Protein structure determination

short title: cb1_3dexp

lecture: Protein Prediction 1 - Protein structure
Computational Biology 1
TUM summer 2014
Announcements

Videos: YouTube / www.rostlab.org

THANKS:
Tim Karl + Jonas Reeb

Special lectures:
- Apr 15 - Andrea Schafferhans

No lecture:
- Apr 17/22 Easter
- May 01 Thu May day
- May 06 Tue Student assembly
- May 29 Thu Ascension day
- Jun 03 Tue no room
- Jun 10 Tue Whitsun holidays
- Jun 19 Thu Corpus Christi

LAST lecture: July 1

Examen: July 8
- Makeup: Oct 21 - morning
Recap: 3D prediction by comparative modeling
Zones

- Midnight Zone
- Twilight Zone
- Daylight Zone

Sequence - sequence
Profile - profile
Sequence - profile
Profile - profile

sequence similar
structure similar

human - fly - bacteria

green: 3gft K-Ras - human
lime: 3lbn Rash - human
orange: 2y8e Rab6 - fly
purple: 2y8e hydroxylase

P putida

PIDE:
pairwise identical residues

Slide from:
Andrea Schafferhans

3gft: Y Tong et al. & H Park (unpublished) / 4IW3: JS Scotti (unpublished)
Comparative modeling methods

- **MODELLER**
  lots of whistles and bells, downloadable, very accurate

- **SWISS-MODEL**
  automated, increasingly comprehensive and flexible
Comparative modeling applicable to about 1/3 of all proteins
Comparative modeling 2
Goal of structure prediction

Epstein & Anfinsen, 1961: sequence uniquely determines structure

- **INPUT:** sequence
- **OUTPUT:**

3D structure and function
protein folding from first principles should then be possible
Protein structure prediction problem solved!

Problem: predict the 3D structure of a protein from sequence alone

- 60s - Washington Post
- 70s - New York Times
- 90s - Washington Post
How would you assess prediction performance?
How to get those into the prediction?
Critical Assessment of Structure Prediction

April-May (Organizers):
- collect experimental structures
  (since 2004 from structural genomics)

June-August: Prediction season
- deadline: predictions in before experimental structures are published

September-November: Assessors divine

December: Meeting to discuss results
Protein Structure Prediction

CASP

© Burkhard Rost (TUM Munich)
Protein Structure Prediction

Only homology modeling good

CASP
Protein Structure Prediction

- Only homology modeling good
- No general prediction of 3D from sequence, yet

CASP

© Burkhard Rost (Columbia New York)
© Burkhard Rost (TUM Munich)
Protein Structure Prediction

- Only homology modeling good
- No general prediction of 3D from sequence, yet
- Important improvements in many fields!

CASP

© Burkhard Rost (Columbia New York)
© Burkhard Rost (TUM Munich)
Servers, META-servers, META-META, …
CASP 9 results

- TBM
  - overall good

Problems with CASP
Problems of CASP

- Comparisons based on apples and oranges
- Analysis of irrelevant types of test cases
- Inappropriate ranking
- Conclusions based on insignificant differences
- Different categories evaluated differently
- Too few targets
Ranking not stable!

29 different worse than 11 identical

Pairwise comparison matrix
Conclusion: Comparative modeling

- Comparative modeling/homology modeling most accurate way to predict structure
- as good and as complete as the template
  - depends on quality and similarity of template
- mostly driven by accuracy of alignment
  - driven by alignment quality
- loop modeling: still not fully there, yet
- side chain modeling: unclear how well we do
3D from experiment 2 co-ordinates
3D details - 3D cartoon
Structure by experiment
Experiments determine protein structure

<table>
<thead>
<tr>
<th></th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB</td>
<td>84,413</td>
<td>1</td>
</tr>
<tr>
<td>Xray</td>
<td>75,068</td>
<td>89</td>
</tr>
<tr>
<td>NMR</td>
<td>8,723</td>
<td>10</td>
</tr>
<tr>
<td>EM</td>
<td>428</td>
<td>1</td>
</tr>
</tbody>
</table>

**PDB (Protein Data Bank)**
Helen Berman (Rutgers Univ, New Brunswick) & Phil Bourne (UCSD San Diego)
Protein structure by X-ray crystallography

<table>
<thead>
<tr>
<th>Method</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB</td>
<td>84,413</td>
<td>1</td>
</tr>
<tr>
<td>Xray</td>
<td>75,068</td>
<td>89</td>
</tr>
<tr>
<td>NMR</td>
<td>8,723</td>
<td>10</td>
</tr>
<tr>
<td>EM ElectronMic</td>
<td>428</td>
<td>1</td>
</tr>
</tbody>
</table>

