Beta membrane, accessibility, 2D

pp1_tmb_acc_2d

Protein Prediction 1 - Protein structure
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 22 Thu short lecture
• May 29 Thu Ascension day
• Jun 03 Tue no lecture
• Jun 10 Tue Whitsun holidays
• Jun 19 Thu Corpus Christi

LAST lecture: July 1

Examen: July 8 - this room
• Makeup: Oct 21 - morning
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. Biastica

Phospholipase A
OmpX
OmpA
porin from R. Capsulatiae

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
## 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>import of ferrichrome, signal transduction</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</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>deacylates LPS, perforating foreign membranes</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>binds β-strand-containing foreign proteins</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</td>
<td>docking points for bacteriophage</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></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>
PROFtmb: Structure-based labels

“loop” out

“loop” in

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 (TUM Munich)
Arrows denote allowed transitions in the HMM. Dotted arrow/region indicates one connection per enclosed state.
Amino Acid Statistics

Bilayer

Pore

© Burkhard Rost (TUM Munich)
## PROFtmb: Per-residue results

<table>
<thead>
<tr>
<th>Method</th>
<th>Data</th>
<th>(Q_2)</th>
<th>(Q^%_{prb})</th>
<th>(Q^%_{obs})</th>
<th>C</th>
<th>SOV_2</th>
<th>SOV_(\beta)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated (\sigma)</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>SetTMB</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>SetTMBfull</td>
<td>86</td>
<td>86</td>
<td>85</td>
<td>0.74</td>
<td>88</td>
<td>94</td>
</tr>
</tbody>
</table>

### AUC confusion matrix

<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>82</td>
<td>82</td>
<td>90</td>
<td>88</td>
<td>3154</td>
<td>86</td>
</tr>
</tbody>
</table>

H Bigelow, D Petrey, J Liu, D Przybylski & B Rost (2004) *NAR* 32, 2566
Testing 2

Whole Protein Discrimination
set_SWISS_locexp

Fraction Accuracy (or Coverage)

Accuray
Coverage

Bits Score

H Bigelow, D Petrey, J Liu, D Przybylski & B Rost (2004) NAR 32, 2566
Whole Genome Results

H Bigelow, D Petrey, J Liu, D Przybylski & B Rost (2004) NAR 32, 2566
Whole Genome Results

H Bigelow, D Petrey, J Liu, D Przybylski & B Rost (2004) NAR 32, 2566
Whole Genome Results

H Bigelow, D Petrey, J Liu, D Przybylski & B Rost (2004) *NAR* 32, 2566
1D: solvent accessibility
Defining residue solvent accessibility
How to predict ASA

- **absolute accessibility**
  \[ \text{ASA} = \text{square Ångstrøm} \]

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

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

- **what is best?**
PHDace

- Local alignment
- 13 adjacent residues
- Global statistic
- Whole protein

### Input local in sequence

```
<table>
<thead>
<tr>
<th>A</th>
<th>C</th>
<th>L</th>
<th>I</th>
<th>G</th>
<th>S</th>
<th>V</th>
<th>ins</th>
<th>del</th>
<th>cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1.17</td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>33</td>
<td>0</td>
<td>0.42</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>33</td>
<td>0.92</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>33</td>
<td>66</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>0</td>
<td>0</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.17</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>66</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>66</td>
<td>0</td>
<td>0.48</td>
<td></td>
</tr>
</tbody>
</table>
```

### 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)

© Burkhard Rost (TUM Munich)
Evolution for accessibility prediction

- Detailed prediction problematic

- Significant gain by evolutionary information:
  in/out with > 75% accuracy!
Prediction of protein function

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

- Proteins
- Decreasing 'globularity'
- Fragments

Domain → Chain → Decreasing 'globularity'
Proteins are amazingly cubic ...
Short yeast ORFs
Ooi number
number of C-alpha atoms surrounding a residue within a sphere of 14 Å
K Nishikawa & T Ooi (1986) J Biochem 100:1043-7
“The correlation between the experimental and computed quantities is as high as 0.50 on the average over 92 proteins of known 3D structure” (Abstract)
Optimize choice of 1D feature by prediction success  
best: Ooi number on C-beta within 14Å sphere
2D prediction
Notation: protein structure 1D, 2D, 3D

