Membrane helix prediction

pp1_TMpred_1

Protein Prediction 1 - Protein structure
TUM Summer 2013
Announcements

- Videos: SciVe / www.rostlab.org

THANKS:
  - Tim Karl + ?

Special lectures:
  - May 16, 23, 28: Andrea Schafferhans
  - June 27: Thomas Hopf

No lecture:
  - May 9 Thu (Ascension)
  - May 14 Tue (Student assembly)
  - May 21 Tue (Whitsun break)
  - May 30 Thu (Corpus Christi)

LAST lecture: Jul 4

Examen: Jul 9, 13:15 (likely this room + others)
  - Makeup: Oct 17 - morning

CONTACT: Marlena Drabik assistant@rostlab.org
Today: Secondary structure prediction 1

- LAST WEEKs
  - 3D->1D: secondary structure assignment & prediction

- THIS WEEK
  - Secondary structure prediction methods - details & TMH

- NEXT WEEK
  - 1D/2D prediction contd.
Notation: protein structure 1D, 2D, 3D

<table>
<thead>
<tr>
<th>P</th>
<th>PP</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q</td>
<td>QQQ</td>
<td>Q</td>
</tr>
<tr>
<td>I</td>
<td>FPF</td>
<td>I</td>
</tr>
<tr>
<td>T</td>
<td>SSIVR</td>
<td>T</td>
</tr>
<tr>
<td>L</td>
<td>LLSTL</td>
<td>L</td>
</tr>
<tr>
<td>W</td>
<td>WWQED</td>
<td>W</td>
</tr>
<tr>
<td>Q</td>
<td>RKQAK</td>
<td>Q</td>
</tr>
<tr>
<td>R</td>
<td>RRRPQ</td>
<td>R</td>
</tr>
<tr>
<td>P</td>
<td>PPPPP</td>
<td>P</td>
</tr>
<tr>
<td>L</td>
<td>VVTKFE</td>
<td>L</td>
</tr>
<tr>
<td>V</td>
<td>VVLIE</td>
<td>V</td>
</tr>
<tr>
<td>T</td>
<td>TTKEE</td>
<td>T</td>
</tr>
<tr>
<td>I</td>
<td>AALIV</td>
<td>I</td>
</tr>
<tr>
<td>K</td>
<td>HYKKF</td>
<td>K</td>
</tr>
<tr>
<td>I</td>
<td>IILVE</td>
<td>I</td>
</tr>
<tr>
<td>G</td>
<td>EENGGE</td>
<td>G</td>
</tr>
<tr>
<td>G</td>
<td>GGGTGG</td>
<td>G</td>
</tr>
<tr>
<td>Q</td>
<td>QQKRR</td>
<td>Q</td>
</tr>
<tr>
<td>L</td>
<td>PPLWWE</td>
<td>L</td>
</tr>
<tr>
<td>K</td>
<td>VVFKVE</td>
<td>K</td>
</tr>
<tr>
<td>E</td>
<td>EESSKE</td>
<td>E</td>
</tr>
<tr>
<td>A</td>
<td>VVGOLVE</td>
<td>A</td>
</tr>
<tr>
<td>L</td>
<td>LLLLEV</td>
<td>L</td>
</tr>
<tr>
<td>L</td>
<td>LLLVVE</td>
<td>L</td>
</tr>
<tr>
<td>D</td>
<td>DDDDDD</td>
<td>D</td>
</tr>
<tr>
<td>T</td>
<td>TTTTTT</td>
<td>T</td>
</tr>
<tr>
<td>G</td>
<td>GGGGG</td>
<td>G</td>
</tr>
<tr>
<td>A</td>
<td>AAAAAA</td>
<td>A</td>
</tr>
<tr>
<td>D</td>
<td>DDDDDD</td>
<td>D</td>
</tr>
<tr>
<td>D</td>
<td>DDAKE</td>
<td>D</td>
</tr>
<tr>
<td>T</td>
<td>SSTTV</td>
<td>T</td>
</tr>
<tr>
<td>V</td>
<td>IIIVVE</td>
<td>V</td>
</tr>
<tr>
<td>L</td>
<td>VVIVL</td>
<td>L</td>
</tr>
</tbody>
</table>

© Burkhard Rost (TU Munich)
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)

© Kaz Wrzeszczynski: Thesis
1D: TM transmembrane helix prediction
TMH (Transmembrane helix) background
Cytoplasm (stromal side)  cytoplasm

1JB0  Cyanobacterial Photosystem I
Jordan P, Krauss N

1E7P  Fumarate Reductase
Lancaster CD, Michel H

Monday June 17, 2013
Quinol Fumarate Reductase
Iverson TM, Rees DC

Mechanosensitive Channel
MscL Homolog
Chang G, Rees DC

1K FY
1M S L
Transmembrane vs. membrane anchored/associated

2BCC
Cytochrome bc1
Zhang Z, Kim S

1QCR
Cytochrome bc1 complex from Bovine Heart Mitochondria
Xia D, Diesenhofer J
Membrane prediction
HTM prediction wait for db growth ...
HTM prediction wait for db growth ...
Topology for membrane helical proteins

ex tra -cy to plasm ic

protein A

out

lipid membrane bilayer

protein B

C-term

protein C

in

C-term

intra-cytoplasmic

C-term
TMH prediction
PHDsec success on Poly-Valine

HEADER LIPOPROTEIN(SURFACE FILM)
COMPND PULMONARY SURFACTANT-ASSOCIATED POLYPEPTIDE C(SP-C)
SOURCE PIG (SUS SCROFA)
AUTHOR J.JOHANSSON, T.SZYPERSKI, T.CURSTEDT, K.WUTHRICH

AA LRIPCCPVLKRLLVVVVVVLVVVVTVGALLMGL
OBS sec HHHHHHHHHHHHHHHHHHHHHHHHHHH
PHD sec EEEEEEEEEEEEEEEEEEEEEEE
How 2 separate outside/inside?
Lipid bilayer