**Myoglobin structure**
* THIS 1mbo: SE Philips JMB 142:531-54
(image Wikipedia Aza Toth)
(Hemoglobin: Max Perutz 1959)

© Wikipedia
Protein structure by NMR* spectroscopy

* NMR: Nuclear Magnetic Resonance

© Wikipedia
Protein structure by NMR* spectroscopy

* NMR: Nuclear Magnetic Resonance

<table>
<thead>
<tr>
<th>Method</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB</td>
<td>84,413</td>
<td>1</td>
</tr>
<tr>
<td>Xray</td>
<td>75,068</td>
<td>89</td>
</tr>
<tr>
<td>NMR</td>
<td>8,723</td>
<td>10</td>
</tr>
<tr>
<td>EM ElectronMic</td>
<td>428</td>
<td>1</td>
</tr>
</tbody>
</table>

900 MHz NMR machine

NYSBC - City College New York City


© Wikipedia
Protein structure by cryo-EM

* EM: Cryo-Electron Microscopy

<table>
<thead>
<tr>
<th></th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB</td>
<td>84,413</td>
<td>1</td>
</tr>
<tr>
<td>Xray</td>
<td>75,068</td>
<td>89</td>
</tr>
<tr>
<td>NMR</td>
<td>8,723</td>
<td>10</td>
</tr>
<tr>
<td>EM</td>
<td>428</td>
<td>1</td>
</tr>
</tbody>
</table>

4 Ångstrøm 8Å 16Å 32Å
GroEL - J Wang & DC Bosvert (2004) 1j4z

© Wikipedia
Protein structure by cryo-EM

* EM: Cryo-Electron Microscopy

<table>
<thead>
<tr>
<th></th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB</td>
<td>84,413</td>
<td>1</td>
</tr>
<tr>
<td>Xray</td>
<td>75,068</td>
<td>89</td>
</tr>
<tr>
<td>NMR</td>
<td>8,723</td>
<td>10</td>
</tr>
<tr>
<td>EM ElectronMic</td>
<td>428</td>
<td>1</td>
</tr>
</tbody>
</table>

Structure resolution

© PDB
Notation: protein structure 1D, 2D, 3D

```
P  PP P 128 110
Q  QQQY 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  4   E  0
G  EENGG  46   41
G  GGGTG  62   53
Q  QQKAR  68   71
L  PLLMW E 118  E  59
K  VVFKV E  31  E  73
E  EESKK E 124  E  95
A  VVGLG E  1   E  0
L  LLLIL E  29  E  0
L  LLLLV E 24   E  0
D  DDDDD 49   E  58
T  TTTTT  72   51
G  GGGGG  62   30
A  AAAAA  17   0
D  DDDDD 102  79
D  DDAKE  69   58
T  SSTTV  1   69
V  IVIVV E 14   E  0
L  VVIVL E  0   E  0
```

**1D**

**2D**

**3D**

---

© Burkhard Rost (TUM Munich)
Secondary structure stabilized by hydrogen bonds
Linus Pauling: introduce concept

- L Pauling, RB Corey and HR Branson (1951) The Structure of Proteins: Two Hydrogen-bonded Helical Configurations of the Polypeptide Chain PNAS 37: 205-34
- L Pauling, RB Corey and HR Branson (1951) Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain PNAS 37: 205-11
- L Pauling and RB Corey (1953) Two Rippled-sheet Configurations of Polypeptide Chains, and a Note About the Pleated Sheets PNAS 39: 253-6

Nobel Foundation:
The Nobel Prize in Chemistry 1954 was awarded to Linus Pauling "for his research into the nature of the chemical bond and its application to the elucidation of the structure of complex substances".

http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1954/
Hydrogen-bond formation

3D details - 3D cartoon
3D details - 3D cartoon
Pauling Nobel Prize 1954 – first protein structure when?
First protein structures

John Kendrew


myoglobin

Max Perutz


hemoglobin
Secondary structure assignment

Different evaluation criteria applied:

- Assignment coverage: define
  Geometry (fitting ideal sec str segments)

- Enthalpic energy: dssp
DSSP
Dictionary of Secondary Structure of Proteins
Secondary Structure Assignment: DSSP

Dictionary of protein Secondary Structure for Proteins


Chris Sander

Wolfgang Kabsch
Pauling’s H-bond pattern used in DSSP

\[ \alpha\text{-helix} \]

DSSP \[ E_{\text{HB}} < -0.5 \text{ kcal/mol} \]

L Pauling & RB Corey (1953) PNAS 39:247-252
L Pauling, RB Corey & HR Branson (1951) PNAS 37:205-234
Pauling’s H-bond pattern used in DSSP

