1D: TM transmembrane helix prediction
TMHs (helices) correctly predicted?

Observed Helix 1 (O1)  O2  O3
Predicted Helix 1 (P1)  P2  P3
TMHs (helices) correctly predicted:
if at most ±5 residues overlap

here=0
Prediction of membrane helices

Q_{ok}: % of protein with all TMH right

TMH* prediction 2

*TransMembrane Helix
New problems, new methods
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
TMH proteins: reminders

- statistics for PDB in June 2010:
  - 67,086 structures in PDB (June 2010)
  - 1,197 transmembrane
  - 246 unique* (<99%PIDE!)

- 1,014 alpha helical
- 179 beta barrel

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

S Jayasinghe, K Hristova, SH White (2001) Protein Sci 10:455-8
GE Tusnady, ZS Dosztanyi & I Simon (2005) Bioinformatics 21:1276-7
Generic TMH protein

N_{out}

H=hydrophobic
L=hydrophilic

© Arne Elofsson (Stockholm Univ)

© Burkhard Rost
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


© Arne Elofsson (Stockholm Univ)
36 reentrant helices
- 20 in new classification
24% contain reentry
72% on the outside
Length 3-32 residues
Loops 11-117 residues


© Arne Elofsson (Stockholm Univ)
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

The Z-coordinate

Z-coordinate: distance residue 2 membrane center


© Arne Elofsson (Stockholm Univ)
New methods for old & new problems: TMSEG
Last time...

Per-segment: may be improvement possible?

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

![Graph showing comparison of different methods with error bars](image)
Yet another transmembrane predictor?

- More data available
  - Re-training old methods is viable but not done
- Less extensive machine learning
- Runtime

TMSEG overview

© M Bernhofer et al 2016 Proteins

Step 1: Compute initial scores for each residue.

Step 2: Filter scores and assign signal peptide and TMHs

Step 3: Adjust and split TMHs

Step 4: Predict in/out topology

Dataset – Transmembrane helices I

166 membrane protein sequences *(TMP166)*

- TMH assignment from 3D-structure by OPM & PDBTM
  - Assignments differ, both used for training
- Map to UniProt sequence using SIFTS
- Redundancy reduction with UniqueProt at HVAL > 0

**References:***


Dataset – Transmembrane helices II

Inside/Outside topology assignment OPM

Split into 4 subsets, maintaining distribution of TMPs, SPs and sequence lengths

Use 3 sets for cross-validation, keep one for final independent evaluation (Blind set)
TMSEG step 1: Random Forest

Step 1: Compute initial scores for each residue.

Step 2: Filter scores and assign signal peptide and TMHs

Step 3: Adjust and split TMHs

Step 4: Predict in/out topology

TMSEG step 1: Random Forest

Step 1: Compute initial scores for each residue.

Step 2: Filter scores and assign signal peptide and TMHs

Step 3: Adjust and split TMHs

Step 4: Predict in/out topology

TMSEG step 1: Random Forest

- Model: Random Forest (T=100, m=9)
- Sliding window of 19 residues (w=19)
- 3 scores for each residue (0-1000)
  - signal peptide
  - transmembrane helix
  - soluble
- Scores scaled [0-1] -> [0-1000]

---

## TMSEG step 1: Features

### Global features:
- Global amino acid composition $2\times 20$
- Protein length $1$

### Local features:
- PSSM score $21\times 19$
- Distance to N- and C-terminus $2$
- Average hydrophobicity (Kyte-Doolittle) $2\times 1$
- % hydrophobic $2\times 1$
- % charged (positive & negative) $2\times 2$
- % polar $2\times 1$

TMSEG step 2: Empirical filters

Step 1: Compute initial scores for each residue.

