title: Computational Biology 1 - Protein structure:

Beta membrane, accessibility, 2D

short title: cb1_tmb_acc_2d

lecture: Protein Prediction 1 - Protein structure
Computational Biology 1 - TUM Summer 2015
CONTACT: Inga Weise assistant@rostlab.org

Announcements

Videos: YouTube / www.rostlab.org

THANKS:

Tim Karl + Carlo Di Domenico

Special lectures:

- 06/11 Jonas Reeb (TMH prediction)
- 06/18 Thomas Hopf (contacts)

No lecture:

- 05/12 Student assembly (SVV)
- 05/14 Ascension day
- 05/26 Whitsun holiday
- 06/04 Corpus Christi

LAST lecture: Jul 7

Examen: Jul 9

- Makeup: Oct 13, 2015 - morning/noon
Assessing performance of TMH predictions

JONAS REEB
which methods to use?
Which methods to evaluate?

☐ All is not a feasible answer:

• many hydrophobicity scales exist
• more than 30 advanced methods
• consensus/meta methods

☐ how to choose? which ones?
Focus on some important methods

- Getting methods to run is not trivial, e.g., old compilers/libraries, differing shells/operating systems.
- More methods make analysis more challenging, e.g., common data sets, display of results.
Important methods are those that are:

- recently published
- being used
- interesting/unusual
Random TMH prediction to have perspective

randomly predict residue as TMH or not
Random TMH prediction to have perspective

☐ randomly predict residue as TMH or not
☐ predict segments (TMH) of fixed lengths at random positions
Random TMH prediction to have perspective

☐ randomly predict residue as TMH or not
☐ predict segments (TMH) of fixed lengths at random positions
☐ adjust segment length according to distribution in database
Random TMH prediction to have perspective

- randomly predict residue as TMH or not
- predict segments (TMH) of fixed lengths at random positions
- adjust segment length according to distribution in database
- adjust segment length AND number of helices to reflect database distribution (dubbed RANDOM in the following)
Random TMH prediction to have perspective

- randomly predict residue as TMH or not
- predict segments (TMH) of fixed lengths at random positions
- adjust segment length according to distribution in database
- adjust segment length AND number of helices to reflect database distribution (dubbed RANDOM in the following)
- place known number of helices at random positions with random lengths (dubbed SEMIRANDOM in following)
Random TMH prediction to have perspective

- randomly predict residue as TMH or not
- predict segments (TMH) of fixed lengths at random positions
- adjust segment length according to distribution in database
- adjust segment length AND number of helices to reflect database distribution (dubbed RANDOM in the following)
- place known number of helices at random positions with random lengths (dubbed SEMIRANDOM in following)
- HYDRO: straightforward thresholding for hydrophobicity
Data set

- mind you: few structures known!

- low-resolution: biochemical labelling
  - many of those, but “low-resolution”
  - not much better than prediction…
High-resolution data set

Just use 3D structures from the PDB!
All problems solved?

Bacteriorhodopsin
(PDB id 1c3wA)

3D structures also not “perfect”

- we need high-enough resolution
- 3D structures do not show the membrane layer
- proteins are dynamic: they move
Annotate transmembrane helices

Bacteriorhodopsin (PDB id 1c3wA)

Fumarate Reductase
CD Lancaster & H Michel

1E7P
TMH annotation from 3D

- two databases annotating TMH from 3D:
  - OPM: Orientation of Proteins in the Membrane
    Lomize et al NAR (2012)
  - PDBTM: Protein Data Bank of Transmembrane Proteins
    Kozma et al NAR (2013)

- they differ->
  we used both and considered predictions correct that mapped to either of the two
High-resolution data set + annotation

Just use 3D structures from the PDB!
All problems solved?

