Computational Biology 1 - Protein structure:

Intro into protein structure

cb1_intro2_structure

Protein Prediction 1 - Protein structure

Computational Biology 1 - TUM Summer 2016
Videos: YouTube / www.rostlab.org

THANKS:

Special lectures:
- TBD - Thomas Hopf
- TBD - Jonas Reeb

No lecture:
- 05/10 Student assembly (SVV)
- 05/17 Ascension day
- 05/26 Whitsun holiday
- 06/04 Corpus Christi

LAST lecture: June 28 (followed by 2 wrap-up sessions)

Examen: June 30, 2016: lecture time room TBD

Makeup: TBC: Oct 18 & 20, 2016 - lecture time

CONTACT: Lothar Richter richter@rostlab.org
Questions about last lecture?

© Wikipedia
Recap
Recap: genes - proteins

- common to life: DNA/cells
- DNA->RNA->proteins = machinery of life
### Central dogma as information flow

- **Dogma**

Sequential information cannot be transferred back from protein to protein, RNA or DNA.

<table>
<thead>
<tr>
<th>General</th>
<th>Special</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA → DNA</td>
<td>Replication</td>
<td>RNA → DNA</td>
</tr>
<tr>
<td>DNA → RNA</td>
<td>Transcription</td>
<td>RNA → RNA</td>
</tr>
<tr>
<td>RNA → protein</td>
<td>Translation</td>
<td>DNA → protein</td>
</tr>
</tbody>
</table>
Central dogma

- DNA Polymerase: DNA replication (DNA -> DNA)
- RNA Polymerase: transcription (DNA -> RNA)
- Ribosome: translation (RNA -> Protein)

Function

Structure

slide: Andrea Schafferhans

© Burkhard Rost

ROSTLAB.
Codon wheel
Illumina: MiSeq

- run: 6 hours
- full capacity: ~5-10 TB data / day

Illumina - Early 2012
Illumina: HiSeq 2000

- run: 1-8 days, 24 GB/day
- 1 human genome @ 30x / day
- full capacity: 25 TB data / day
- 120 x 76 x 94 cm

BGI - Shenzhen

Illumina - 2011
Recap: genes - proteins

- common to life: DNA/cells
- DNA->RNA->proteins = machinery of life
- proteins made up of amino acids (20 different: like pearl chains with pearls of 20 different sizes & “colors”, i.e. biophysical features)
- 20k proteins in human
- ~65M proteins known
- protein length (number of amino acids): 35-30K
scale reduced
Cells: outside & inside

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

- Previous lecture
  - Organisms, genes, central dogma

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

- NEXT lectures
  - domains
  - 3D comparisons
Protein Prediction I: Beginners

1 Introduction
1.2 - Proteins/domains
1.3 - 3D comparisons
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
Alcohol dehydrogenase (ADH)

homodimer ADH5


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

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

Photograph: Filippo Monteforte/AFP/Getty Images
Different levels of abstraction

(a)

(b)

Photograph: Filippo Monteforte/AFP/Getty Images

Umberto Boccioni - Dynamism of a soccer player

Wu et al. unpublished
Constituents of proteins: amino acids
Amino acid

side-chain

backbone
Joining amino acids into proteins

isolated amino acid

side-chain

backbone
Joining amino acids into proteins

A dipeptide

From Wikipedia
Joining amino acids into proteins

A dipeptide

From Wikipedia

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

A dipeptide

From Wikipedia

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

a dipeptide

backbone

side-chain

R

backbone

side-chain

H N C H OH

H N H C O OH

R

From Wikipedia

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

a dipeptide

From Wikipedia

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

a dipeptide

From Wikipedia

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

polypeptide chain
Joining amino acids into proteins

side-chain

backbone

© Burkhard Rost
ROSTLAB.
© TUM
Joining amino acids into proteins
Joining amino acids into proteins
Joining amino acids into proteins
Joining amino acids into proteins
Peptides

\[
\text{R-COOH} + \text{H}_{2}\text{N-}R' \rightarrow \text{R-NH-}R' + \text{H}_2\text{O}
\]

\[
\begin{align*}
\text{N-terminus} & : \text{R}^1 \quad \text{R}^2 \\
\text{C-terminus} & : \text{R}^3 \quad \text{R}^4
\end{align*}
\]