Fully hydrated
Fully dehydrated
Intermediate
Lipid head
Lipid tail

Wikipedia
© http://en.wikipedia.org/wiki/Lipid_bilayer
Example 1: GPCR*: signaling

* G-Protein Coupled Receptor

40% of all drugs target GPCRs

Wikipedia
© http://en.wikipedia.org/wiki/Lipid_bilayer
Example 2: pumps and channels

Shaker: voltage gated potassium channel

SB Long, EB Campbell & R Mackinnon (2005)
Science 309: 897–903

Wikipedia
© http://en.wikipedia.org/wiki/Lipid_bilayer
Example 3: cell fusion

Wikipedia
© http://en.wikipedia.org/wiki/Lipid_bilayer
Lipid bilayer: hydrophobic in inside

Fully hydrated
Fully dehydrated
Intermediate
Lipid head
Lipid tail

Wikipedia
© http://en.wikipedia.org/wiki/Lipid_bilayer
Hydrophobic core of a protein

Monday June 17, 2013
Topology for membrane helical proteins

extra-cytoplasmic

protein A
out

protein B
C-term

lipid membrane bilayer

protein C
C-term

intra-cytoplasmic

C-term

in

in
Hydrophobic side chains
Eisenberg hydrophobicity index

<table>
<thead>
<tr>
<th>AA-3</th>
<th>AA-1</th>
<th>Eisenberg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ile</td>
<td>I</td>
<td>1.380</td>
</tr>
<tr>
<td>Phe</td>
<td>F</td>
<td>1.190</td>
</tr>
<tr>
<td>Val</td>
<td>V</td>
<td>1.080</td>
</tr>
<tr>
<td>Leu</td>
<td>L</td>
<td>1.060</td>
</tr>
<tr>
<td>Trp</td>
<td>W</td>
<td>0.810</td>
</tr>
<tr>
<td>Met</td>
<td>M</td>
<td>0.640</td>
</tr>
<tr>
<td>Ala</td>
<td>A</td>
<td>0.620</td>
</tr>
<tr>
<td>Gly</td>
<td>G</td>
<td>0.480</td>
</tr>
<tr>
<td>Cys</td>
<td>C</td>
<td>0.290</td>
</tr>
<tr>
<td>Tyr</td>
<td>Y</td>
<td>0.260</td>
</tr>
<tr>
<td>Pro</td>
<td>P</td>
<td>0.120</td>
</tr>
<tr>
<td>Thr</td>
<td>T</td>
<td>-0.050</td>
</tr>
<tr>
<td>Ser</td>
<td>S</td>
<td>-0.180</td>
</tr>
<tr>
<td>His</td>
<td>H</td>
<td>-0.400</td>
</tr>
<tr>
<td>Glu</td>
<td>E</td>
<td>-0.740</td>
</tr>
<tr>
<td>Asn</td>
<td>N</td>
<td>-0.780</td>
</tr>
<tr>
<td>Gln</td>
<td>Q</td>
<td>-0.850</td>
</tr>
<tr>
<td>Asp</td>
<td>D</td>
<td>-0.900</td>
</tr>
<tr>
<td>Lys</td>
<td>K</td>
<td>-1.500</td>
</tr>
<tr>
<td>Arg</td>
<td>R</td>
<td>-2.530</td>
</tr>
</tbody>
</table>

David Eisenberg, UCLA

© https://www.uclaaccess.ucla.edu/uploads/image/faculty/134.jpg

5 Hydrophobicity/tm/occupancy scales

hydrophobicity scales

Monday June 17, 2013
Many indices exist

K Tomii and M Kanehisa (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng 9:27-36: Fig. 2 (402 indices)
PHDsec success on Poly-Valine

HEADER LIPOPROTEIN(SURFACE FILM)
COMPND PULMONARY SURFACTANT-ASSOCIATED POLYPEPTIDE C(SP-C)
SOURCE PIG (SUS SCROFA)
AUTHOR J.JOHANSSON, T.SZYPERSKI, T.CURSTEDT, K.WUTHRICH

AA LRIPCCPVNLKRLLVVVVVVLVVVVTGVALLMG
OBS sec HHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
PHD sec EEEEEEEEEEEEEEEEEEEEEEEEEEE

h: hydrophobic NLKRLLLVVVVVVVLVVVVTGVALL
h hhhhhhhhhhhhhhh hh hh
Membrane protein prediction

Gunnar von Heijne
H-index: 75
7 papers cited >1,000 times
63 papers cited >100 times

Gunnar von Heijne
Stockholm University
Royal Swedish Academy of Sciences

http://www.abo.fi/public/fi/media/6765/heijne_von_gunnar_2.jpg
Membrane protein prediction

Gunnar von Heijne
H-index: 75
7 papers cited >1,000 times
63 papers cited >100 times

Gunnar von Heijne
Stockholm Univ
Royal Swedish Academy of Sciences

2009 van Deenen Medal for his outstanding contributions to the field of biomembrane research

Gunnar von Heijne announcing Chemistry Nobel Prize 2008

Monday June 17, 2013
Membrane protein prediction

Gunnar von Heijne


Gunnar von Heijne
Stockholm Univ
Royal Swedish Academy of Sciences
©http://www.abo.fi/public/fi/media/6765/heijne_von_gunnar_2.jpg
Identify hydrophobic regions

G von Heijne (1992)
Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J Mol Biol 225: 487-94: Fig. 4
Identify hydrophobic regions

G von Heijne (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J Mol Biol 225: 487-94: Fig. 1