Bacteriorhodopsin (PDB id 1c3waA)
Performance per-residue

\[ Q_{2T}^{\%\text{obs}} = \frac{TP}{TP + FN} \]

Recall

\[ Q_{2T}^{\%\text{pred}} = \frac{TP}{TP + FP} \]

Precision
Performance per-residue

\[
Q_{2T}^{\%_{\text{obs}}} = \frac{TP}{TP + FN}
\]
Recall

\[
Q_{2T}^{\%_{\text{pred}}} = \frac{TP}{TP + FP}
\]
Precision
Per-residue similar for methods

Q_2T^obs

Q_2T^pred

TopPred2
PHDhtm
HMMTOP2
TMHMM2
SOSUI
Phobius
PolyPhobius
MEMSAT3
Philius
SCAMPI
SPOCTOPUS
MEMSAT–SVM
RANDOM
SEMIRANDOM
HYDRO

Jonas Reeb & Burkhard Rost

Per-residue comparisons: issues

- methods look very similar
- not all predictions make sense, e.g.
  
  . . . . , . . . . 1 . . . . , . . . . 2
  . . . . . . HHHHHHH . . . . . . H H . . .

- not fully supported by resolution
  e.g. 10 TMH, each by +/-1 residue off-
  10 x 2 x 2 = 40 mistakes in a protein of typically 400
  residues => 10%
Performance per-segment

\[ Q_{\text{tmh}}^{\% \text{obs}} = \frac{\#\text{correctly predicted TMHs}}{\#\text{observed TMHs}} \]

Recall

\[ Q_{\text{tmh}}^{\% \text{pred}} = \frac{\#\text{correctly predicted TMHs}}{\#\text{predicted TMHs}} \]

Precision

\[ Q_{\text{ok}} = \frac{100}{N_{\text{prot}}} \times \sum_{i=0}^{N_{\text{prot}}} \delta_i \]

\% correct proteins

\[ \delta_i = \begin{cases} 1, & \text{if } Q_{\text{tmh}}^{\% \text{obs}} = 1 = Q_{\text{tmh}}^{\% \text{pred}} \\ 0, & \text{else} \end{cases} \]
Performance per-segment: correctly predicted?

- TMH observed
- TMH predicted
- allowed range

- Less than 5 residues deviation on each side
- Only one-to-one matches between predicted and observed TMHs
Per-segment: most mostly work

Prediction is generally successful for the full set of 95 proteins.
Per-segment: performance drops for new proteins

![Bar chart showing performance drops for new proteins.

- **New** proteins
  - 44 proteins
  - TopPred2
  - PHDhtm
  - HMMTOP2
  - TMHMM2
  - SOSUI
  - Phobius
  - PolyPhobius
  - MEMSAT3
  - Philius
  - SCAMPI
  - SPOCTOPUS
  - MEMSAT–SVM
  - RANDOM
  - SEMIRANDOM
  - HYDRO

- **Old** proteins
  - 146 proteins

Q_{ok} vs. New/Old proteins.
Per-segment: performance varies between kingdoms

![Graph showing performance varies between protein origins](image)
Per-segment: **more TMH-> lower performance**

![Graph showing the relationship between the number of TMHs per protein and performance](image)

- 92 (184) proteins (helices) with 1 TMH
- 60 (371) proteins (helices) with 2-5 TMHs
- 33 (570) proteins (helices) with >5 TMHs

### Performance across methods:

- TOPPRED 2
- PHDHTM
- HMMTOP 2
- TMHMM 2.0
- SOSUI
- PHOBIUS
- POLYPHOBIUS
- MEMSAT3
- PHILIUS
- SCAMPI
- SPOCTOPUS
- MEMSAT SVM
- RANDOM
- SEMIRANDOM
- HYDRO

Per-segment: >90% predictions right within +/- 1 TMH

Number of protein annotations

-6  -4  -2  0  2  4
#TMHs predicted – #TMHs observed

-  0  50  100  150  200  250  300
Number of protein annotations

-  TopPred2
-  PHDhtm
-  HMMTOP2
-  TMHMM2
-  SOSUI
-  Phobius
-  PolyPhobius
-  MEMSAT3
-  Philius
-  SCAMPI
-  SPOCTOPUS
-  MEMSAT–SVM
Per-protein: **Good identification of TMH proteins**