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
# Amino acid structure

<table>
<thead>
<tr>
<th>Name</th>
<th>Formula</th>
<th>Abbreviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycine</td>
<td>$\text{H}<em>2\text{C}\text{O}</em>{\text{H}}\text{NH}_2$</td>
<td>Gly, G</td>
</tr>
<tr>
<td>Alanine</td>
<td>$\text{H}<em>3\text{C}\text{C}\text{O}</em>{\text{H}}\text{NH}_2$</td>
<td>Ala, A</td>
</tr>
<tr>
<td>Valine</td>
<td>$\text{H}<em>3\text{C}</em>{\text{CH}}\text{C}\text{O}_{\text{H}}\text{NH}_2$</td>
<td>Val, V</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Formula</th>
<th>Abbreviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cysteine</td>
<td>$\text{HS}\text{C}<em>{\text{S}}\text{O}</em>{\text{H}}\text{NH}_2$</td>
<td>Cys, C</td>
</tr>
<tr>
<td>Methionine</td>
<td>$\text{H}<em>3\text{C}</em>{\text{S}}\text{C}<em>{\text{H}}\text{O}</em>{\text{H}}\text{NH}_2$</td>
<td>Met, M</td>
</tr>
<tr>
<td>Lysine</td>
<td>$\text{H}<em>2\text{N}</em>{\text{NH}}\text{C}<em>{\text{CH}}\text{O}</em>{\text{H}}\text{NH}_2$</td>
<td>Lys, K</td>
</tr>
</tbody>
</table>

**Predominant Species at pH=6.0**

- Alanine $\text{pI}=6.01$
- Isoleucine $\text{pI}=6.02$

---

**Images:** [https://www2.chemistry.msu.edu/faculty/reusch/virxtxtjml/proteins.htm](https://www2.chemistry.msu.edu/faculty/reusch/virxtxtjml/proteins.htm)

---

**slide:** Andrea Schafferhans

© Burkhard Rost

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

- Alanine
- Arginine
- Asparagine
- Aspartic Acid
- Cysteine
- Glycine
- Glutamine
- Glutamic Acid
- Histidine
- Isoleucine
- Leucine
- Lysine
- Methionine
- Phenylalanine
- Proline
- Serine
- Threonine
- Trytophan
- Tyrosine
- Valine
Side chain properties

- Arginine
- Asparagine
- Aspartic Acid
- Glutamine
- Glutamic Acid
- Lysine
- Methionine
- Phenylalanine
- Proline
- Serine
- Threonine
- Tryptophan
- Tyrosine
- Valine
Negatively charged amino acids

- D, E
Polar amino acids

- Alanine
- Arginine
- Asparagine
- Aspartic Acid
- Cysteine
- Glycine
- Glutamine
- Glutamic Acid
- Histidine
- Isoleucine
- Leucine
- Lysine
- Methionine
- Phenylalanine
- Proline
- Serine
- Threonine
- Tryptophan
- Tyrosine
- Valine
amino acids
“components” of protein structure: domains
Domain from introns?

RNA splicing

Gene product = protein

domais = introns?
Domain merger

prokaryote P, protein A

prokaryote P, protein B

prokaryote P2, protein C
Domains
Domains
Domains
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
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 proteins are multi-domain?</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>
Prokaryotes and eukaryotes similar

Average domain length
- in proteins $\geq 2$ 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
3D comparisons: principle idea how to?
Matching shapes

How to match?
How to match?
Differences for corresponding points
Differences for corresponding points
Differences for corresponding points
Differences for corresponding points
Differences for corresponding points
Differences for corresponding points

Difference

\[ d_1 + d_2 + d_3 + \ldots + d_8 \]

\[ = |r_{1a} - r_{1b}| + \ldots + |r_{8a} - r_{8b}| \]

RMSD (root mean square deviation)

\[ = \sqrt{ \sum (r_{iA} - r_{iB})^2 } \]

\[ RMSD(A, B) = \sqrt{ \sum_i (r_{iA}^A - r_{iB}^B)^2 } \]
Differences for corresponding points

$$RMSD(A, B) = \sqrt{\sum_i (r_i^A - r_i^B)^2}$$
Actual algorithm inverted

1st: find corresponding points
2nd: superimpose

$$RMSD(A, B) = \sqrt{\sum_i (r^A_i - r^B_i)^2}$$
fit now?
Scaling easy for simple shapes

\[ x^2 + y^2 = r^2 \]
Proteins: points are defined -> no scaling

Global vs. local comparisons
Global vs. local comparisons
Global vs. local comparisons

global solution 1:

global solution 2:
cut into “units”
cut into “units”
trouble: where to stop?

valid “unit” for comparison?
How to decide what is a valid unit?
Decision upon validity

valid “unit”
for comparison?
Valid or not?

