>
<td>(\text{H}_2\text{N}\text{C}\text{C}\text{O}\text{OH})</td>
<td>(\text{Lys} \text{ K})</td>
</tr>
</tbody>
</table>

**Predominant Species at pH=6.0**

- Alanine \(p_I=6.01\)
- Isoleucine \(p_I=6.02\)

**Side-chain**

Images: [https://www2.chemistry.msu.edu/faculty/reusch/virttxtjml/proteins.htm](https://www2.chemistry.msu.edu/faculty/reusch/virttxtjml/proteins.htm)
Side chain properties
Side chain properties

- Alanine
- Cysteine
- Histidine
- Methionine
- Threonine
- Tyrosine
- Tryptophan
- Valine

- Arginine
- Asparagine
- Glutamine
- Leucine
- Lysine
- Proline
- Serine
- Glycine

Chemical structures of the amino acids are shown, indicating their side chain properties.
Negatively charged amino acids
Polar amino acids

Ala

Asp

Arg

Glu

Gly

His

Ile

Leu

Lys

Met

Phe

Pro

Ser

Thr

Trp

Tyr

Val
amino acids
Macroscopic sub-structures
Protein sequence

>gi|16128674|ref|NP_415226.1| potassium translocating ATPase, subunit A [Escherichia coli K12]
MAAQGFLLIATFLLVLMVLMVLRPLGLGLARLINDIPLPGTTGVERVLFRALGVSR
EMNWKQYLCAILGLNMLGLAVLFFMLLGQHYLPPLNPQQLPGLSWDLALNTAVSF
VTNTNWQSYLSYGFSMQAGLTVQNFLSAASGIAVIFALIRAFTRQQSMSTL
GNAWVDLLRILWVLVPVALLIALFFIQQGALQNFLPYQAVNTEGAQQLLPMG
PVASQEAIKMLGTNGGMFNANSSHPFENPTALTNFVQMLAIFLIPTALCFAFG
EVMGDRRQQGRMLLWAMSIVFICVGVVMWAEVQGNPHLLALGTDSINMEGKES
RFGVLVSSLFAUVTTAASCAGAVIAMHDSFTALGGMVPMWLMQIGEVVFGGVSGL
LYGMMLFVLLAVFIAGLMIGRTPEYLGKKIDVREMKLTALAILVTPTLVLMGAA
LAMMTDAGRSSLNPQPHGFSEVLYAVSSAANNNSAFAAGLSANSFVWNCLLAF
CMFVGRRFGVIIIPVMAIAGSLVSKSQAAASSGTLPTHGPLLFGVLIGTVLLVGA
LTFIPALALGPVAEYLS
Protein structure

Primary

Secondary

Tertiary

Quaternary
Secondary structure: \( \alpha \)-Helix

Dihedral Angles

\[
\begin{align*}
\Phi &= C-N-C^\alpha-C \\
\Psi &= N-C-C^\alpha-N
\end{align*}
\]

one turn of Ala-Thr-Gly-Ala-Phe-Leu-Ala-Phe-Ser-Ile-Gly

Images: https://www2.chemistry.msu.edu/faculty/reusch/virttxtjml/protein2.htm
Secondary structure: $\beta$-Sheet

Images: https://www2.chemistry.msu.edu/faculty/reusch/virttxtjml/protein2.htm
Spot the secondary structure
Spot the secondary structure
Spot the secondary structure
GN Ramachandran (Gopalasamudram Narayana Iyer Ramachandran)

- 8 October 1922 - 7 April 2001
- 1942: Electrical engineering -> physics (Master with Raman) / optics
- 1949: PhD: Cavendish Lab in Cambridge: crystallography
- 1952: Madras: shift to biological molecules: triple helical structure of collagen
- 1963 (JMB): analysis of backbone confirmations: Ramachandran plots
Secondary structure: Ramachandran

from: http://boscoh.com/protein/the-mysterious-regions-of-the-ramachandran-plot

Ramachandran plot (Ramachandran, 1963, JMB 7:99)

From Wikipedia
Author: Dcrjsr
Protein structure cartoons
Protein structure
Disulfide bridges

Cys -> Cys

From Wikipedia, author: Benjah-bmm27

From Wikipedia, Personal communication
“components” of protein structure: domains
Domain from introns?