<table>
<thead>
<tr>
<th>Eukaryotes</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>TopPred2</td>
<td>100</td>
</tr>
<tr>
<td>PHDhtm</td>
<td>97</td>
</tr>
<tr>
<td>HMMTOP2</td>
<td>96</td>
</tr>
<tr>
<td>TMHMM2</td>
<td>97</td>
</tr>
<tr>
<td>SOSUI</td>
<td>97</td>
</tr>
<tr>
<td>Phobius</td>
<td>99</td>
</tr>
<tr>
<td>PolyPhobius</td>
<td>97</td>
</tr>
<tr>
<td>MEMSAT3</td>
<td>100</td>
</tr>
<tr>
<td>Philius</td>
<td>93</td>
</tr>
<tr>
<td>SCAMPI</td>
<td>100</td>
</tr>
<tr>
<td>SPOCTOPUS</td>
<td>97</td>
</tr>
<tr>
<td>MEMSAT-SVM</td>
<td>99</td>
</tr>
</tbody>
</table>
### Per-protein: Good distinction TMH / not-TMH

<table>
<thead>
<tr>
<th></th>
<th>Euk FPR</th>
<th>Gram- FPR</th>
<th>Gram+ FPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>TopPred2</td>
<td>62</td>
<td>53</td>
<td>52</td>
</tr>
<tr>
<td>PHDhtm</td>
<td>11</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>HMMTOP2</td>
<td>21</td>
<td>17</td>
<td>15</td>
</tr>
<tr>
<td>TMHMM2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SOSUI</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Phobius</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>PolyPhobius</td>
<td>6</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>MEMSAT3</td>
<td>8</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Philius</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SCAMPI</td>
<td>30</td>
<td>27</td>
<td>21</td>
</tr>
<tr>
<td>SPOCTOPUS</td>
<td>11</td>
<td>24</td>
<td>25</td>
</tr>
<tr>
<td>MEMSAT-SVM</td>
<td>6</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>
**Per-protein: Distinction worse for signal peptides**

<table>
<thead>
<tr>
<th>Method</th>
<th>Euk FPR</th>
<th>Gram- FPR</th>
<th>Gram+ FPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>TopPred2</td>
<td>97</td>
<td>99.8</td>
<td>98</td>
</tr>
<tr>
<td>PHDhtm</td>
<td>31</td>
<td>17</td>
<td>39</td>
</tr>
<tr>
<td>HMMTOP2</td>
<td>56</td>
<td>77</td>
<td>89</td>
</tr>
<tr>
<td>TMHMM2</td>
<td>20</td>
<td>25</td>
<td>63</td>
</tr>
<tr>
<td>SOSUI</td>
<td>62</td>
<td>32</td>
<td>44</td>
</tr>
<tr>
<td>Phobius</td>
<td>4</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>PolyPhobius</td>
<td>5</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>MEMSAT3</td>
<td>60</td>
<td>47</td>
<td>67</td>
</tr>
<tr>
<td>Philius</td>
<td>3</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>SCAMPI</td>
<td>92</td>
<td>95</td>
<td>97</td>
</tr>
<tr>
<td>SPOCTOPUS</td>
<td>13</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>MEMSAT-SVM</td>
<td>25</td>
<td>6</td>
<td>24</td>
</tr>
</tbody>
</table>
Per-segment: may be improvement possible?

\[ Q_{\text{top}} = (Q_{\text{ok}} = 1 \text{ and correct inside/outside topology}) \]

Bernhofer et al., unpublished
Recap
Lipid bilayer

Wikipedia

© http://en.wikipedia.org/wiki/Lipid_bilayer
Hydrophobic core of a protein
Membrane prediction
TMH proteins: reminders

Edda Kloppmann & Marco Punta:

1,035 PDB unique TM structures (Jan 2012) -> 107 Pfam families

i.e. although about 15-25% of all proteins AND 60% of all drug targets

only about 2% of all “unique” structures have membrane helices

... but, prediction methods for 1D very successful!
TMH prediction 1D+

TopPred
G von Heijne (1992)
JMB 225: 487-94

PHDhtm/PROFhtm
B Rost (1996) Methods
Enzymol 266:525-39

TMHMM
A Krogh, B Larsson, G von Heijne, EL Sonnhammer
(2001) JMB 305:567-80
Eight best HTM's

\[ \mu = 0: \text{0 HTM} \]
\[ \mu = 1: \text{1 HTM} \]
\[ \mu = 2: \text{2 HTM} \]
\[ \mu = 3: \text{3 HTM} \]

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

final prediction:
\[ \Delta = (5+1) - (2+3) > 0 \]
\[ \Rightarrow \text{first loop out} \]
1D: TM-beta
Beta-barrel predictions
Predicting transmembrane beta barrels

Extracellular

Sucrose Specific Porin
Maltoporin
OmpF Matrix Porin

FhuA receptor
FepA active transporter
porin from R. Blastica

Phospholipase A
OmpX
OmpA
porin from R. Capsulatis

Computer Simulation of the Rough Lipopolysaccharide Membrane of Pseudomonas aeruginosa

Biophys J, August 2001, p. 1037-1046, Vol. 81, No. 2

H Bigelow, D Petrey, J Liu, D Przybylski & B Rost (2004) NAR 32, 2566
# TMB proteins 2004: structures & functions

<table>
<thead>
<tr>
<th>Pdb</th>
<th>Ref.</th>
<th>Name</th>
<th>Source</th>
<th>#str</th>
<th>N</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a0s</td>
<td>NSB 5(1) pp. 37-46 Jan 98</td>
<td>Sucrose Specific Porin</td>
<td>S. Typh</td>
<td>18</td>
<td>3</td>
<td>rapid transport of sucrose</td>
</tr>
<tr>
<td>1af6</td>
<td>JMB 272 pp. 56-63 '97</td>
<td>Maltoporin</td>
<td>E. Coli</td>
<td>18</td>
<td>3</td>
<td>selective transport of glucose</td>
</tr>
<tr>
<td>1bt9</td>
<td>Biochem 37 pp. 15663 '98</td>
<td>OmpF Matrix Porin</td>
<td>E. Coli</td>
<td>16</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>1by5</td>
<td>Cell 95 pp. 771 '98</td>
<td>FhuA receptor</td>
<td>E. Coli</td>
<td>22</td>
<td>1</td>
<td>import of ferrichrome, signal transduction</td>
</tr>
<tr>
<td>1fep</td>
<td>NSB 6 pp 56 '99</td>
<td>FepA active transporter</td>
<td>E. Coli</td>
<td>22</td>
<td>1</td>
<td>import of ferrichrome</td>
</tr>
<tr>
<td>1prn</td>
<td>JMB 243 pp. 891 '94</td>
<td>'porin'</td>
<td>R. Blastica</td>
<td>16</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>1qd5</td>
<td>Nature 401 pp. 717 '99</td>
<td>phospholipase A</td>
<td>E. Coli</td>
<td>12</td>
<td>1</td>
<td>deacylates LPS, perforating foreign membranes</td>
</tr>
<tr>
<td>1qj9</td>
<td>Structure 7 pp. 1301 '99</td>
<td>OmpX</td>
<td>E. Coli</td>
<td>8</td>
<td>1</td>
<td>binds β-strand-containing foreign proteins</td>
</tr>
<tr>
<td>1qjp</td>
<td>JMB 298 pp 273 '00</td>
<td>OmpA membrane domain</td>
<td>E. Coli</td>
<td>8</td>
<td>1</td>
<td>docking points for bacteriophage</td>
</tr>
<tr>
<td>3por</td>
<td>JMB 231 pp. 817 '93</td>
<td>'porin'</td>
<td>R. Capsulatis</td>
<td>16</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

*all outer-membrane*
PROFtmb: Structure-based labels

“loop” out

“loop” in

Henry Bigelow & BR Columbia Univ

Legend:
- I: periplasmic hairpins
- O: extracellular loops
- U[A-Z]: upward strand, facing inward
- U[a-z]: upward strand, facing outward
- D[a-z]: downward strand, facing inward
- D[A-Z]: downward strand, facing towards bilayer

© Burkhard Rost

48/63
PROFtmb: Model design

Arrows denote allowed transitions in the HMM. Dotted arrow/region indicates one connection per enclosed state.
model seems to make sense, but is it “right”?
Answer the question in groups

How to assess whether model makes sense?