Step 2: Filter scores and assign signal peptide and TMHs

Step 3: Adjust and split TMHs

Step 4: Predict in/out topology

TMSENG step 2: Empirical filters

- Smooth scores with median filter (w=5)
- Adjust scores to avoid over-prediction
  - soluble (-185)
  - TMH (-60)
- Assign each residue to state with top score
- Remove short signal peptides (<4 residues)
- Remove short TMHs (<7 residues)

## TMSEG step 2 – Example

**SEQ:** M G P R A R P A R L L L L L ...

**SIG:** 400 400 100 100 800 600 700 900 100 600 100 800 ...

**SOL:** 500 400 600 500 100 100 100 000 500 100 100 200 ...

**TMH:** 100 200 300 400 100 300 200 100 400 300 800 000 ...

→ Median filter

**SIG:** 400 400 400 400 600 700 700 600 600 600 ...

**SOL:** 500 500 500 400 100 100 100 100 100 ...

**TMH:** 100 200 200 300 300 200 200 300 300 ...

→ Adjust for overprediction

**SIG:** 400 400 400 400 600 700 700 600 600 600 ...

**SOL:** 315 315 315 215 -85 -85 -85 -85 -85 ...

**TMH:** 040 140 140 240 240 140 140 240 240 ...

**OUT:** S S S S S S S S S S S S ...

TMSEG step 3: Refine TMH prediction

Step 1: Compute initial scores for each residue.

Step 2: Filter scores and assign signal peptide and TMHs

Step 3: Adjust and split TMHs

Step 4: Predict in/out topology

TMSEG step 3: Refine TMH prediction I

- neural network (single ANN 25 hidden units)
- input: TMH segments of variable lengths
- features (for each segment):
  - amino acid composition \(2 \times 20\)
  - average hydrophobicity (Kyte-Doolittle) \(2 \times 1\)
  - % hydrophobic residues \(2 \times 1\)
  - % charged residues \(2 \times 1\)
  - segment length (exact number) \(1\)

split long TMHs (≥35 residues) into 2 (≥17)
• keep two TMHs if higher average score after split
adjust TMH endpoints by up to 3 residues in both directions
TMSEG step 4: Topology prediction

Step 1: Compute initial scores for each residue.

Step 2: Filter scores and assign signal peptide and TMHs

Step 3: Adjust and split TMHs

Step 4: Predict in/out topology

Random Forest \( (T=100, \ m=7) \)

assign double segments to sides 1 or 2

features:

- amino acid composition \( 2*2*20 \)
- % positive charge \( 2*2*1 \)
- % abs. difference of pos. charge side1/side2 \( 2*1 \)
TMSEG step 4: Topology prediction II

- consider only residues close to TMHs
  - 15 residues next to TMHs and 8 residues into TMHs
- predict topology of N-terminus and extrapolate
- if SP (signal peptide) predicted:
  residues following SP predicted as “outside”
TMH predictions on *blind* test set

![Bar chart showing TMH predictions on blind test set](chart.jpg)

### TMP classification

- **Very low misclassification rates**

<table>
<thead>
<tr>
<th>Method</th>
<th>TMP sensitivity</th>
<th>TMP FPR</th>
<th>Topology correct</th>
<th>Misclassified in human</th>
<th>More mistakes than TMSEG in human</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMSEG</td>
<td>98 ± 2</td>
<td>3 ± 1</td>
<td>93 ± 4</td>
<td>558</td>
<td>-</td>
</tr>
<tr>
<td>PolyPhobius</td>
<td>100 ± 0</td>
<td>5 ± 1</td>
<td>78 ± 7</td>
<td>770</td>
<td>212</td>
</tr>
<tr>
<td>MEMSAT3</td>
<td>100 ± 0</td>
<td>28 ± 2</td>
<td>93 ± 4</td>
<td>4,313</td>
<td>3,755</td>
</tr>
<tr>
<td>MEMSAT-SVM</td>
<td>98 ± 2</td>
<td>14 ± 2</td>
<td>88 ± 5</td>
<td>2,253</td>
<td>1,695</td>
</tr>
<tr>
<td>Baseline</td>
<td>95 ± 3</td>
<td>31 ± 2</td>
<td>75 ± 7</td>
<td>5,015</td>
<td>4,457</td>
</tr>
</tbody>
</table>

Applying TMSEG to other methods I

- High modularity (steps 1-4)

- Apply steps 3 and 4 to other methods
  - Step3: NN-based TMH prediction improvement
  - Step4: RF-based topology prediction