**RNA splicing**

Gene product = protein

domains = introns?
Domain merger

prokaryote P, protein A

prokaryote P, protein B

prokaryote P2, protein C
3D modules
3D modules

Multiple 3D alignment identifies consensus secondary structure
Guessing domains from sequence

protein A
protein B
protein C
protein D
protein E
protein F
Guessing domains from sequence

protein A
protein B
protein C
protein D
protein E
protein F

domain 1
domain 2
Guessing domains from sequence

protein A
protein B
protein C
protein D
protein E
protein F

domain 1     domain 2
Problem: biology is not that simple
Problem: biology is not that simple
Domain length distribution

some common assumptions and facts about proteins
Facts about proteins & domains

prokas=prokaryotic proteins / eukas: eukaryotic proteins

Longer proteins are rarer than shorter ones? yes

Prokas shorter than eukas? ?
Kingdoms similar in length

## Facts about proteins & domains

prokas = prokaryotic proteins / eukas: eukaryotic proteins

<table>
<thead>
<tr>
<th>Fact</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Longer proteins are rarer than shorter ones?</td>
<td>yes</td>
</tr>
<tr>
<td>Prokas shorter than eukas?</td>
<td>yes</td>
</tr>
<tr>
<td>Some proteins are multi-domain, most are not</td>
<td>?</td>
</tr>
<tr>
<td>Most multi-domain proteins have more than 3 domains?</td>
<td>?</td>
</tr>
<tr>
<td>There are more proteins with 3 domains than with 1?</td>
<td>?</td>
</tr>
</tbody>
</table>
short insert:
CHOP - domain guessing
How to identify domains?
CHOP proteins into structural domains

CUT 1: 3D domains

CUT 2: Pfam regions

CUT 3: SWISS-PROT

Final 5 domains “today”

J Liu & Rost 2004 Proteins 55:678-686
Most proteins multi-domain

Liu, Hegyi, Acton, Montelione & Rost 2003 Proteins 56:188-200
Liu & Rost 2004 Proteins 55:678-686
Most proteins multi-domain

Single-domain proteins:
61% in PDB
28% in 62 proteomes

Liu, Hegyi, Acton, Montelione & Rost 2003 Proteins 56:188-200
Liu & Rost 2004 Proteins 55:678-686
### Facts about proteins & domains

**prokas=prokaryotic proteins / eukas: eukaryotic proteins**

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Longer proteins are rarer than shorter ones?</td>
<td>yes</td>
</tr>
<tr>
<td>Prokas shorter than eukas?</td>
<td>yes</td>
</tr>
<tr>
<td>Some proteins are multi-domain, most are not</td>
<td>no</td>
</tr>
<tr>
<td>Most multi-domain proteins have 3 or more domains?</td>
<td>yes</td>
</tr>
<tr>
<td>There are more proteins with 3 domains than with 1?</td>
<td>yes</td>
</tr>
<tr>
<td>Do prokas have more domains than eukas?</td>
<td></td>
</tr>
</tbody>
</table>
Prokaryotes and eukaryotes similar

Liu, Hegyi, Acton, Montelione & Rost 2003 *Proteins* 56:188-200
Liu & Rost 2004 *Proteins* 55:678-686
Prokaryotes and eukaryotes similar

**Average domain length**
- in proteins ≥ 2 domains: ~100 residues
- in proteins with 1 domain: 1.7-3 times longer

Liu, Hegyi, Acton, Montelione & Rost 2003 *Proteins* 56:188-200
Liu & Rost 2004 *Proteins* 55:678-686
### Facts about proteins & domains

**prokas** = prokaryotic proteins / **eukas**: eukaryotic proteins

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Longer proteins are rarer than shorter ones?</td>
<td>yes</td>
</tr>
<tr>
<td><strong>Prokas shorter than eukas?</strong></td>
<td>yes</td>
</tr>
<tr>
<td>Some proteins are multi-domain, most are not</td>
<td>no</td>
</tr>
<tr>
<td>Most multi-domain proteins have 3 or more domains?</td>
<td>yes</td>
</tr>
<tr>
<td>There are more proteins with 3 domains than with 1?</td>
<td>yes</td>
</tr>
<tr>
<td>Do prokas have more domains than eukas?</td>
<td>no</td>
</tr>
<tr>
<td>Do prokas have more proteins with very many domains eukas?</td>
<td></td>
</tr>
</tbody>
</table>
Eukaryotes have more very long proteins