<p>| | | | | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

© Burkhard Rost (TUM Munich)
Predict 2D: How

☐ Predict all inter-residue
  • contacts
  • distances

☐ or focus on some strong interactions?
Secondary structure
Beta-sheet pairing potentials


Tim Hubbard, Sanger

Beta-sheet pairing potentials
Beta sheet pairing: idea older

C Sander and S Lifson 1979 In Molecular mechanisms of biological recognition (M. Balaban) Biomedical Press
Beta-sheet pairing potentials

Prediction in “full” 2D

Prediction of (long-range) inter-residue contacts

long-range
Prediction in “full” 2D

- Prediction of (long-range) inter-residue contacts:
  - statistics
  - correlated mutations
  - neural networks

- CONTACT defined as $C_\alpha < 0.8$ nm (8 Ångstrøm)
Concept of compensatory mutations

<table>
<thead>
<tr>
<th>1</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>fyn_human VTLFVALID EARTEDLSF HKGEKFOILN SSEGDFWEAR SLTTGETQYI</td>
<td></td>
</tr>
<tr>
<td>yrk_chick VTLFIALYDY EARTEDDDSF QKGEKPHIIN NTEGDFWEAR SLLSGATQYI</td>
<td></td>
</tr>
<tr>
<td>fgr_human VTLFIALYDY EARTEDDLTF TKGEKMPILN NTEGDFWEAR SLLSGKTQCI</td>
<td></td>
</tr>
<tr>
<td>yes_chick VTVFVALDY EARTTDDLSF KGERFQIIN NTEGDFWEAR SIATGKTQYI</td>
<td></td>
</tr>
<tr>
<td>src_avis2 VTTFVALDY ESRTETDLFN KGERIQQIVN NTEGDFWHAH SLTTGQTQYI</td>
<td></td>
</tr>
<tr>
<td>src_aviss VTTFVALDY ESRTETDLFN KGERIQQIVN NTEGDFWHAH SLTTGQTQYI</td>
<td></td>
</tr>
<tr>
<td>src_avisr VTTFVALDY ESRTETDLFN KGERIQQIVN NTEGDFWHAH SLTTGQTQYI</td>
<td></td>
</tr>
<tr>
<td>src_chick VTTFVALDY ESRTETDLFN KGERIQQIVN NTEGDFWHAH SLTTGQTQYI</td>
<td></td>
</tr>
<tr>
<td>stk_hydat VTFIVALDY EARISEDLSF KKGERIQIIN TADGDWWAR SLITNSEQI</td>
<td></td>
</tr>
<tr>
<td>src_rsvp ...</td>
<td></td>
</tr>
<tr>
<td>hck_human ..IVVALDY EAIHHEDSLQ QKGDQMVLE ES.GEWKAR SLATRKEQYI</td>
<td></td>
</tr>
<tr>
<td>blk_mouse ..FVVALFDY AAVNRDLQV LKGEKIQVLR .STGDWWAR SLVTGREGQV</td>
<td></td>
</tr>
<tr>
<td>hck_mouse ..TIVVALDY EAIHREDLSF QKGDQMVLE .EAGEWKAQ SLATKKEQYI</td>
<td></td>
</tr>
<tr>
<td>lyn_human ..IVVALYDY DGIHPPDSLQ KGEKMKVLE .EHGEWKAQ SLLTKKEQFI</td>
<td></td>
</tr>
<tr>
<td>lck_human ..LVIALHSY EPSHGDLGF EKEQIRILE QS.GEWKAQ SLTTGQEQFI</td>
<td></td>
</tr>
<tr>
<td>ss81_yeast ....ALYPY DADDDeISF EONEIQVSD .IEGRWKAR R.AGETGQI</td>
<td></td>
</tr>
<tr>
<td>ab1_mouse ..LFVALYDF VASGNTLSI TKGEKIRVLG YnnGEWCAQ ..TKNGQGVV</td>
<td></td>
</tr>
<tr>
<td>abll_human ..LFVALYDF VASGNTLSI TKGEKIRVLG YnnGEWCAQ ..TKNGQGVV</td>
<td></td>
</tr>
<tr>
<td>src1 drome ..VVVALDY KSRDESDSLQ MKGDMEVID DTESDWRVV NLTTRQEGLI</td>
<td></td>
</tr>
<tr>
<td>mysd_dicdi ....ALYDF DAESEMELSFT KEGDIIYVLD QSSGDFWWAE L ..KGRRKV</td>
<td></td>
</tr>
<tr>
<td>yfj4_yeast ....ALYSF AGEESGDLPF RKGDVITIKL ksQNDWWTGR V ..NGREGIF</td>
<td></td>
</tr>
<tr>
<td>abl2 human ..LFVALYDF VASGNTLSI TKGEKIRVLG YNQNGESEV RSKNG. QGVV</td>
<td></td>
</tr>
<tr>
<td>tec_human .EIVVAMYDF QAAGHDLRL ERGQYLYLLE KNDVHWRRAR D.KYGENQY</td>
<td></td>
</tr>
<tr>
<td>abl1_caeel ..LFVALYDF HVGGEQQLSL RKGDVIRILG YNKNNECEA RlrLGEIGV</td>
<td></td>
</tr>
<tr>
<td>txk_human ...ALYDF LPERPCNLAR RRAEYLYLLE KYNPHWKAR D.RLGNQYI</td>
<td></td>
</tr>
<tr>
<td>yha2 yeastr VRRVALYDL TTNEPDELSF RKGDVITVLE QVYRDWWGA L ..RGNMGF</td>
<td></td>
</tr>
<tr>
<td>abpl_sacex ....AEYDY EAGEDNELTF AENDKIINIE FVDDDWLGE LETTGQKGLF</td>
<td></td>
</tr>
</tbody>
</table>