- Can this improve other methods?
Applying TMSEG to other methods II

Performance Improvement

- PolyPhobius
- MEMSAT-SVM
- MEMSAT3
- Baseline

Q_{ok}  Q_{top}

© Burkhard Rost
Availability

- Github: [github.com/Rostlab/TMSEG](http://github.com/Rostlab/TMSEG)
- PredictProtein: [predictprotein.org](http://predictprotein.org)

Helical transmembrane region (192, 214; length 23)
Type: helical transmembrane region - SO:0001812 -
Evidence: Prediction (TMSEG) - See Citing Info Below
How many membrane proteins are there?
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
Kingdoms similar in length

Kingdoms similar in amino acids usage

<table>
<thead>
<tr>
<th>Kingdoms</th>
<th>Number of codons</th>
<th>Hypothetical age</th>
</tr>
</thead>
<tbody>
<tr>
<td>archaea</td>
<td>4 2 2 2 2 4 2 3 2 6 1 2 4 2 6 6 4 4 4 1 2</td>
<td></td>
</tr>
<tr>
<td>bacteria</td>
<td>1 15 4 6 16 1 17 12 13 8 18 10 7 14 11 5 9 3 20 19</td>
<td></td>
</tr>
<tr>
<td>eukaryotes</td>
<td>10% 5% 15% 10% 5%</td>
<td></td>
</tr>
</tbody>
</table>

Family sizes similar

Cumulative percentage of proteins

Number of proteins in family
Inventory of life: membrane proteins

eukaryotes

bacteria

archaea

2013 note: some issues with data (incomplete sequences?) e.g. human has more than 18%

More TM -> more complexity?

Membrane LEGO

Inventory of life: compartments

TMH proteins: reminders

Edda Kloppmann & Marco Punta:

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

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

Henry Bigelow & BR Columbia Univ

© Burkhard Rost
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 toward
- D[a-z]: downward strand, facing inward
- D[A-Z]: downward strand, facing towards bilayer

porin from R. blastica, pdb code 1pm

Henry Bigelow & BR Columbia Univ

© Burkhard Rost

54/122
PROFtmb: Model design

Arrows denote allowed transitions in the HMM. Dotted arrow/region indicates one connection per enclosed state.
1D: solvent accessibility
Get accessibility from 3D structures
Defining residue solvent accessibility
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
**PHDacc**

- **Local alignment**:
  - **AAA**: A C L I G S V
  - **LLL**: ins del cons
  - **AAG**
  - **CCS**
  - **GVV**

- **Global statistics**:
  - **%AA**: 100.0000000000 1.17
  - **Length**: 100.000000 0.42
  - **ΔN-term**: 0 33 66 0 0 0 0 0.74
  - **ΔC-term**: 66 0 0 33 0 0 0 1.17
  - **%AA**: 0 66 0 0 33 0 0 0.74
  - **Length**: 0 0 33 0 0 66 0 0.48

**Input local in sequence**

**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)
1D: Natively unstructured Disordered proteins

IDP: Intrinsically Disordered Proteins
Structure determines function
Order
Protein structure determines function

Method 1: predict B-values (flexibility)
Protein dynamics determine function

Figure from Predrag Radivojac, Indiana Univ

Coat protein of a tobacco mosaic virus

Tertiary structure  B-factors
Flexibility of proteins

superposition of 44 hen-white lysozyme structures

© Wikipedia
Danielkeedy
Backbone flexibility: B-value

where to threshold?
PROFbval

- Predict flexible/rigid residues through B-value data
- Can predict ‘X-ray disorder’
- Residues predicted to be **rigid** and **accessible** are correlated with the location of active sites (see output for RNAase HI)

![Graph](image)

![3D Structure](image)

Red/ - flexible
Yellow/green- intermediate
Blue - rigid

PROFbval: predict flexibility/rigidity

© COVER of Proteins

red = flexible  blue = rigid

ras

switch II
Gln61

beta-propeller
B-factor capture aspects of protein dynamics NOT directly of disorder!
Method 3: predict contact-deprived regions
Ucon: unstructured regions from contact prediction
Myosin serves as a glue

myosin

A Schlessinger, M Punta & B Rost 2007 submitted
MAX transcription factor (date hub)
Important to remember: so far we have NOT assumed that we know what disorder is!
Experimental
“handle”
on disorder
Types of natively unstructured regions