DSSP $E_{HB} < -0.5$ kcal/mol

L Pauling & RB Corey (1953) PNAS 39:247-252
L Pauling, RB Corey & HR Branson (1951) PNAS 37:205-234
Pauling’s H-bond pattern used in DSSP

DSSP

$E_{\text{HB}} < -0.5 \text{ kcal/mol}$

L Pauling & RB Corey (1953) PNAS 39:247-252
L Pauling, RB Corey & HR Branson (1951) PNAS 37:205-234

© Burkhard Rost (TUM Munich)
DSSP assigns 7 “states”

<table>
<thead>
<tr>
<th>State</th>
<th>Description</th>
<th>7 DSSP states</th>
<th>3 state map</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>helix (i,i+4)</td>
<td>helix</td>
<td>H</td>
</tr>
<tr>
<td>G</td>
<td>3(^1)0 helix (i,i+3)</td>
<td>helix</td>
<td>H</td>
</tr>
<tr>
<td>T</td>
<td>turn of helix</td>
<td>other</td>
<td>L</td>
</tr>
<tr>
<td>E</td>
<td>extended/strand</td>
<td>strand</td>
<td>E</td>
</tr>
<tr>
<td>B</td>
<td>beta-bulge</td>
<td>strand</td>
<td>E</td>
</tr>
<tr>
<td>S</td>
<td>bend (no H-bond)</td>
<td>other</td>
<td>L</td>
</tr>
<tr>
<td>“</td>
<td>loop</td>
<td>other</td>
<td>L</td>
</tr>
</tbody>
</table>
Secondary structure assignment

Different evaluation criteria applied:

- Assignment coverage: **DEFINE**
  Geometry (fitting ideal sec str segments)

- Enthalpic energy: **DSSP**

- Expert assignment: **STRIDE**

- Predictability: **NNass**
1D: secondary structure prediction
Notation: protein structure 1D, 2D, 3D

P PP P 128 110
Q QQQY 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 RRRPQ 200  62
P PPPPP 24  48
L VVTKF E 71  E  59
V VVLI E 14  E  0
T TTKEK E 74  E  69
I AALIV E  0 E  0
K HYKKF E 90  E  73
I IILVI  4 E  0
G EENGG 46  41
G GGTTG 62  53
Q QQKAR 68  71
L PPLMW E 118 E  59
K VVFKV E 31  E  73
E EESKK E 124 E  95
A VVGLG E 1  E  0
L LLLL E 29  E  0
L LLLVV E 24  E  0
D DDDDD 49  E  58
T TTITT 72  51
G GGGGG 62  30
A AAAAA 17  0
D DDDDD 102  79
D DDAKE 69  58
T SSTTV 1  69
V IVIV E 14  E  0
L VIVVL E 0  E  0
Secondary structure prediction
2ndary structure prediction
2D prediction
Protein function classification

Protein Space:

X = Positive

Y = Negative

- Close Homology (Sequence Id. > 60%
  Psi-Blast Eval < 10^-20)
- Distant Homology (Domain, Motif)
- Machine Learning (NN, SVM)
Coverage of structure space

- Percentage proteins: 
  - Experimental 3D: 20% 
  - Comparative modelling: 80%

- Percentage residues: 
  - Experimental 3D: 20% 
  - Comparative modelling: 80%

The art of being humble
DSSP secondary assignment has 8 “states”

- H = Helix
- G = 3$_{10}$ helix
- I = Pi helix
- E = Extended (strand)
- B = beta-bridge, single strand residue
- T = Turn, i.e. one turn of helix
- S = bent
- “ “ = loop
How pentapeptides occur in 2 states?
Secondary structure prediction methods

- L Pauling, RB Corey and HR Branson (1951) Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. PNAS 37:205-211.

... some are more equal than others ...
Sec str pred methods: single residues

- Pauling, RB Corey and HR Branson (1951) Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. PNAS 37:205-211.
First step (Szent-Györgyi)
Proline breaks a helix
Helices span several turns, i.e. >4 residues
-> identify helices/non-helices
Simple prediction: frequency

- First step (Szent-Györgyi)
  Proline breaks a helix
  Helices span several turns, i.e. >4 residues
  -> identify helices/non-helices

- From Proline to odds for all ....
Simple prediction: frequency

from Proline to odds for all

.....,.....1.....,.....2....
QEKS PREVTMKKGDILTLLLNSTNK
E...E       EEEEEEE

<table>
<thead>
<tr>
<th>AA</th>
<th>D</th>
<th>E</th>
<th>G</th>
<th>I</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>P</th>
<th>Q</th>
<th>R</th>
<th>S</th>
<th>T</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>L</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Secondary structure prediction methods