© Wikipedia
Amino Acid Statistics

Bilayer

Pore

Bilayer
put observation into the priors for the HMM, train for all others
## TMB proteins 2004: structures & functions

<table>
<thead>
<tr>
<th>Pdb</th>
<th>Ref.</th>
<th>Name</th>
<th>Source</th>
<th>#str N</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a0s</td>
<td>NSB 5(1) pp. 37-46 Jan 98</td>
<td>Sucrose Specific Porin</td>
<td>S. Typh</td>
<td>18</td>
<td>3 rapid transport of sucrose</td>
</tr>
<tr>
<td>1af6</td>
<td>JMB 272 pp. 56-63 '97</td>
<td>Maltoporin</td>
<td>E. Coli</td>
<td>18</td>
<td>3 selective transport of glucose</td>
</tr>
<tr>
<td>1bt9</td>
<td>Biochem 37 pp. 15663 '98</td>
<td>OmpF Matrix Porin</td>
<td>E. Coli</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>1by5</td>
<td>Cell 95 pp. 771 '98</td>
<td>FhuA receptor</td>
<td>E. Coli</td>
<td>22</td>
<td>1 import of ferrichrome, signal transduction</td>
</tr>
<tr>
<td>1fep</td>
<td>NSB 6 pp. 56 '99</td>
<td>FepA active transporter</td>
<td>E. Coli</td>
<td>22</td>
<td>1 import of ferrichrome</td>
</tr>
<tr>
<td>1prn</td>
<td>JMB 243 pp. 891 '94</td>
<td>'porin'</td>
<td>R. Blastica</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>1qd5</td>
<td>Nature 401 pp. 717 '99</td>
<td>phospholipase A</td>
<td>E. Coli</td>
<td>12</td>
<td>1 deacylates LPS, perforating foreign membranes</td>
</tr>
<tr>
<td>1q9</td>
<td>Structure 7 pp. 1301 '99</td>
<td>OmpX</td>
<td>E. Coli</td>
<td>8</td>
<td>1 binds β-strand-containing foreign proteins</td>
</tr>
<tr>
<td>1qjp</td>
<td>JMB 298 pp. 273 '00</td>
<td>OmpA membrane domain</td>
<td>E. Coli</td>
<td>8</td>
<td>1 docking points for bacteriophage</td>
</tr>
<tr>
<td>3por</td>
<td>JMB 231 pp. 817 '93</td>
<td>'porin'</td>
<td>R. Capsulatis</td>
<td>16</td>
<td>3</td>
</tr>
</tbody>
</table>
## PROFtmb: per-residue results

<table>
<thead>
<tr>
<th>Method</th>
<th>Data</th>
<th>$Q_2$</th>
<th>$Q_{β}^{% prd}$</th>
<th>$Q_{β}^{% obs}$</th>
<th>C</th>
<th>SOV$_2$</th>
<th>SOV$_β$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated $σ$</td>
<td></td>
<td>±6</td>
<td>±9</td>
<td>±14</td>
<td>±0.11</td>
<td>±13</td>
<td>±11</td>
</tr>
<tr>
<td>Martelli</td>
<td>SetTMBComp</td>
<td>84</td>
<td>87</td>
<td>80</td>
<td>0.69</td>
<td>91</td>
<td>94</td>
</tr>
<tr>
<td>PROFtmb</td>
<td>SetTMBComp</td>
<td>83</td>
<td>87</td>
<td>80</td>
<td>0.69</td>
<td>79</td>
<td>93</td>
</tr>
<tr>
<td>PROFtmb</td>
<td>SetTMBComp</td>
<td>83</td>
<td>84</td>
<td>85</td>
<td>0.70</td>
<td>87</td>
<td>92</td>
</tr>
<tr>
<td>PROFtmb</td>
<td>SetTMBComp</td>
<td>86</td>
<td>86</td>
<td>85</td>
<td>0.74</td>
<td>88</td>
<td>94</td>
</tr>
</tbody>
</table>