- Unstructured (conformational ensemble)
  - For example, ACTR (no NCBD)

- Molten globule (conformational ensemble)
  - For example, NCBD (no ACTR)

- Linked folded domains (beads on a string)
  - For example, zinc fingers (no DNA)

- Mostly folded, local disorder
  - For example, eIF-4E (N terminus is unfolded)

Folding on target binding:
- ACTR–NCBD complex
- Zinc-finger-1–3–DNA complex
- eIF-4E–eIF-4G complex

Dunker-hypothesis

Residues not visible in 3D structures share disorder
Different “types” of “experimental” “disorder” similar


- dis XRAY (2844)
- dis NMR (4019)
- dis CD (10554)
DisProt

Database of Protein Disorder

The Database of Protein Disorder (DisProt) is a curated database that provides information about proteins that lack fixed 3D structure in their putatively native states, either in their entirety or in part. DisProt is a collaborative effort between Center for Computational Biology and Bioinformatics at Indiana University School of Medicine and Center for Information Science and Technology at Temple University.

Latest additions:
- Human serum albumin
- Nogo-B
- Thymidylate synthase

Download DisProt
Download DisProt in FASTA or XML format.

(http://www.disprot.org/)
Method 4: MetaDisorder (MD)
Many methods predicting disorder

<table>
<thead>
<tr>
<th>Group</th>
<th>Method name</th>
<th>Definition of disorder</th>
<th>approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sussman &amp; Uversky</td>
<td>FoldIndex</td>
<td>DisProt</td>
<td>Hydrophobicity/net charge</td>
</tr>
<tr>
<td>David Jones</td>
<td>DISOPRED1</td>
<td>Xray</td>
<td>Neural Network</td>
</tr>
<tr>
<td>David Jones</td>
<td>DISOPRED2</td>
<td>Xray</td>
<td>SVM</td>
</tr>
<tr>
<td>Rob Russell</td>
<td>GlobPlot</td>
<td>'Hot' loops (High Bfactor loops)</td>
<td>Amino Acid propensities from PDB structures</td>
</tr>
<tr>
<td>Rob Russell</td>
<td>DisEMBL</td>
<td>Xray</td>
<td>Neural Network</td>
</tr>
<tr>
<td>Robert Esnouf</td>
<td>RONN</td>
<td>Invisible residues in Xray and NMR</td>
<td>Neural Network</td>
</tr>
<tr>
<td>Istevan Simon</td>
<td>IUPRED</td>
<td>Xray &amp; DisProt</td>
<td>Energy potentials</td>
</tr>
<tr>
<td>Pierre Baldi</td>
<td>DISpro</td>
<td>Xray</td>
<td>Neural Network</td>
</tr>
<tr>
<td>Robert MacCallum</td>
<td>DRIP-PRED</td>
<td>Xray</td>
<td>Self organizing maps and evolutionary information</td>
</tr>
<tr>
<td>Softberry</td>
<td>PreLink</td>
<td>Xray</td>
<td>Neural Network</td>
</tr>
<tr>
<td>Gianluca Pollastri</td>
<td>SPRITZ</td>
<td>Xray</td>
<td>SVM</td>
</tr>
<tr>
<td>Oxana Galzitskaya</td>
<td>FoldUnfold</td>
<td>DisProt</td>
<td>Average contact number</td>
</tr>
<tr>
<td>Keith Dunker</td>
<td>DisProt VL2</td>
<td>Different sets (NMR, CD, Xray)</td>
<td>Linear regression</td>
</tr>
<tr>
<td>Keith Dunker</td>
<td>DisProt VL3</td>
<td>DisProt</td>
<td>Neural Network</td>
</tr>
<tr>
<td>Keith Dunker</td>
<td>DisProt VL3H</td>
<td>DisProt</td>
<td>Neural Network + homology</td>
</tr>
<tr>
<td>Keith Dunker</td>
<td>DisProt VL3E</td>
<td>DisProt</td>
<td>Neural Network + evolutionary info</td>
</tr>
<tr>
<td>Keith Dunker</td>
<td>PONDR VL3BA</td>
<td>DisProt</td>
<td>Neural Network</td>
</tr>
<tr>
<td>Molecular kinetics</td>
<td>PONDR VSL1</td>
<td>DisProt + Xray</td>
<td>Logistic regression models</td>
</tr>
<tr>
<td>Molecular kinetics</td>
<td>PONDR VLXT</td>
<td>Fully disordered and fully ordered</td>
<td>Several machine learning methods</td>
</tr>
<tr>
<td>Chen-Ming Hsu</td>
<td>DisPSSM</td>
<td>Xray</td>
<td>PSSM + SVMs</td>
</tr>
<tr>
<td>iPDA</td>
<td>DisPSSM2</td>
<td>Xray</td>
<td>PSSM + SVMs + amino acid propensities</td>
</tr>
</tbody>
</table>
Simple average slightly improves prediction