Figure 1. Sliding window used in the hydrophobicity analysis. For a given window-position, the hydrophobicity value $h_i$ for each residue in the window is multiplied by the corresponding window-weight, $w_i$, and the sum over the window is taken. $w_i$ is given by $w_i = \{i/S$ for $1 \leq i \leq n-q+1; (n-q+1)/S$ for $(n-q+1) < i < (n+q+1); (2n+2-i)/S$ for $(n+q+1) \leq i \leq 2n+1\}$, where $S = (1+n)^2 - q^2$ is a normalization factor to make $\sum_{i=1}^{2n+1} w_i = 1$

In the calculations reported here, $n = 10$ and $q = 5$. 
Identify hydrophobic regions

G von Heijne (1992)
Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J Mol Biol 225: 487-94: Fig. 4

Figure 4. (a) Hydrophobicity plot for the SecY protein. The upper and lower cutoffs are marked. A tentative transmembrane segment with a mean hydrophobicity falling between the 2 cutoffs is marked by an arrow. (b) Two possible topologies for the SecY protein based on the hydrophobicity plot. The putative transmembrane segment is shown in black. The number of Arg+Lys residues is shown next to each polar segment. Note that the correct alternative (bottom, including the putative transmembrane segment) has a much higher charge-bias than the incorrect one.
Topology for membrane helical proteins

**Diagram:**
- **Extra-cytoplasmic**
- **Intra-cytoplasmic**
- **Lipid membrane bilayer**

**Proteins:**
- Protein A
- Protein B
- Protein C

**Terms:**
- C-term
G von Heijne (1986)
The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology.
EMBO J 5:3021-7
Fig. 2

Fig. 2. Distribution of the number of positively charged residues in periplasmic connecting loops (open squares, 54 loops in total) and cytosolic connecting loops (solid squares, 56 loops in total) in the 20 inner membrane proteins listed in Table I.
Topology for membrane helical proteins

extra-cytoplasmic

protein A

protein B

protein C

lipid membrane bilayer

in

out

intra-cytoplasmic

C-term
Heijne rule: positive inside out

Loop lengths
Charge: Number of R+K in loops 1-4

\[ \Delta = (5+1) - (2+3) > 0 \]
=> first loop out

final prediction:

lipid membrane bilayer
extra-cytoplasmic
intra-cytoplasmic

\[ \sum = 2 \]
\[ \sum = 5 \]
\[ \sum = 3 \]
\[ \sum = 1 \]
Identify hydrophobic regions

- 1. predict <H>
- 2. assign positive inside-out
- 3. choose threshold to optimize inside-out difference

G von Heijne (1992)
Membrane protein structure prediction.
Hydrophobicity analysis and the positive-inside rule. J Mol Biol 225: 487-94: Fig. 4

Figure 4. (a) Hydrophobicity plot for the SecY protein. The upper and lower cutoffs are marked. A tentative transmembrane segment with a mean hydrophobicity falling between the 2 cutoffs is marked by an arrow.
(b) Two possible topologies for the SecY protein based on the hydrophobicity plot. The putative transmembrane segment is shown in black. The number of Arg+Lys residues is shown next to each polar segment. Note that the correct alternative (bottom, including the putative transmembrane segment) has a much higher charge-bias than the incorrect one.
Hydrophobicity-based


idea: optimize hydrophobicity scale for prediction
PHDsec success on Poly-Valine

HEADER: LIPOPROTEIN(SURFACE FILM)
COMPND: PULMONARY SURFACTANT-ASSOCIATED POLYPEPTIDE C(SP-C)
SOURCE: PIG (SUS SCROFA)
AUTHOR: J. JOHANSSON, T. SZYPERSKI, T. CURSTEDT, K. WUTHRICH

AA: LRIPCPVNLKRLLVVVFVVLVVFVTVGALLMGL
OBS sec: HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
PHD sec: EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
PHDhtm

**input local in sequence**

A C L I G S V ins del cons
100 0 0 0 0 0 0 0 0 0 1.17
100 0 0 0 0 0 0 33 0 0 0.42
0 0 100 0 0 0 0 0 0 33 0.92
0 0 33 66 0 0 0 0 0 0 0.74
66 0 0 0 33 0 0 0 0 0 1.17
0 66 0 0 0 33 0 0 0 0 0.74
0 0 0 33 0 0 66 0 0 0 0.48

**input global in sequence**

percentage of each amino acid in protein
length of protein  (≤60, ≤120, ≤240, >240)
distance: centre, N-term  (≤40, ≤30, ≤20, ≤10)
distance: centre, C-term  (≤40, ≤30, ≤20, ≤10)
Dynamic programming on NN ‘energy’

residue number

© Burkhard Rost (TU Munich)
### PHDhtm refine topology prediction

<table>
<thead>
<tr>
<th>N-term</th>
<th>C-term</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.92</td>
<td>0.91</td>
</tr>
<tr>
<td>0.95</td>
<td>0.90</td>
</tr>
<tr>
<td>0.93</td>
<td>0.87</td>
</tr>
<tr>
<td>0.92</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Eight best HTM's:
- \( \mu = 0 \): 0 HTM
- \( \mu = 1 \): 1 HTM
- \( \mu = 2 \): 2 HTM
- \( \mu = 3 \): 3 HTM

**Loop lengths**
- Charge: Number of R+K in loops 1-4
- final prediction: \( \Delta = (5+1) - (2+3) > 0 \) => first loop out

- Loop 1: \( \sum = 2 \) (R+K) out
- Loop 2: \( \sum = 5 \) (R+K) extra-cytoplasmic
- Loop 3: \( \sum = 3 \) (R+K) lipid membrane bilayer
- Loop 4: \( \sum = 1 \) (R+K) intra-cytoplasmic
### PHDhtm on Poly-Valine