### Predicted vs Observed Counts

<table>
<thead>
<tr>
<th></th>
<th>up-strand</th>
<th>down-strand</th>
<th>peri-loop</th>
<th>outer-loop</th>
<th>SUM</th>
<th>Pok</th>
</tr>
</thead>
<tbody>
<tr>
<td>up-strand</td>
<td>563</td>
<td>9</td>
<td>27</td>
<td>57</td>
<td>656</td>
<td>86</td>
</tr>
<tr>
<td>down-strand</td>
<td>5</td>
<td>594</td>
<td>58</td>
<td>33</td>
<td>690</td>
<td>86</td>
</tr>
<tr>
<td>peri-loop</td>
<td>39</td>
<td>20</td>
<td>763</td>
<td>13</td>
<td>835</td>
<td>91</td>
</tr>
<tr>
<td>outer-loop</td>
<td>77</td>
<td>104</td>
<td>2</td>
<td>790</td>
<td>973</td>
<td>81</td>
</tr>
<tr>
<td>SUM</td>
<td>684</td>
<td>727</td>
<td>850</td>
<td>893</td>
<td>3154</td>
<td>86</td>
</tr>
<tr>
<td>Pok</td>
<td>82</td>
<td>82</td>
<td>90</td>
<td>88</td>
<td>86</td>
<td>86</td>
</tr>
</tbody>
</table>
Remarkable performance, but: can we distinguish proteins with and without TMB?
Testing 2: per protein performance

Whole Protein Discrimination
set_SWISS_locexp

Fraction Accuracy (or Coverage)

Bits Score

Accuracy
Coverage

H Bigelow, D Petrey, J Liu, D Przybylski & B Rost (2004) NAR 32, 2566
Testing 2: per protein performance

Where to put threshold?
Testing 2: per protein performance

Whole Protein Discrimination
set_SWISS_locexp

Fraction Accuracy (or Coverage)

Bits Score

Accuracy
Coverage

H Bigelow, D Petrey, J Liu, D Przybylski & B Rost (2004) NAR 32, 2566
One question: how well method does it do for proteomes?

Another: what DOES it do for proteomes?
Apply PROFtmb to predict entire organisms
Apply PROFtmb to predict entire organisms

Species

<table>
<thead>
<tr>
<th>Species</th>
<th>low threshold</th>
<th>high threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species 1</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Species 2</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>Species 3</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>
Apply PROFtmb to predict entire organisms

- **Species 1**
  - High threshold: 4
  - Low threshold: 8

- **Species 2**
  - High threshold: 8
  - Low threshold: 8

- **Species 3**
  - High threshold: 4
  - Low threshold: 8

Hypothetical data
Apply PROFtmb to predict entire organisms

- high threshold
- low threshold

![Graph showing hypothetical data for species 1, 2, and 3. The x-axis represents Bits Score, and the y-axis shows Fraction Accuracy or Coverage. The graph compares Accuracy and Coverage for different threshold levels.](image)
Apply PROFtmb to predict entire organisms

What else is missing?
In order to gauge **METHOD**
we need to know what it is "up against"
Apply PROFtmb to predict entire organisms

How much is annotated today?

- by experiment &
- by homology
Whole organism view of TMB
Whole organism view of TMB

Gram-Negative

H Bigelow, D Petrey, J Liu, D Przybylski & B Rost (2004) NAR 32, 2566
Whole organism view of TMB

H Bigelow, D Petrey, J Liu, D Przybylski & B Rost (2004) NAR 32, 2566
Whole organism view/prediction of TMB

Gram-negative (2 membranes)
Gram-positive (1 membrane)

H Bigelow, D Petrey, J Liu, D Przybylski & B Rost (2004) NAR 32, 2566
Whole organism view/prediction of TMB