**Method** | **Area under the curve**  
--- | ---  
IUPred+NORSnet+Ucon+DISOPRED2 | 0.765  
Ucon | 0.761  
IUPred | 0.752  
DISOPRED2 | 0.731  
NORSnet | 0.693

**True positive rate**: fraction of proteins with disorder correctly identified  
**False positive rate**: fraction of well-structured proteins mis-predicted with disorder
Meta disorder predictor (MD)

- Profiles
- Prediction methods:
  - DISOPRED2
  - NORSnet
  - Ucon (contacts only)
  - Ucon (contacts + energy)
  - PROFbval (predicted normalized B-values)
- Properties:
  - Predicted solvent accessibility
  - Predicted secondary structure
  - Predicted domain borders
  - Low complexity regions
  - Amino acid composition
  - Hydrophobicity/net charge
  - Length
  - Fraction of exposed residues
  - Secondary structure content

1D Jones et al JMB 2004 26:635-45
MD (meta disorder) most accurate

**True positive rate**: fraction of proteins with disorder correctly identified

**False positive rate**: fraction of well-structured proteins mis-predicted with disorder

<table>
<thead>
<tr>
<th>Method</th>
<th>Area under the curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD</td>
<td>0.809</td>
</tr>
<tr>
<td>IUPred+NORSnet+Ucon+DISOPRED2</td>
<td>0.765</td>
</tr>
<tr>
<td>Ucon</td>
<td>0.761</td>
</tr>
<tr>
<td>IUPred</td>
<td>0.752</td>
</tr>
<tr>
<td>RONN</td>
<td>0.746</td>
</tr>
<tr>
<td>DISOPRED2</td>
<td>0.731</td>
</tr>
<tr>
<td>NORSnet</td>
<td>0.693</td>
</tr>
<tr>
<td>PROFbval</td>
<td>0.691</td>
</tr>
</tbody>
</table>
Main findings

☐ Specific contacts are important for disorder prediction
☐ Hub proteins are abundant with unstructured loops
☐ Different methods focus on different aspects of protein disorder
☐ Combining predictors substantially improves overall prediction
Some findings & applications
Different methods find different proteins

**NORS:** J Liu, H Tan & B Rost 2002 J Mol Biol 322:53-64  
**PROFbval:** A Schlessinger & B Rost 2005 Proteins 61: 115-126  
**UCon:** A Schlessinger, M Punta & B Rost 2007 Bioinformatics 21:2376-84  
**MD:** A Schlessinger & B Rost 2009 PLoS One, 4: doi10.1371

**UCon**  
(contacts)

Max transcription factor (1an2)  

Capsid protein from cricket paralysis virus (1b35_C)  
Eukaryotes dominate disorder (4-10x)

Prediction method: MD IUPred

<table>
<thead>
<tr>
<th>Domain</th>
<th>Percentage of proteins with ≥30 consecutive residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virusea</td>
<td>20-30%</td>
</tr>
<tr>
<td>Eukaryota</td>
<td>36-43%</td>
</tr>
<tr>
<td>Bacteria</td>
<td>7-13%</td>
</tr>
<tr>
<td>Archaea</td>
<td>7-13%</td>
</tr>
</tbody>
</table>

A Schlessinger et al & B Rost 2011 *Curr Opin Struc Biol* 21:412-8
Molecular Recognition Element: MoRE

A Keith Dunker
Molecular Recognition Element (MoRE)
Dunker et al: α-MoRE predictions across 3 kingdoms

- **Eukaryotes**: High percentage of proteins with predicted MoREs, indicating a high prevalence of these residues in eukaryotic proteins.
- **Bacteria**: Moderate percentage of proteins with predicted MoREs, showing a moderate prevalence of these residues in bacterial proteins.
- **Archaea**: Lower percentage of proteins with predicted MoREs, indicating a lower prevalence of these residues in archaeal proteins.