- **single residues** (1. generation)
  - Chou-Fasman, GOR

how to assess performance?

problem 1: where to get secondary structure from?
how to assess performance?

problem 2: how to measure?
Secondary structure prediction accuracy

- Q3 : three-state per-residue accuracy

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

Secondary structure prediction methods

single residues (1. generation)

- Chou-Fasman, GOR
  published: 63% accuracy

Secondary Structure Assignment: DSSP

- Dictionary of protein Secondary Structure for Proteins
- ASSESSING secondary structure prediction

Secondary structure prediction methods

- single residues (1. generation) 1957-70/80
  - Chou-Fasman, GOR
  - 50-55% accuracy (assessed in 1994)

<table>
<thead>
<tr>
<th>No.</th>
<th>Date</th>
<th>Lecture Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>2014/04/08</td>
<td>Tue: sorry</td>
</tr>
<tr>
<td>02</td>
<td>2014/04/10</td>
<td>Thu: welcome: who we are</td>
</tr>
<tr>
<td>03</td>
<td>2014/04/15</td>
<td>Tue: Intro I - acids/structure (Andrea Schafferhans)</td>
</tr>
<tr>
<td>04</td>
<td>2014/04/17</td>
<td>Thu: SKIP: Easter vacation</td>
</tr>
<tr>
<td>05</td>
<td>2014/04/22</td>
<td>Tue: SKIP: Easter vacation</td>
</tr>
<tr>
<td>06</td>
<td>2014/04/24</td>
<td>Thu: Intro 2 - domains</td>
</tr>
<tr>
<td>07</td>
<td>2014/04/29</td>
<td>Tue: Intro 3 - 3D comparisons</td>
</tr>
<tr>
<td>08</td>
<td>2014/05/01</td>
<td>Thu: SKIP: “May day” - (NOT to be confused with “m’aidez”)</td>
</tr>
<tr>
<td>09</td>
<td>2014/05/06</td>
<td>Tue: SKIP: student assembly (SVV)</td>
</tr>
<tr>
<td>10</td>
<td>2014/05/08</td>
<td>Thu: Alignment 1</td>
</tr>
<tr>
<td>11</td>
<td>2014/05/13</td>
<td>Tue: Alignment 2</td>
</tr>
<tr>
<td>12</td>
<td>2014/05/15</td>
<td>Thu: Comparative modeling 1</td>
</tr>
<tr>
<td>13</td>
<td>2014/05/20</td>
<td>Tue: CM2 + Experimental structure determination + Secondary structure prediction</td>
</tr>
<tr>
<td>14</td>
<td>2014/05/22</td>
<td>Thu: Secondary structure prediction 2</td>
</tr>
<tr>
<td>15</td>
<td>2014/05/27</td>
<td>Tue: 1D: Secondary structure prediction 1</td>
</tr>
<tr>
<td>16</td>
<td>2014/05/29</td>
<td>Thu: SKIP: holiday (Ascension Day)</td>
</tr>
<tr>
<td>17</td>
<td>2014/06/03</td>
<td>Tue: SKIP: no room</td>
</tr>
<tr>
<td>18</td>
<td>2014/06/05</td>
<td>Thu: 1D: Secondary structure prediction 2</td>
</tr>
<tr>
<td>19</td>
<td>2014/06/10</td>
<td>Tue: SKIP: Whitsun holidays</td>
</tr>
<tr>
<td>20</td>
<td>2014/06/12</td>
<td>Thu: 1D: Transmembrane helix prediction</td>
</tr>
<tr>
<td>21</td>
<td>2014/06/17</td>
<td>Tue: Nobel prize symposium</td>
</tr>
<tr>
<td>22</td>
<td>2014/06/19</td>
<td>Thu: SKIP: Corpus Christi (Fronleichnam)</td>
</tr>
<tr>
<td>23</td>
<td>2014/06/24</td>
<td>Tue: 1D: Transmembrane strand prediction, solvent accessibility</td>
</tr>
<tr>
<td>24</td>
<td>2014/06/26</td>
<td>Thu: 2D prediction</td>
</tr>
<tr>
<td>25</td>
<td>2014/07/01</td>
<td>Tue: 3D prediction/wrap up</td>
</tr>
<tr>
<td>26</td>
<td>2014/07/03</td>
<td>Thu: wrap up again</td>
</tr>
<tr>
<td>27</td>
<td>2014/07/08</td>
<td>Tue: examen, no lecture</td>
</tr>
<tr>
<td>28</td>
<td>2014/07/10</td>
<td>Thu: no lecture</td>
</tr>
</tbody>
</table>