Gram-negative (2 membranes)

Gram-positive (1 membrane)

H Bigelow, D Petrey, J Liu, D Przybylski & B Rost (2004) NAR 32, 2566
1D: solvent accessibility
Get accessibility from 3D structures
Defining residue solvent accessibility
Different ways to cast value into accessibility

- **absolute accessibility**
  
  $ASA = \text{square } \text{Ångstrøm} \ (1\text{Å}=0.1\text{nm})$
Joining amino acids into proteins
Joining amino acids into proteins

polypeptide chain
Joining amino acids into proteins
Some residues have short side-chains, others long ones.
Some residues have short side-chains, others long ones
Some residues have short side-chains, others long ones

A = short side chain

K = long side chain
Long side chains may appear more accessible
Different ways to cast value into accessibility

- **Absolute accessibility**
  \[ \text{ASA} = \text{square Ångstrøm} \ (1\text{Å} = 0.1\text{nm}) \]

- **Relative accessibility**
  \[ \frac{\text{ASA}}{\text{max ASA}} \]

100% = \text{max ASA}
Different ways to cast value into accessibility

- absolute accessibility
  \[ \text{ASA} = \text{square Ångström} \ (1\text{Å}=0.1\text{nm}) \]
- relative accessibility
  \[ \frac{\text{ASA}}{\text{max ASA}} \]
- “states”
  - buried, exposed
  - buried, intermediate, exposed
- what is best?
Different ways to cast value into accessibility

- **absolute accessibility**
  \[ \text{ASA} = \text{square Ångstrøm} \ (1Å=0.1\text{nm}) \]

- **relative accessibility**
  \[ \frac{\text{ASA}}{\text{max ASA}} \]

- **“states”**
  - buried, exposed
  - buried, intermediate, exposed

what is best??
100% and 80% more similar than 20 and 0
100% and 80% more similar than 20 and 0

how to reflect this in the design of a prediction method?
Simple function realizing objective

“States” to predict

Percentage solvent accessibility
Predict solvent accessibility
historically: by hydrophobicity
**PHDace**

Local alignment

- **AAA**
- **AA.**
- **LLL**
- **LII**
- **AAG**
- **CCS**
- **GWW**

Global statistics

- **%AA**
- **Length**
- **ΔN-term**
- **ΔC-term**

### Input Local in Sequence

```
ACLIGSV ins del cons
```

```
100 0 0 0 0 0 0 0 0 1.17
100 0 0 0 0 0 0 33 0 0.42
0 0 100 0 0 0 0 0 33 0.92
0 0 33 66 0 0 0 0 0 0.74
66 0 0 0 33 0 0 0 0 1.17
0 66 0 0 0 33 0 0 0 0.74
0 0 0 33 0 0 66 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)
Correct prediction of correctly predicted residues

![Graph](image)

- PHDse
- PHDac
- PHDhtm

Overall per-residue accuracy vs. percentage of residues predicted.
Evolution for accessibility prediction

- Detailed prediction problematic

- Significant gain by evolutionary information:
  
in/out with > 75% accuracy!
Challenges to predict accessibility

- Training max to 80%
  why?
Challenges to predict accessibility

- Training max to 80%: data inconsistent
- Evolutionary conservation much lower than of secondary structure
Conservation of solvent accessibility

Challenges to predict accessibility

- Training max to 80%: data inconsistent
- Evolutionary conservation much lower than of secondary structure
Challenges to predict accessibility

☐ Training max to 80%: data inconsistent
☐ Evolutionary conservation much lower than of secondary structure
☐ May be:
   redefine according to what describes accessibility and IS conserved?
What to get from predictions?
Residue conservation on surface

ConSurf

Cavities

Cavity: but NOT conserved
Conserved cavities: different substrates
Accessibility helps in predicting protein function

☐ Sub-cellular localization
☐ Protein-protein interactions
☐ Flexibility/motion from structure
More globular - more likely expressed?

proteins → fragments

domain

chain

decreasing 'globularity'
Proteins are amazingly cubic ...
Lecture plan (CB1 structure)

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