AEYDY KAVEDNLETF DENDKIINIE FVDDDWLGE LETTGQKGLF

© Burkhard Rost (TUM Munich)
Concept of compensatory mutations

fyn_human VTLFVALYDY BARTEDDLSE HKGEKFQILN SSEGDWEAR SLTTGETGQYI
yrk_chick VTLFIALYDY BARTEDDLSE QKGEKFHIIN NTEGDWEAR SLSSGATQI
fgr_human VTLFIALYDY BARTEDDLTE TKGEKFHILN NTEGDWEAR SLSSGKTQI
yes_chick VTVFVALYDY BARTTDDLSF KGERFQIIN NTEGDWEAR SIATGTKQI
src_avis2 VTTFVALYDY ERRTETDLSE KGERIQUIVN NTEGDWLAH SLTTGQTQI
src_aviss VTTFVALYDY ERRTETDLSE KGERIQUIVN NTEGDWLAH SLTTGQTQI
src_avisr VTTFVALYDY ERRTETDLSE KGERIQUIVN NTEGDWLAH SLTTGQTQI
src_chick VTTFVALYDY ERRTETDLSE KGERIQUIVN NTEGDWLAH SLTTGQTQI
stk_hydat VTIFVALYDY ERAISEDLSF KGERIQQIVN TADGDWEAR SLITSEQQI
src_rsvpa ............
src_rsvpa ............
hck_human ..IVVALYDY EAIHHLDELSF QKGDQMVLE ES.GEWKAR SLATRKEQYI
blk_mouse ..FVVALDY AAENDRDLQVL KGEKIQLVLR .STGDWLAH SLVTGREGQV
hck_mouse .TIVVALYDY EAIHREDLSF QKGDQMVLE .EAGEWKAR SLATKKEQYI
lyn_human ..IVVALPYD GGIHPDLSF KGEKMKVLE .EHGEWKAK SLTKKKEQYI
lck_human ..LVIALHSY EPHSIGDLGF KEGQRIILE QS.GEWKAQ SLTTGQEQQI
ss81_yeast .......ALYPY DADDDdeISF EONEILQVSD .IEGRWARK R.ANGETGI
abl_mouse ..LFVALYDF VASGNTLSI TKEGIKRVLG YnnGEWCAQ ..TNGQGQV
abl1_human ..LFVALYDF VASGNTLSI TKEGIKRVLG YnnGEWCAQ ..TNGQGQV
src1_drome ..VVVLHYDY KSRDESLSF MKGDMVEVID DTESDWRVV NLTTRQEGI
mysd_dicdi.....ALYDF DAESEMELSF KEGDIITVLD QSSGQDWAEL ...KGRRKV
yfj4_yeast .......ALYSF AGSESGDLPF RKGDVTILK ksQNDWARTV ...NGREGF
abl2_human ..LFVALYDF VASGNTLSI TKEGIKRVLG YNQNGEWEV RSKNG.QGV
tec_human .EIVVAMYDF QAAEGHDLRI ERGQEYMILE KNDVHWRAR D.KYNGEQY
abl1_caeel ..LFVALYDF HSVGEEQLSL RKGDVQRIILG YNKNWCEA RlrLGEIGV
txk-human .EIVVAMYAF QAAEGHDLRI ERGQEYMILE KNDVHWRAR D.KYNGEQY
yha2_yeastVRRVALYDI TTNPEDELGF RKGDVTVLQE QVYRDWKGALA ...RGNMGF
abpl_sacex....AEYDY EAGEDNLTE AENDKIINIE FVDDWLGGE LETTGQKGLF

