title: Computational Biology 1 - Protein structure:

Intro into protein structure

short title: cb1_intro2_structure

lecture: Protein Prediction 1 - Protein structure

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

THANKS:

Videos:

EXERCISES:

Special lectures:
- TBN

No lecture:
- 04/26 Security check Rostlab (exercise WILL be)
- 05/01 May Day (also no exercise)
- 05/10 Ascension Day (also no exercise)
- 05/22 Whitsun holiday (also no exercise)
- 05/31 Corpus Christi (also no exercise)
- 06/21 no lecture (but exercise)

LAST lecture:
- bef: Jul 14
- after: Jul xx

Examen:
- Makeup:
  - Oct 16 (TBC)

CONTACT: pp1ex@rostlab.org

Announcements
Questions about last lecture?
Recap
Recap: genes - proteins

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

DNA Polymerase

replication
(DNA -> DNA)

RNA Polymerase

transcription
(DNA -> RNA)

Ribosome

translation
(RNA -> Protein)

Protein

Function

Structure

slide: Andrea Schafferhans
© Burkhard Rost
ROSTLAB. TUM
Codon wheel

Amino acid code

- A = Alanine
- C = Cysteine
- D = Aspartic Acid
- E = Glutamic Acid
- F = Phenylalanine
- G = Glycine
- H = Histidine
- I = Isoleucine
- K = Lysine
- L = Leucine
- M = Methionine
- N = Asparagine
- P = Proline
- Q = Glutamine
- R = Arginine
- S = Serine
- T = Threonine
- V = Valine
- W = Tryptophan
- Y = Tyrosine

© yourgenome.org

slide: Andrea Schafferhans

© Burkhard Rost
Illumina: MiSeq

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

Illumina - Early 2012
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
- ~11M protein sequences known
- protein length (number of amino acids): 35-30K
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
  - 3D comparisons

- NEXT lectures
  - sequence comparisons/alignments
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
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

slide: Marco Punta

© Burkhard Rost

ROSTLAB. TUM

21/105
Alcohol dehydrogenase (ADH)


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

homodimer ADH5
Umberto Boccioni - Dynamism of a soccer player
Umberto Boccioni - Dynamism of a soccer player
Different levels of abstraction

(a)

(b)

Umberto Boccioni - Dynamism of a soccer player

Wu et al. unpublished

Photograph: Filippo Monteforte/AFP/Getty Images
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

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](http://www.webchem.net/notes/chemical_bonding/covalent_bonding.htm)

slide: Marco Punta

© Burkhard Rost

ROSTLAB. TUM
Joining amino acids into proteins

**a dipeptide**

From Wikipedia

```latex
\begin{align*}
\text{R} & \quad \text{D} \\
\text{OH} & \quad \text{H} \\
\text{H} & \quad \text{N} \\
\text{H} & \quad \text{P} \\
\text{R} & \quad \text{R}' \\
\text{H}_2\text{O} &
\end{align*}
```

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
Joining amino acids into proteins
Joining amino acids into proteins
Joining amino acids into proteins
Rationalizing biophysical features of constituents
Side chain properties
Side chain properties
Negatively charged amino acids

- Aspartic Acid
- Glutamic Acid
Polar amino acids
amino acids
“components” of protein structure: domains
Domain from introns?

**RNA splicing**

![Diagram of RNA splicing](https://en.wikipedia.org/wiki/RNA_splicing)

© Wikipedia

Gene product = protein
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

domain 1  domain 2
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
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)
\[ = \sqrt{\sum (r_{i}^A - r_{i}^B)^2} \]
Differences for corresponding points

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

1st: find corresponding points
2nd: superimpose

\[ \text{RMSD}(A, B) = \sqrt{\sum_{i} (r_i^A - r_i^B)^2} \]
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 distribution with signal and background]
Cut, match, compare by RMSD

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

\[ RMSD(A, B) = \sqrt{\sum_i (r^A_i - r^B_i)^2} \]
# 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>

![Diagram](image)
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
Root mean square displacement (rmsd)