<table>
<thead>
<tr>
<th>HEADER</th>
<th>LIPOPROTEIN(SURFACE FILM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPND</td>
<td>PULMONARY SURFACTANT-ASSOCIATED POLYPEPTIDE C(SP-C)</td>
</tr>
<tr>
<td>SOURCE</td>
<td>PIG (SUS SCROFA)</td>
</tr>
<tr>
<td>AUTHOR</td>
<td>J.JOHANSSON, T.SZYPERSKI, T.CURSTEDT, K.WUTHRICH</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AA</th>
<th>LRIPCCPVNLKRLLVVVVVLVVVVTVGALLMGL</th>
</tr>
</thead>
<tbody>
<tr>
<td>OBS htm</td>
<td>TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT</td>
</tr>
<tr>
<td>PHD htm</td>
<td>TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT</td>
</tr>
</tbody>
</table>
Membrane helix prediction: TMHMM

A Krogh, B Larsson, G von Heijne, EL Sonnhammer (2001)
J Mol Biol 305:567-80

Erik Sonnhammer
Stockholm Univ

Gunnar von Heijne
Stockholm Univ
Royal Swedish Academy of Sciences
©http://www.abo.fi/public/fi/media/6765/heijne_von_gunnar_2.jpg

Anders Krogh
Copenhagen Univ
Membrane helix prediction: TMHMM

A Krogh, B Larsson, G von Heijne, EL Sonnhammer (2001) 305:567-80, Fig. 1

TMHMM: sketch

details: inside/outside loop

details: TM core
Membrane helix prediction: HMMTOP


Gabor E Tusnady
Inst Enzymology, Budapest

Istvan Simon
Inst Enzymology, Budapest
Membrane helix prediction: HMMTOP


---

Figure 3. Architecture of HMM used for topology prediction. States with the same transition matrices are colored in the same way: white, helix states; light gray, tail states; dark gray, loop states. Rectangular areas FL type states; hexagonal ones, NFL type states. The observation-symbol probabilities used by states are marked in each state. The structure of substrates in the case of the FL type is drawn within states. Lines and arrows show the possible transition between states or substrates.
## Prediction of membrane helices

<table>
<thead>
<tr>
<th>Method</th>
<th>$Q_{ok}$</th>
<th>$Q_2$</th>
<th>TOPO</th>
<th>Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHDpsihtm08</td>
<td>84</td>
<td>80</td>
<td>66</td>
<td>NN output post-processed by dynamic programming</td>
</tr>
<tr>
<td>HMMTOP2</td>
<td>83</td>
<td>80</td>
<td>61</td>
<td>HMM based</td>
</tr>
<tr>
<td>DAS</td>
<td>79</td>
<td>72</td>
<td>-</td>
<td>Optimized use of Hydrophobicity plots</td>
</tr>
<tr>
<td>TopPred2</td>
<td>75</td>
<td>77</td>
<td>54</td>
<td>Averages GES scale of hydrophobicity (Engelman et al. 1986)</td>
</tr>
<tr>
<td>SOSUI</td>
<td>71</td>
<td>75</td>
<td>-</td>
<td>Combines Hydrophobicity and Amphiphilicity (Hirokawa et. Al. 1998)</td>
</tr>
<tr>
<td>TMHMM1</td>
<td>71</td>
<td>80</td>
<td>45</td>
<td>Special Architecture HMM trained by Baum-Welch with decreasing Noise. (Profile-fed?)</td>
</tr>
</tbody>
</table>
TMH: re-evaluation
“We were very wrong!”
Example IS representative

<table>
<thead>
<tr>
<th>Method/Subset</th>
<th>Nprot</th>
<th>Q</th>
<th>% correct segments</th>
<th>% correct topology</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHDhtm_fil</td>
<td>131</td>
<td>94.4</td>
<td>88.5 ±3.1</td>
<td>82.4 ±3.8</td>
</tr>
<tr>
<td>PHDhtm_ref</td>
<td>131</td>
<td>93.8</td>
<td>89.3 ±3.1</td>
<td>86.3 ±3.1</td>
</tr>
<tr>
<td>PHDhtm_ref</td>
<td>83</td>
<td>93.6</td>
<td>88.0 ±3.6</td>
<td>85.5 ±4.8</td>
</tr>
<tr>
<td>Jones et al., 1994</td>
<td>83</td>
<td>93.6</td>
<td>88.0 ±3.6</td>
<td>85.5 ±4.8</td>
</tr>
<tr>
<td>Eukaryotes</td>
<td>99</td>
<td>95.8</td>
<td>93.5 ±3.2</td>
<td>90.3 ±3.2</td>
</tr>
<tr>
<td>Prokaryotes</td>
<td>33</td>
<td>85.6</td>
<td>75.8 ±9.1</td>
<td>72.7 ±9.1</td>
</tr>
</tbody>
</table>

all HTM correct: 89.3 ± 3.1
topology correct: 86.3 ± 3.1

B Rost, P Fariselli & R Casadio 1996 *Prot Science* 5, 1704-1718
### Table: Method/Subset, Nprot, Q, % correct segments, % correct topology

<table>
<thead>
<tr>
<th>Method/Subset</th>
<th>Nprot</th>
<th>Q</th>
<th>% correct segments</th>
<th>% correct topology</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHDhtm_fil</td>
<td>131</td>
<td>94.4</td>
<td>88.5 ± 3.1</td>
<td>82.4 ± 3.3</td>
</tr>
<tr>
<td>PHDhtm_ref</td>
<td>131</td>
<td>93.7</td>
<td>89.3 ± 3.1</td>
<td>86.3 ± 3.1</td>
</tr>
<tr>
<td>PHDhtm_ref</td>
<td>83</td>
<td>93.6</td>
<td>88.0 ± 3.7</td>
<td>85.5 ± 4.8</td>
</tr>
<tr>
<td>Jones et al., 1994</td>
<td>83</td>
<td>93.6</td>
<td>88.0 ± 3.7</td>
<td>85.5 ± 4.8</td>
</tr>
<tr>
<td>Eukaryotes</td>
<td>99</td>
<td>95.8</td>
<td>93.5 ± 3.2</td>
<td>90.3 ± 3.2</td>
</tr>
<tr>
<td>Prokaryotes</td>
<td>33</td>
<td>85.6</td>
<td>75.8 ± 9.1</td>
<td>72.7 ± 9.1</td>
</tr>
</tbody>
</table>