Scientifically significant:
some expert says
How can a machine decide what is a valid unit?
Valid or not?

- **Scientifically significant:**
  some expert says

- **Statistically significant:**
  background

![Graph showing signal and background with distribution](image)
Cut, match, compare by RMSD

\[ RMSD(A, B) = \sqrt{\sum_{i} (r_i^A - r_i^B)^2} \]
Only Cartesian RMSD comparison?

\[ RMSD(A, B) = \sqrt{\sum_i (r_i^A - r_i^B)^2} \]
### 2D: difference matrix

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![Graph with numbered stars arranged in a circle]
Comparison 2D: differences of differences

Total of 8 x 8 differences
3D comparisons: biology
Structure alignment

*Slides taken from Patrice Koehl, UC Davis*

Patrice Koehl
Structure alignment: two steps

1. Identify equivalent positions (residues that match in 3D)
2. find superposition independent of domain movements

Step 1: find corresponding points in proteins A and B

\( d(i) \) are the distances between all corresponding points (typic: \( C_{\alpha} \), all atoms)

\[
\text{rmsd}(A,B) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} d_i^2}
\]
RMSD is not a metric

A similar B
B similar C
NOT implying:
A similar C

cRMSD = 2.8 Å
= 0.28 nm

cRMSD = 2.85 Å
= 0.285 nm
SSAP
3D alignment
Taylor & Orengo
SSAP concept

WR Taylor & CA Orengo (1989)
Protein structure alignment
JMB 208:1-22

\[ S_{ik} = \sum_{m=-n}^{m=+n} \frac{a}{|d_{i,i+m}^A - d_{k,k+m}^B|} + b \]
SSAP concept

WR Taylor & CA Orengo (1989)
Protein structure alignment
JMB 208:1-22

$$S_{ik} = \sum_{m=-n}^{m=+n} a + b$$

Problem: loss of information about direction
DALI
3D alignment
Holm & Sander
Notation: protein structure 1D, 2D, 3D

<table>
<thead>
<tr>
<th></th>
<th>1D</th>
<th>2D</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>128</td>
<td>110</td>
</tr>
<tr>
<td>Q</td>
<td>175</td>
<td>97</td>
</tr>
<tr>
<td>I</td>
<td>70</td>
<td>E 60</td>
</tr>
<tr>
<td>T</td>
<td>77</td>
<td>E 69</td>
</tr>
<tr>
<td>L</td>
<td>120</td>
<td>E 14</td>
</tr>
<tr>
<td>W</td>
<td>238</td>
<td>E 81</td>
</tr>
<tr>
<td>Q</td>
<td>169</td>
<td>E 97</td>
</tr>
<tr>
<td>R</td>
<td>200</td>
<td>E 62</td>
</tr>
<tr>
<td>P</td>
<td>24</td>
<td>E 48</td>
</tr>
<tr>
<td>L</td>
<td>71</td>
<td>E 59</td>
</tr>
<tr>
<td>V</td>
<td>14</td>
<td>E 0</td>
</tr>
<tr>
<td>T</td>
<td>74</td>
<td>E 69</td>
</tr>
<tr>
<td>I</td>
<td>0</td>
<td>E 0</td>
</tr>
<tr>
<td>K</td>
<td>90</td>
<td>E 73</td>
</tr>
<tr>
<td>I</td>
<td>4</td>
<td>E 0</td>
</tr>
<tr>
<td>G</td>
<td>46</td>
<td>E 41</td>
</tr>
<tr>
<td>G</td>
<td>62</td>
<td>E 53</td>
</tr>
<tr>
<td>Q</td>
<td>68</td>
<td>E 71</td>
</tr>
<tr>
<td>L</td>
<td>118</td>
<td>E 59</td>
</tr>
<tr>
<td>K</td>
<td>31</td>
<td>E 73</td>
</tr>
<tr>
<td>E</td>
<td>124</td>
<td>E 95</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>E 0</td>
</tr>
<tr>
<td>L</td>
<td>29</td>
<td>E 0</td>
</tr>
<tr>
<td>L</td>
<td>24</td>
<td>E 0</td>
</tr>
<tr>
<td>D</td>
<td>49</td>
<td>E 58</td>
</tr>
<tr>
<td>T</td>
<td>72</td>
<td>E 51</td>
</tr>
<tr>
<td>G</td>
<td>62</td>
<td>E 30</td>
</tr>
<tr>
<td>A</td>
<td>17</td>
<td>E 0</td>
</tr>
<tr>
<td>D</td>
<td>102</td>
<td>E 79</td>
</tr>
<tr>
<td>D</td>
<td>69</td>
<td>E 58</td>
</tr>
<tr>
<td>T</td>
<td>1</td>
<td>E 69</td>
</tr>
<tr>
<td>V</td>
<td>14</td>
<td>E 0</td>
</tr>
<tr>
<td>L</td>
<td>0</td>
<td>E 0</td>
</tr>
</tbody>
</table>