The graph illustrates the distribution of predicted MoREs/Residue and the percentage of proteins with predicted MoREs across the three kingdoms, with Eukaryotes having the highest prevalence and Archaea the lowest.
2D prediction
Notation: protein structure 1D, 2D, 3D

1D

| P | PP | P  | 128 | 110 |
| Q | QQQ | Q  | 175 | 97  |
| I | PPQV | I  | 70  | E 60 |
| T | SSIVR | T  | 77  | E 69 |
| L | LLSTL | L  | 120 | E 14 |
| W | WWQED | W  | 238 | E 81 |
| Q | RKQAK | Q  | 169 | E 97 |
| R | RRRFPQ | R  | 200 | E 62 |
| P | PPPPP | P  | 24  | 48  |
| L | VVTKF | L  | 71  | E 59 |
| V | VVLI | V  | 14  | E 0  |
| T | TTKEK | T  | 74  | E 69 |
| I | AALIV | I  | 0   | E 0  |
| K | HYKKF | K  | 90  | E 73 |
| I | IIIV | I  | 4   | E 0  |
| G | EEENG | G  | 46  | E 41 |
| G | GGGTG | G  | 62  | E 53 |
| Q | QQKRR | Q  | 68  | E 71 |
| L | PLLW | L  | 118 | E 59 |
| K | VVFKE | K  | 31  | E 73 |
| E | EESKK | E  | 124 | E 95 |
| A | VVGLG | A  | 1   | E 0  |
| L | LLILL | L  | 29  | E 0  |
| L | LLLV | L  | 24  | E 0  |
| D | DDDDD | D  | 49  | E 58 |
| T | TTTTT | T  | 72  | 51  |
| G | GGGG | G  | 62  | 30  |
| A | AAAAA | A  | 17  | 0   |
| D | DDDDD | D  | 102 | 79  |
| D | DDAKE | D  | 69  | 58  |
| T | SSTTV | T  | 1   | 69  |
| V | IIIV | V  | 14  | E 0  |
| L | VVIVL | L  | 0   | E 0  |
Predict 2D: How

☐ Predict all inter-residue
  • contacts
  • distances

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

- C Sander and S Lifson 1979 In Molecular mechanisms of biological recognition (M. Balaban) Biomedical Press
  original idea

  advanced potentials

  better statistics

  topology
Prediction in “full” 2D

Prediction of (long-range) inter-residue contacts

long-range
Long- and short-range interactions
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)
Correlated mutations
Compensated mutations
Evolutionary couplings
Correlated mutations/Coevolution

correlated mutation in 1D

contact in 3D

doi: 10.1016/j.cell.2012.04.012

© Debbie Marks - HMS
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
Also works for membrane proteins!

predicted  experimental

Hopf et al., Cell (2012)
References for further reading

Protein 3D Structure Computed from Evolutionary Sequence Variation

Debora S. Marks¹, Lucy J. Colwell², Robert Sheridan³, Thomas A. Hopf¹, Andrea Pagnani⁴, Riccardo Zechini⁵, Chris Sander⁶

¹Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, United States of America, ²MRC Laboratory of Molecular Biology, Hills Road, Cambridge, United Kingdom, ³Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, New York, United States of America, ⁴Human Genomics Foundation, Tereno, Italy, ⁵Politecnico di Torino, Tereno, Italy

Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing

Thomas A. Hopf¹,², Lucy J. Colwell², Robert Sheridan³, Burkhard Rost⁴, Chris Sander⁶ and Debora S. Marks¹,⁶