**Below the table:**

- **all HTM correct:** 89.3 ± 3.1
- **topology correct:** 86.3 ± 3.1

---

B Rost, P Fariselli & R Casadio 1996 *Prot Science* 5, 1704-1718

---

© Burkhard Rost (TU Munich)
Localization: helical membrane proteins

**Past:**
80-99% accuracy, problem solved

**Today:**
- All methods over-estimated: 65% for best
- No single method is best TMHMM, HMMTop, PHDhtm
- Hydrophobicity-based, really worse
- Most methods confuse signal peptides with membrane helices
- Eukaryotes and prokaryotes predicted equally well
Good news: most get most right

C-P Chen, A Kernytsky & B Rost 2002 Protein Science 11, 2774-91
accuracy

observed TM O1
predicted TM P1

overlap

overlap

overlap

O2
O3
P2
P3
Membrane prediction myth: accurate

<table>
<thead>
<tr>
<th>Method</th>
<th>$Q_{ok}$</th>
<th>$Q_{ob,htm}$</th>
<th>$Q_{prd,htm}$</th>
<th>TOPO</th>
<th>$Q_2$</th>
<th>$Q_{ob,2T}$</th>
<th>$Q_{prd,2T}$</th>
<th>$Q_{ob,2N}$</th>
<th>$Q_{prd,2N}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERROR</td>
<td>± 13</td>
<td>± 8</td>
<td>± 8</td>
<td>± 14</td>
<td>± 5</td>
<td>± 7</td>
<td>± 8</td>
<td>± 5</td>
<td>± 6</td>
</tr>
<tr>
<td>LOW-RES</td>
<td>70</td>
<td>94</td>
<td>96</td>
<td>63</td>
<td>82</td>
<td>70</td>
<td>90</td>
<td>92</td>
<td>71</td>
</tr>
<tr>
<td>DAS</td>
<td>52</td>
<td>94</td>
<td>93</td>
<td>70</td>
<td>46</td>
<td>96</td>
<td>98</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td>HMMTOP2</td>
<td>76</td>
<td>95</td>
<td>99</td>
<td>61</td>
<td>79</td>
<td>66</td>
<td>91</td>
<td>88</td>
<td>66</td>
</tr>
<tr>
<td>PHDhtm08</td>
<td>68</td>
<td>88</td>
<td>92</td>
<td>61</td>
<td>80</td>
<td>74</td>
<td>86</td>
<td>75</td>
<td>70</td>
</tr>
<tr>
<td>PHDhtm07</td>
<td>76</td>
<td>96</td>
<td>100</td>
<td>68</td>
<td>80</td>
<td>74</td>
<td>86</td>
<td>75</td>
<td>70</td>
</tr>
<tr>
<td>PRED-TMR</td>
<td>38</td>
<td>77</td>
<td>92</td>
<td>71</td>
<td>51</td>
<td>84</td>
<td>92</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>SOSUI</td>
<td>70</td>
<td>87</td>
<td>92</td>
<td>79</td>
<td>66</td>
<td>81</td>
<td>86</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>TMHMM1</td>
<td>69</td>
<td>85</td>
<td>92</td>
<td>56</td>
<td>81</td>
<td>65</td>
<td>84</td>
<td>90</td>
<td>68</td>
</tr>
<tr>
<td>TopPred2</td>
<td>70</td>
<td>93</td>
<td>100</td>
<td>61</td>
<td>77</td>
<td>64</td>
<td>90</td>
<td>87</td>
<td>64</td>
</tr>
<tr>
<td>WW</td>
<td>46</td>
<td>85</td>
<td>87</td>
<td>73</td>
<td>68</td>
<td>69</td>
<td>66</td>
<td>65</td>
<td></td>
</tr>
</tbody>
</table>

## Membrane prediction: best method?

<table>
<thead>
<tr>
<th>Score</th>
<th>Rank 1</th>
<th>Rank 2</th>
<th>Rank 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q&lt;sub&gt;ok&lt;/sub&gt;</td>
<td>DAS, HMMTOPv2, SOSUI, TMHMM, TopPred2</td>
<td>PHDhtm08, PHDhtm07, PRED-TMR, WW</td>
<td></td>
</tr>
<tr>
<td>Q&lt;sub&gt;htm&lt;/sub&gt;&lt;sup&gt;%obs&lt;/sup&gt;</td>
<td>DAS, HMMTOPv2</td>
<td>PHDhtm08, PHDhtm07, PRED-TMR, SOSUI, TMHMM, TopPred2, WW</td>
<td></td>
</tr>
<tr>
<td>Q&lt;sub&gt;htm&lt;/sub&gt;&lt;sup&gt;%prd&lt;/sup&gt;</td>
<td>DAS, HMMTOPv2</td>
<td>PHDhtm08, PHDhtm07, PRED-TMR, SOSUI, TMHMM, TopPred2, WW</td>
<td></td>
</tr>
<tr>
<td>TOPO</td>
<td>HMMTOPv2</td>
<td>PHDhtm08, PHDhtm07, TopPred2, WW</td>
<td></td>
</tr>
<tr>
<td>Q&lt;sub&gt;2&lt;/sub&gt;</td>
<td>HMMTOPv2, PHDhtm08, PHDhtm07, TMHMM, TopPred2</td>
<td>DAS, PRED-TMR, SOSUI, WWW</td>
<td></td>
</tr>
<tr>
<td>Q&lt;sub&gt;2T&lt;/sub&gt;&lt;sup&gt;%obs&lt;/sup&gt;</td>
<td>HMMTOPv2, PHDhtm08, PHDhtm07, SOSUI, TMHMM, WW</td>
<td>TopPred2</td>
<td></td>
</tr>
<tr>
<td>Q&lt;sub&gt;2T&lt;/sub&gt;&lt;sup&gt;%prd&lt;/sup&gt;</td>
<td>DAS, PRED-TMR</td>
<td>HMMTOPv2, TMHMM, TopPred2, WW</td>
<td></td>
</tr>
</tbody>
</table>