# Facts about proteins & domains

**prokas** = prokaryotic proteins  /  **eukas** = eukaryotic proteins

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Longer proteins are rarer than shorter ones?</td>
<td>yes</td>
</tr>
<tr>
<td>Prokas shorter than eukas?</td>
<td>yes</td>
</tr>
<tr>
<td>Some proteins are multi-domain, most are not</td>
<td>no</td>
</tr>
<tr>
<td>Most multi-domain proteins have 3 or more domains?</td>
<td>yes</td>
</tr>
<tr>
<td>There are more proteins with 3 domains than with 1?</td>
<td>yes</td>
</tr>
<tr>
<td>Do prokas have more domains than eukas?</td>
<td>no</td>
</tr>
<tr>
<td>Do prokas have more proteins with very many domains eukas?</td>
<td>yes</td>
</tr>
</tbody>
</table>
Clustering proteins
Grouping is biology
Grouping is biology
Recap: proteins/cells
Cells: outside & inside

Illustration of Mycoplasma genitalium by David S. Goodsell, the Scripps Research Institute, UCSD, USA
A gallery of proteins
HIV-1 and a Human T-cell

HIV-1, envelope glycoprotein

CD4

T-cell surface

gp120

CCR5

slide: Natasha Wood, Cape Town

© Burkhard Rost
HIV-1 and a Human T-cell

Image: http://www.sciencemag.org/content/320/5877/760/F3.large.jpg

Slide: Natasha Wood, Cape Town
N-linked glycans

IMAGE adapted from: http://4.bp.blogspot.com/-60YQVlGcSn-c/U31O6C8YMOI/AAAAAAAANg0/8taBzk9cen8/s1600/primary-structure.png

slide: Natasha Wood, Cape Town
Protein secondary structure: helices & strands/sheets

from 3goe, structure at 0.97 Å resolution
# Lecture plan (PP1 structure/comp biol)

- **01**: 04/23 Tue: No lecture
- **02**: 04/25 Thu: No lecture
- **03**: 04/30 Tue: No lecture
- **04**: 05/02 Thu: Intro 1: organization of lecture: intro into cells & biology
- **05**: 05/07 Tue: Intro 2: amino acids, protein structure (comparison), domains
- **06**: 05/09 Thu: Alignment 1
- **07**: 05/14 Tue: Alignment 2
- **08**: 05/16 Thu: Comparative modeling & exp structure determination & secondary structure assignment
- **09**: 05/21 Tue: SKIP: Student Representation (SVV)
- **10**: 05/23 Thu: 1D: Secondary structure prediction 1
- **11**: 05/28 Tue: 1D: Secondary structure prediction 2
- **12**: 05/30 Thu: SKIP: Ascension Day
- **13**: 06/04 Tue: 1D: Secondary structure prediction 3 - Deep Learning
- **14**: 06/06 Thu: 1D: Secondary structure prediction 4
- **15**: 06/11 Tue: SKIP: Whitsun
- **16**: 06/13 Thu: No lecture (but exercises)
- **17**: 06/18 Tue: 1D: Transmembrane structure prediction 1
- **18**: 06/20 Thu: SKIP: Corpus Christi
- **19**: 06/25 Tue: 1D: Transmembrane structure prediction 2 / Solvent accessibility prediction
- **20**: 06/27 Thu: 1D: Disorder prediction; 2D prediction / 3D prediction
- **21**: 07/02 Tue: 2D: contact prediction - Deep Learning
- **22**: 07/04 Thu: Recap 1
- **23**: 07/09 Tue: Recap 2
- **24**: 07/11 Thu: TBA
- **25**: 07/16 Tue: TBA
- **26**: 07/18 Thu: TBA
- **27**: 07/23 Tue: TBA
- **28**: 07/25 Thu: TBA

**EXAM: 07/17 18:00-20:00**