¹Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA
²Department of Informatics, Technische Universität München, 85748 Garching, Germany
³MRC Laboratory of Molecular Biology, Hills Road, CB2 0QH Cambridge, UK
⁴Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York City, NY 10065, USA
⁵Correspondence: a.pagnani@memorial Sloan-kettering-cancer-center.org
DOI 10.1016/j.cellet.2012.04.012

Protein structure prediction from sequence variation

Debora S Marks¹, Thomas A Hopf⁶ & Chris Sander²

Sequence co-evolution gives 3D contacts and structures of protein complexes

Thomas A Hopf¹,², Charlotte P I Schärfe³,⁴, Joao P G L M Rodrigues⁵, Anna G Green¹, Oliver Kohlbacher²,⁴, Chris Sander⁶, Alexandre M J J Bonvin⁵, Debora S Marks¹⁵

© Burkhard Rost

Rostlab
PROFcon contact prediction

Two-fragment input:
- 9 adjacent residues around $i$ and $j$
- Region of interaction between $i$ 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
PROFcon CASP6 T230
2D prediction useful or not?
PROFcon correlates with folding rates

M Punta and B Rost (2005) *J Mol Biol* 348: 507-12
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
- contact in 3D

doi: 10.1016/j.cell.2012.04.012

© Debbie Marks - HMS
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
11 medically important TMH predicted

OCTN1
Crohn’s disease, rheumatoid arthritis

Adiponectin receptor 1
diabetes, obesity, cancer

MT-ND1
LHON, MELAS, Alzheimer, Parkinson


© Thomas Hopf - TUM & Debbie Marks - HMS
3D prediction
Molecular dynamics

- **2019**: folding@home: distributed computing (~100 peta FLOPS: world’s fastest computing system)
  Vijay Pande et al
- **2018**: HBV (hepatitis B virus)
  JA Hadden et al & K Schulten (2018) eLife 7:e32478
- **Anton (DE Shaw Research)**: special purpose MD computing

  also: use low-level predictions:
  A Raval, S Piana, MP Eastwood, DE Shaw (2016)
Complex of bacteria-infecting viral proteins modeled in CASP 13. The complex contains four separate subunits that were modeled individually. PROTEIN DATA BANK

Google’s DeepMind aces protein folding

By Robert F. Service | Dec. 6, 2018, 12:05 PM
Lecture plan (PP1 structure/comp biol)

- 01: 04/23 Tue: No lecture
- 02: 04/25 Thu: No lecture
- 03: 04/30 Tue: No lecture
- 04: 05/02 Thu: Intro 1: organization of lecture: intro into cells & biology
- 05: 05/07 Tue: Intro 2: amino acids, protein structure (comparison), domains
- 06: 05/09 Thu: Alignment 1
- 07: 05/14 Tue: Alignment 2
- 08: 05/16 Thu: Comparative modeling & exp structure determination & secondary structure assignment
- 09: 05/21 Tue: SKIP: Student Representation (SVV)
- 10: 05/23 Thu: 1D: Secondary structure prediction 1
- 11: 05/28 Tue: 1D: Secondary structure prediction 2
- 12: 05/30 Thu: SKIP: Ascension Day
- 13: 06/04 Tue: 1D: Secondary structure prediction 3
- 14: 06/06 Thu: 1D: Secondary structure prediction 4 - Deep Learning - Michael Heinzinger
- 15: 06/11 Tue: SKIP: Whitsun
- 16: 06/13 Thu: No lecture (but exercises)
- 17: 06/18 Tue: No lecture (but exercises)
- 18: 06/20 Thu: SKIP: Corpus Christi
- 19: 06/25 Tue: 1D: Transmembrane structure prediction 1
- 20: 06/27 Thu: 1D: Disorder prediction / Sketch: 2D & 3D prediction
- 21: 07/02 Tue: 2D: contact prediction - Deep Learning - Konstantin Weissenow
- 22: 07/04 Thu: Recap 1
- 23: 07/09 Tue: Recap 2
- 24: 07/11 Thu: TBA
- 25: 07/16 Tue: TBA
- 26: 07/18 Thu: TBA
- 27: 07/23 Tue: TBA
- 28: 07/25 Thu: TBA

EXAM: 07/17 18:00-20:00