© Burkhard Rost (TU Munich)
Accuracy depends on number of TMH

High-resolution
Low-resolution

Percentage correctly predicted proteins

Number of transmembrane helices

C-P Chen, A Kernytsky & B Rost 2002 Protein Science 11, 2774-91
© Burkhard Rost (TU Munich)
Membrane helices NOT 17-25 residues

C-P Chen & B Rost 2002 Protein Science 11, 2766-73

© Burkhard Rost (TU Munich)
Monday June 17, 2013
Long membrane helices in 3D

1bgy: cytochrome BC complex
S Iwata, JW Lee, K Okada, JK Lee, M Iwata, B Rasmussen, TA Link, S Ramaswamy & BK Jap
1998 Science 281, 64

1eul: calcium ATPAase
C Toyoshima, M Nakasako, H Nomura & H Ogawa 2000 Nature 405, 647

C-P Chen & B Rost 2002 Protein Science 11, 2766-73
Long TMH difficult to predict

C-P Chen & B Rost 2002 *Protein Science* **11**, 2766-73
'Loops' between TMHs often very short

C-P Chen & B Rost 2002 *Protein Science* 11, 2766-73
Short loops very difficult to predict

C-P Chen & B Rost 2002 *Protein Science* 11, 2766-73
Types of mistakes for short loops

![Bar chart showing percentage of helices for different types of mistakes in predicting membrane helices with ≥ 33 residues.]

- Correct
- Incorrectly split
- Not predicted

Legend:
- Advanced methods
- Simple methods
- All
Evolution did it!

Sequence identity implies structural similarity!

Don't know region

B Rost 1999 Prot Engin 12, 85-94

Monday June 17, 2013
Less conserved -> less accurate

Two-state per-residue accuracy

HSSP-threshold

Q2-swiss
Q2-big

B Rost 2003 unpublished
Amazing PSI-BLAST

log2(H/not)-swiss
log2(H/not)-big
log2(H/not)-psi
Good conservation -> good prediction
To be or not to be (HTM)

\[ \vartheta_{\text{strict}} = 0.8 \quad \text{and} \quad \vartheta_{\text{loose}} = 0.7 \]

B Rost, P Fariselli & R Casadio 1996 Prot Science 5, 1704-1718
# Myth: really bad at distinguishing!

<table>
<thead>
<tr>
<th>Method</th>
<th>False positives</th>
<th>False negatives (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High-res</td>
<td>Low-resolution</td>
</tr>
<tr>
<td>ERROR</td>
<td>± 2</td>
<td>± 9</td>
</tr>
<tr>
<td>SOSUI</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>TMHMM1</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Wolfenden</td>
<td>2</td>
<td>39</td>
</tr>
<tr>
<td>PHDpsihtm</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>PHDhtm08</td>
<td>2</td>
<td>19</td>
</tr>
<tr>
<td>Ben-Tal</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>PRED-TMR</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>HMMTOP2</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>TopPred2</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>DAS</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>WW</td>
<td>32</td>
<td>0</td>
</tr>
<tr>
<td>GES</td>
<td>53</td>
<td>0</td>
</tr>
<tr>
<td>Eisenberg</td>
<td>66</td>
<td>0</td>
</tr>
<tr>
<td>KD</td>
<td>81</td>
<td>0</td>
</tr>
<tr>
<td>Sweet</td>
<td>84</td>
<td>0</td>
</tr>
<tr>
<td>Hopp-Woods</td>
<td>89</td>
<td>0</td>
</tr>
<tr>
<td>Heijne</td>
<td>92</td>
<td>0</td>
</tr>
</tbody>
</table>

C-P Chen, A Kernytsky & B Rost 2002 *Protein Science* 11, 2774-91
## Bad at identifying signal peptides

<table>
<thead>
<tr>
<th>Method</th>
<th>Percentage of proteins with signal peptides</th>
<th>Method</th>
<th>Percentage of proteins with signal peptides</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERROR</td>
<td>± 1</td>
<td>ERROR</td>
<td>± 1</td>
</tr>
<tr>
<td>PHDpsihtm08</td>
<td>23</td>
<td>GES</td>
<td>98</td>
</tr>
<tr>
<td>PHDhtm08</td>
<td>24</td>
<td>Eisenberg</td>
<td>99</td>
</tr>
<tr>
<td>Wolfenden</td>
<td>26</td>
<td>KD</td>
<td>99</td>
</tr>
<tr>
<td>Sweet</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TMHMM1</td>
<td>34</td>
<td>Hopp-Woods</td>
<td>99</td>
</tr>
<tr>
<td>Nakashima</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHDhtm07</td>
<td>45</td>
<td>Heijne</td>
<td>99</td>
</tr>
<tr>
<td>PRED-TMR</td>
<td>41</td>
<td>Levitt</td>
<td>99</td>
</tr>
<tr>
<td>HMMTOP2</td>
<td>48</td>
<td>Roseman</td>
<td>99</td>
</tr>
<tr>
<td>Ben-Tal</td>
<td>57</td>
<td>A-Cid</td>
<td>99</td>
</tr>
<tr>
<td>SOSUI</td>
<td>61</td>
<td>Av-Cid</td>
<td>99</td>
</tr>
<tr>
<td>Lawson</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TopPred2</td>
<td>82</td>
<td>EM</td>
<td>99</td>
</tr>
<tr>
<td>WW</td>
<td>90</td>
<td>Bull-Breese</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Bad at identifying signal peptides