AEYDY KAEDNELTE DENDKIINIE FVDDWLGGE LETTGQKGLF

Monday June 23, 2014
Correlated mutations predict contacts

U Goebel, C Sander, R Schneider & A Valencia
(1994) 18:309-17

(note: U Göbel in PubMed)

Alfonso Valencia
CNIC Madrid
evolutionary constraints on protein sequences

- selective pressure from need to maintain protein function
- consequently, conservation and mutation patterns evidence of functional or structural constraints plus mutational drift
  - functional constraints: surface residues
  - mutational drift: loop regions
  - structural constraints: core

- simplifying assumption:
  residues in contact show correlated mutational behaviour, i.e., if one residue mutates, its contact partners also tend to mutate

Do correlated mutations imply spatial proximity?

- sometimes

  » (Altshuh et al., 1987; Altshul et al., 1988; Neher, 1994; Taylor & Hatrick, 1994; Shindyalov et al., 1994; Goebel et al., 1994)
Weak signal from correlated mutations

U Goebel, C Sander, R Schneider & A Valencia (1994)
18:309-17: Fig. 2
Weak signal from correlated mutations

U Goebel, C Sander, R Schneider & A Valencia (1994) 18:309-17: Fig. 4
www.rostlab.org

Marco Punta
IAS/TU Munich
PROFcon contact prediction

Two-fragment input

9 adjacent residues around $i$ and $j$

region of interaction between $i$ and $j$

region connecting $i$ and $j$

3 central residues between $i$ and $j$

Input general:
- information ± 4 residues around $i$
- information ± 4 residues around $j$
- information about 3 central residues between $i$ and $j$
- average information about connecting region (composition of profile and predicted secondary structure)

Input per-residue:
- residue identity
- evolutionary profile
- conservation weight
- predicted secondary structure
- predicted accessibility

M Punta and B Rost (2005) *Bioinformatics* 21: 2960-8
PROFcon: stronger prediction better

M Punta and B Rost (2005) *Bioinformatics* 21: 2960-8
PROFcon: predict inter-residue contacts

M Punta and B Rost (2005) *Bioinformatics* 21: 2960-8
Take home: 2D prediction

- Prediction hard, but stakes are high
- inter-residue
  - Correlated mutations can imply spatial proximity
  - Distinction between different models, no prediction of 3D, yet
- inter-strand
  - Sometimes good enough for approximate modeling of 3D

Don’t freak out when accuracy is low
- 1) how accurate are these prediction methods on average
- 2) are all important contacts predicted?

for 5% of the best-predicted contacts prediction accuracy about 50%
2D prediction: better read off 3D?