- **Step 1:** find corresponding points in proteins A and B
- **d(i)** are the distances between all corresponding points (typic: $C_{\text{alpha}}$, all atoms)

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

\[ cRMSD = 2.8 \text{ Å} \]
\[ = 0.28 \text{ nm} \]

\[ cRMSD = 2.85 \text{ Å} \]
\[ = 0.285 \text{ nm} \]

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

3D alignment

Holm & Sander
Notation: protein structure 1D, 2D, 3D

| P | PP P | 128 | 110 |
| Q | QQQQ | 175 | 97  |
| I | FPQVI | 70  | E 60 |
| T | SSIVR | 77  | E 69 |
| L | LLSTL | 120 | E 14 |
| W | WWQED | 238 | E 81 |
| Q | RKQAK | 169 | E 97 |
| R | RRRFQ | 200 | 62  |
| P | PPPPP | 24  | 48  |
| L | VVTKF | E 71 | E 59 |
| V | VVLII | E 14 | E 0  |
| T | TTKEK | E 74 | E 69 |
| I | AALIV | E 0  | E 0  |
| K | HYKKF | E 90 | E 73 |
| I | IILVI | E 0  | E 0  |
| G | EENG | 46  | 41  |
| G | GGGTG | 62  | 53  |
| Q | QQKRR | 68  | 71  |
| L | PPLW | E 118 | E 59 |
| K | VVFV | E 31  | E 73 |
| E | EESK | E 124 | E 95 |
| A | VVGLG | E 1  | E 0  |
| L | LLILL | E 29 | E 0  |
| L | LLLL | E 24 | E 0  |
| D | DDDD | E 49 | E 58 |
| T | TTTT | 72  | 51  |
| G | GGGGG | 62  | 30  |
| A | AAAA | 17  | 0   |
| D | DDDD | 102 | 79  |
| D | DDAGE | 69  | 58  |
| T | STTTE | 1   | 69  |
| V | IIIIV | E 14 | E 0  |
| L | VVIVL | E 0  | E 0  |
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

![Diagram showing structural alignment](image-url)
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
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
HIV-1 and a Human T-cell

HIV-1 envelope glycoprotein

HIV-1

gp120

CD4

CCR5

slide: Natasha Wood, Cape Town

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

slide: Natasha Wood, Cape Town

IMAGE: http://www.sciencemag.org/content/320/5877/760/F3.large.jpg
Multiple 3D alignment identifies consensus secondary structure
Lecture plan (CB1 structure: INF)

01: 04/10 Tue: No lecture
02: 04/12 Thu: No lecture
03: 04/17 Tue: No lecture
04: 04/19 Thu: Intro 1: organization of lecture: intro into cells & biology
05: 04/24 Tue: Intro 2: amino acids, protein structure (comparison), domains
06: 04/26 Thu: No lecture
07: 05/01 Tue: SKIP: May Day
08: 05/03 Thu: Alignment 1
09: 05/08 Tue: Alignment 2
10: 05/10 Thu: SKIP: Ascension Day
11: 05/15 Tue: Comparative modeling & exp structure determination & secondary structure assignment
12: 05/17 Thu: 1D: Secondary structure prediction 1
13: 05/22 Tue: SKIP: Whitsun holiday
14: 05/24 Thu: 1D: Secondary structure prediction 2
15: 05/29 Tue: 1D: Secondary structure prediction 3
16: 05/31 Thu: SKIP: Corpus Christi
17: 06/05 Tue: 1D: Transmembrane structure prediction 1
18: 06/07 Thu: 1D: Transmembrane structure prediction 2 / Solvent accessibility prediction
19: 06/12 Tue: 1D: Transmembrane structure prediction 3 / Solvent accessibility prediction
20: 06/14 Thu: 1D: Disorder prediction
21: 06/19 Tue: 2D prediction / 3D prediction
22: 06/21 Thu: No lecture
23: 06/26 Tue: recap 1
24: 06/28 Thu: recap 2
25: 07/03 Tue: TBA
26: 07/05 Thu: TBA
27: 07/10 Tue: TBA
28: 07/12 Thu: TBA