<table>
<thead>
<tr>
<th>Method</th>
<th>Percentage of proteins with signal peptides</th>
<th>Method</th>
<th>Percentage of proteins with signal peptides</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERROR ± 1</td>
<td>ERROR ± 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHDpsihtm08</td>
<td>23</td>
<td>GES</td>
<td>98</td>
</tr>
<tr>
<td>PHDhtm08</td>
<td>24</td>
<td>Eisenberg</td>
<td>99</td>
</tr>
<tr>
<td>Wolfenden</td>
<td>26</td>
<td>KD</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sweet</td>
<td>99</td>
</tr>
<tr>
<td>TMHMM1</td>
<td>34</td>
<td>Hopp-Woods</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nakashima</td>
<td>99</td>
</tr>
<tr>
<td>PHDhtm07</td>
<td>45</td>
<td>Heijne</td>
<td>99</td>
</tr>
<tr>
<td>PRED-TMR</td>
<td>41</td>
<td>Levitt</td>
<td>99</td>
</tr>
<tr>
<td>HMMTOP2</td>
<td>48</td>
<td>Roseman</td>
<td>99</td>
</tr>
<tr>
<td>Ben-Tal</td>
<td>57</td>
<td>A-Cid</td>
<td>99</td>
</tr>
<tr>
<td>SOSUI</td>
<td>61</td>
<td>Av-Cid</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lawson</td>
<td>99</td>
</tr>
<tr>
<td>TopPred2</td>
<td>82</td>
<td>EM</td>
<td>99</td>
</tr>
<tr>
<td>WW</td>
<td>90</td>
<td>Bull-Breese</td>
<td>99</td>
</tr>
</tbody>
</table>
Other problems unravelled by recent structures
Inventory of life: membrane proteins

Eukaryotes

Prokaryotes

Archaea

Kingdoms similar in length

Kingdoms similar in length

Kingdoms similar in amino acids usage

Family sizes similar

Cumulative percentage of proteins

Prokaryotes

Eukaryotes

Number of proteins in family

Monday June 17, 2013
Inventory of life: membrane proteins

Eukaryotes

Prokaryotes

Archaea


Monday June 17, 2013
More TM -> more complexity?

Membrane proteins: kingdoms invented different tricks


Monday June 17, 2013
Membrane LEGO


Monday June 17, 2013
Length of globular regions in membrane proteins

Length of globular regions in membrane proteins

Aeropyrum pernix K1
Archaeoglobus fulgidus

Bacillus subtilis
Escherichia coli

Caenorhabditis elegans
Drosophila melanogaster

Percentage of globular regions

© Burkhard Rost (TU Munich)
Inventory of life: coiled-coil proteins

%mem

%coils

Eukaryotes

Prokaryotes

Archaeans


© Burkhard Rost (TU Munich)
Inventory of life: compartments


© Burkhard Rost (TU Munich)
TMH proteins: reminders

- statistics for PDB in June 2010:
  - 67,086 structures in PDB (June 2010)
  - 1,197 transmembrane
    - 1,014 alpha helical
    - 179 beta barrel

- -> < 2%
- BUT: >20% of all proteins!

GE Tusnady, ZS Dosztanyi & I Simon (2005)
Bioinformatics 21:1276-7
statistics for PDB in June 2010:
67,086 structures in PDB (June 2010)
246 unique* transmembrane

-> < way less than 2%
BUT: >20% of all proteins!

* unique=non-identical sequence (can have PIDE>99.5%)!

S Jayasinghe, K Hristova, SH White (2001)
Protein Sci 10:455-8
TMH proteins: reminders

- Edda Kloppmann & Marco Punta:
  - 1,035 PDB unique TM structures (Jan 2011)
  - 107 Pfam families

Thanks to Arne Elofsson

Following slides taken from Arne Elofsson, Stockholm Univ
Interface helices (Granseth, JMB 2005)

- 78 interface helices
- ~50% of chains contain interface helix
- Average length ~ 9 aa
- Longest is 19 aa
- Most frequent in photosynthetic reaction center

Reentry regions

- 36 reentrant helices
  - 20 in new classification
- 24% contain reentry
- 72% on the outside
- Length 3-32 residues
- Loops 11-117 residues

36 reentry regions in 3 classes

Helix-coil/Coil-helix

Helix-coil-helix

Coil

© Arne Elofsson (Stockholm Univ)
Predict re-entry regions

H Viklund, E Granseth & A Elofsson (2006) J Mol Biol 361:591-603: Fig. 5
# Rentry predicted in entire genomes

<table>
<thead>
<tr>
<th>Genome</th>
<th>Proteins</th>
<th>Reentrant fraction</th>
<th>Reentrants out</th>
<th>Reentrants in</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>H. sapiens</em></td>
<td>4181</td>
<td>0.15</td>
<td>0.46</td>
<td>0.54</td>
</tr>
<tr>
<td><em>S. cerevisiae</em></td>
<td>757</td>
<td>0.10</td>
<td>0.60</td>
<td>0.40</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>773</td>
<td>0.167</td>
<td>0.48</td>
<td>0.52</td>
</tr>
<tr>
<td>Observed in dataset</td>
<td>79</td>
<td>0.240</td>
<td>0.72</td>
<td>0.28</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fraction</th>
<th>Signal receptors</th>
<th>Electron transporters</th>
<th>Active transporters</th>
<th>Channels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraction</td>
<td>0.07</td>
<td>0.11</td>
<td>0.22</td>
<td>0.31</td>
</tr>
</tbody>
</table>

H Viklund, E Granseth & A Elofsson  
Membrane protein structures are complex
- TM-helices ends at different locations
- Different angles
- Neighboring helices often interact
- Interface helices
- Reentrant regions