- option 1: predict contacts from sequence
- option 2: build homology model or *de novo* model and read off contacts

which is better?
Contact specialist better than models

O Grana et al (2005) *Proteins* 7:214-24: Fig. 5

Fig. 5. Mean values of accuracy and \( X_d \) for all groups that made predictions for all 11 NF targets. (a) mean accuracy and (b) mean \( X_d \). Groups labeled "RR" are contact prediction groups, while those labeled "AL" and "TS" are structure prediction groups.
2D prediction useful or not?
2D prediction useful or not?
PROFcon correlates with folding rates

M Punta and B Rost (2005) *J Mol Biol* 348: 507-12
light at the end of the tunnel ...
Membrane protein 3D from sequence ALONE!

Debora Marks
Harvard Medical

Thomas Hopf
TUM Munich

Chris Sander
Sloan Kettering NYC
Correlated mutations/Coevolution

Correlated mutation in 1D

doi: 10.1016/j.cell.2012.04.012

© Debbie Marks - HMS © Burkhard Rost (TUM Munich)
Residue contacts accurately predicted

β2 adrenergic receptor

G-3-P transporter GlpT

doi: 10.1016/j.cell.2012.04.012

© Thomas Hopf - TUM & Debbie Marks - HMS

© Burkhard Rost (TUM Munich)
11 medically important TMH predicted

OCTN1

Adiponectin receptor 1

MT-ND1

Crohn's disease, rheumatoid arthritis

diabetes, obesity, cancer

LHON, MELAS, Alzheimer, Parkinson


© Thomas Hopf - TUM & Debbie Marks - HMS
2D correlated mutations
Thanks for slides to Thomas Hopf
The structure prediction problem

genotype → phenotype
The structure prediction problem

Genotype: ACTGTGCACG
Phenotype: TAATGGGCATC
Structure from sequence alone?

Christian Anfinsen, Nobel Prize for Chemistry 1972
Sequence-structure gap is not closing!

Marks, Hopf & Sander, Nature Biotechnology (2012)
A protein
Evolutionary selection leaves residue covariation signature
Evolutionary selection leaves residue covariation signature
Folding proteins from evolutionary couplings
Folding proteins from evolutionary couplings
Try something simple: correlation between two columns

To what extent do we see a pair of amino acids more/less often than expected by chance?

single column frequencies: \( f_i(A_i) \) \( f_j(A_j) \)

column pair frequencies: \( f_{ij}(A_i, A_j) \)

\[ f_{ij}(A_i, A_j) - f_i(A_i) f_j(A_j) \]
Mutual information measures correlation between two columns

\[ MI_{ij} = \sum_{A_i, A_j=1}^{q} f_{ij}(A_i, A_j) \ln \left( \frac{f_{ij}(A_i, A_j)}{f_i(A_i)f_j(A_j)} \right) \]

- sum all possible amino acid combinations
- weight
- deviation from statistical independence
Bad news: doesn’t work.

local model
main problem: transitivity

PDB structure residue contacts
residue pairs with 100 highest MI values
Solution: use a global model!

- Global probability model explains observed correlations by causative pair interactions.

- Observed correlations:
  - inverse Potts inference (maximum entropy)

- Causal, direct coevolution:
Most global model pairs are close in 3D

local model
(mutual information)

![Diagram of local model]

global model
(MaxEnt, Marks et al., 2011)

![Diagram of global model]
A brief reminder
The breakthrough: 15 proteins folded from sequences alone

Marks et al., PLoS ONE (2011)
Also works for membrane proteins!

predicted  experimental

Hopf et al., Cell (2012)
Acknowledgements I
Acknowledgements I

Alan Lapedes
Acknowledgments II

Debora Marks
(Harvard Medical School)

Chris Sander (MSKCC)

Thomas Hopf

www.evfold.org
Lecture plan PP1: Structure

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

Monday June 23, 2014