Notation: protein structure 1D, 2D, 3D
Structural alignment: **DALI**


- Distance matrix Alignment
- Algorithm: Monte Carlo on all-against-all for hexapeptides (5)
Vorolign
3D alignment
Birzele & Zimmer
Structural alignment: VOROLIGN

- Dynamic programming on Voronoi environments
3D comparisons: local vs. global
2 forms of calcium-bound Calmodulin

Two forms of calcium-bound Calmodulin:

Ligand free

Complexed with trifluoperazine

Global alignment:
RMSD = 15 Å / 143 residues

Local alignment:
RMSD = 0.9 Å / 62 residues

© Patrice Koehl, UC Davis
Many other 3D alignment methods exist
Recap: proteins/cells
Cells: outside & inside

Illustration of Mycoplasma genitalium by David S. Goodsell, the Scripps Research Institute, UCSD, USA
A gallery of proteins

© David Goodsell

scale reduced
HIV-1 and a Human T-cell

T-cell surface

HIV-1 envelope glycoprotein

gp120

CD4

CCR5

slide: Natasha Wood, Cape Town
HIV-1 and a Human T-cell

V3 loop

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

slide: Natasha Wood, Cape Town
3D modules

Multiple 3D alignment identifies consensus secondary structure
Lecture plan (CB1 structure)

- 01: 04/11 Tue: no lecture
- 02: 04/13 Thu: no lecture
- 03: 04/19 Tue: Organization of lecture: intro into cells & biology
- 04: 04/21 Thu: Intro I - acids/structure - domains
- 05: 04/26 Tue: Alignment 1
- 06: 04/28 Thu: Alignment 2
- 07: 05/03 Tue: Alignment 3
- 08: 05/05 Thu: SKIP: Ascension Day
- 09: 05/10 Tue: SKIP: student assembly (SVV)
- 10: 05/12 Thu: Comparative modeling
- 11: 05/17 Tue: SKIP: Whitsun holiday (05/15-17)
- 12: 05/19 Thu: Experimental structure determination / 3D -> 1D: Secondary structure assignment
- 13: 05/24 Tue: 1D: Secondary structure prediction 1
- 14: 05/26 Thu: SKIP: Corpus Christi
- 15: 06/31 Tue: 1D: Secondary structure prediction 2
- 16: 06/02 Thu: 1D: Transmembrane structure prediction 1
- 17: 06/07 Tue: 1D: Transmembrane structure prediction 2 / Solvent accessibility prediction
- 18: 06/09 Thu: 2D prediction 1
- 19: 06/14 Tue: 2D prediction 2 - Thomas Hopf
- 20: 06/16 Thu: 3D prediction / Nobel prize symposium
- 21: 06/21 Tue: 1D: Disorder prediction
- 22: 06/23 Thu: recap 1
- 23: 06/28 Tue: recap 2
- 24: 06/30 Thu: examen, no lecture
- 25: 07/05 Tue: examen alternative, no lecture
- 26: 07/07 Thu: examen, no lecture
- 27: 07/12 Tue: wrap up exercises - no lecture
- 28: 07/14 Thu: wrap up exercises - no lecture