No sheets close to the membrane

More complex structures need new prediction methods

The Z-coordinate

Z-coordinate: distance residue 2 membrane center

H Viklund, E Granseth & A Elofsson
# Lecture plan (PP1: Structure) - now

<table>
<thead>
<tr>
<th>Date</th>
<th>Lecture Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>01/04/16</td>
<td>Tue: welcome; who we are</td>
</tr>
<tr>
<td>02/04/18</td>
<td>Thu: no room, no lecture</td>
</tr>
<tr>
<td>03/04/23</td>
<td>Tue: intro I - acids/structure</td>
</tr>
<tr>
<td>04/04/25</td>
<td>Thu: INSERT - Machine learning in biology (cross-validation asf)</td>
</tr>
<tr>
<td>05/04/30</td>
<td>Tue: intro IIa - 3D comparisons</td>
</tr>
<tr>
<td>06/05/02</td>
<td>Thu: CANCELED alignment 1</td>
</tr>
<tr>
<td>07/05/07</td>
<td>Tue: alignment 1</td>
</tr>
<tr>
<td>08/05/09</td>
<td>Thu: holiday (Ascension Day)</td>
</tr>
<tr>
<td>09/05/14</td>
<td>Tue: no lecture; student assembly</td>
</tr>
<tr>
<td>10/05/16</td>
<td>Thu: no lecture</td>
</tr>
<tr>
<td>11/05/21</td>
<td>Tue: Whitsun holiday</td>
</tr>
<tr>
<td>12/05/23</td>
<td>Thu: alignment 2</td>
</tr>
<tr>
<td>13/05/28</td>
<td>Tue: alignment 3</td>
</tr>
<tr>
<td>14/05/30</td>
<td>Thu: Fronleichnam=Corpus Christi</td>
</tr>
<tr>
<td>15/06/04</td>
<td>Tue: comparative modeling 1</td>
</tr>
<tr>
<td>16/06/06</td>
<td>Thu: comparative modeling 2</td>
</tr>
<tr>
<td>17/06/11</td>
<td>Tue: 3D structure determination</td>
</tr>
<tr>
<td>18/06/13</td>
<td>Thu: 3D-&gt;1D: sec str - sec str pred 1 (chalk)</td>
</tr>
<tr>
<td>19/06/18</td>
<td>Tue: secondary structure prediction 2</td>
</tr>
<tr>
<td>20/06/20</td>
<td>Thu: transmembrane helix prediction</td>
</tr>
<tr>
<td>21/06/25</td>
<td>Tue: transmembrane strand, solvent accessibility, 2D prediction</td>
</tr>
<tr>
<td>22/06/27</td>
<td>Thu: 2D prediction - Thomas Hopf</td>
</tr>
<tr>
<td>23/07/02</td>
<td>Tue: 3D prediction</td>
</tr>
<tr>
<td>24/07/04</td>
<td>Thu: summary; what we do in our group</td>
</tr>
<tr>
<td>25/07/09</td>
<td>Tue: examen, no lecture</td>
</tr>
<tr>
<td>26/07/11</td>
<td>Thu: no lecture</td>
</tr>
<tr>
<td>27/07/16</td>
<td>no lecture</td>
</tr>
<tr>
<td>28/07/18</td>
<td>no lecture</td>
</tr>
</tbody>
</table>
Lecture plan (PP1: Structure)-generic

01: 2013/04/16 Tue: welcome: who we are
02: 2013/04/18 Thu: no room, no lecture
03: 2013/04/23 Tue: intro I - acids/structure
04: 2013/04/25 Thu: INSERT - Machine learning in biology (cross-validation asf)
05: 2013/04/30 Tue: intro IIa - 3D comparisons
06: 2013/05/02 Thu: CANCELED alignment 1
07: 2013/05/07 Tue: alignment 1
08: 2013/05/09 Thu: holiday (Ascension Day)
09: 2013/05/14 Tue: no lecture: student assembly
10: 2013/05/16 Thu: UNK: Andrea Schafferhans: Comparative Modeling 1
11: 2013/05/21 Tue: Whitsun holiday
12: 2013/05/23 Thu: Andrea Schafferhans: Comparative Modeling 2
13: 2013/05/28 Tue: alignment 2
14: 2013/05/30 Thu: Fronleichnam=Corpus Christi
15: 2013/06/04 Tue: 3D->1D: sec str
16: 2013/06/06 Thu: sec str pred 1 (white board)
17: 2013/06/11 Tue: sec str pred 2
18: 2013/06/13 Thu: sec str pred 3
19: 2013/06/18 Tue: transmembrane helix prediction
20: 2013/06/20 Thu: transmembrane strand prediction, solvent accessibility
21: 2013/06/25 Tue: 2D prediction, intro
22: 2013/06/27 Thu: 2D prediction - Thomas Hopf
23: 2013/07/02 Tue: 3D prediction
24: 2013/07/04 Thu: summary: what we do in our group
25: 2013/07/09 Tue: examen, no lecture
26: 2013/07/11 Thu: no lecture
27: 2013/07/16: no lecture
28: 2013/07/18: no lecture
Models and reality

René Margritte (1889-1967)
Models and reality

“A Model must be wrong, in some respects, else it would be the thing itself. The trick is to see where it is right.” (Henry A. Bent)

René Margritte (1889-1967)
“A Model must be wrong, in some respects, else it would be the thing itself. The trick is to see where it is right.” (Henry A. Bent)

“A model is a tool that helps to interpret biochemical data.” (Thorsten Schwede)
Rostlab & friends @ ISMB/ECCB Vienna

www.rostlab.org

© Burkhard Rost (TU Munich)
How to answer the question?
comp sci:
continue here
comp bio: continue here
this needs work!!
PUT in slide: idea of how to align