Protein disorder -
IDP: intrinsically disordered proteins

cb1_1d_disorder
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

CONTACT: Inga Weise assistant@rostlab.org
Natively Unstructured Disordered proteins
Keith Dunker - Indiana Univ

CV
- BS Chemistry  UC Berkeley
- MS Physics  Univ Wisconsin
- PhD Biophysics  Univ Wisconsin
- PD (CompBiol)  Yale
- Indiana Univ

Publications  (2015/06 GoogleScholar)
- > 200 publications
- 2x >1,000
- 40x >200
- H-index 78

Structure determines function
Central dogma

- **DNA Polymerase**
  - DNA replication (DNA -> DNA)

- **RNA Polymerase**
  - Transcription (DNA -> RNA)

- **Ribosome**
  - Translation (RNA -> Protein)

- **Protein**

---

**Function**

**Structure**

slide: Andrea Schafferhans

© Burkhard Rost

ROSTLAB.
Protein structure determines function

Proteins are dynamic

Calmodulin

http://molmovdb.org/
Motions in proteins

- Allosteric regulation
- Enzyme catalysis
- Ligand binding
- Folding

Natively unstructured regions: induced fit

Types of natively unstructured regions

NORS: no regular secondary structure

< 5% helix or strand > 70 residues
Types of NORS in PDB

A: Connecting loops
B: Floppy ends
C: Wrapping loops
D: Floppy domains
10% of biomass weird!
Length distribution of floppy regions

J Liu, H Tan & B Rost 2002 J Mol Biol 322:53-64
Approaches for prediction

- Missing residues from electron density map
- Non Regular Secondary structure (NORS) regions
- Regions with High B-values
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

A Schlessinger & B Rost 2005 *Proteins* 61: 115-126
Backbone flexibility: B-value

where to threshold?
Backbone flexibility: B-value

A Schlessinger & B Rost 2005 *Proteins* 61: 115-126
Conservation of B-values

Table from Predrag Radivojac
Indiana Univ

<table>
<thead>
<tr>
<th>PIDE</th>
<th>All pairs</th>
<th>Same space group</th>
</tr>
</thead>
<tbody>
<tr>
<td>[70, 90)</td>
<td>0.59±0.07</td>
<td>0.63±0.09</td>
</tr>
<tr>
<td>[90, 100)</td>
<td>0.76±0.04</td>
<td>0.82±0.03</td>
</tr>
<tr>
<td>100</td>
<td>0.79±0.02</td>
<td>0.81±0.02</td>
</tr>
</tbody>
</table>
B-values imprinted onto sequence
PROFbval reliability correlates with accuracy

A Schlessinger & B Rost 2005 *Proteins* 61: 115-126
Good prediction for RNase H

RNase H
3D: 2rn2


NMR: Palmer lab

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)

**PROFbval: predict flexibility/rigidity**

© COVER of Proteins

![Beta-propeller](image)

**ras**

red = flexible

blue = rigid

A Schlessinger & B Rost 2005 *Proteins* 61: 115-126
how do B-values relate to disorder?

B-factor not directly proportional to disorder

Figure from Predrag Radivojac
Indiana Univ

B-factor capture aspects of protein dynamics NOT directly of disorder!
Method 2: predict short NORS regions
or:
distinguish unstructured from well-structured loops
Predict NORS (no regular secondary structure)

☐ less than 5% helix or strand over > 70 residues
coverage of protein space
Predict NORS (no regular secondary structure)

- less than 5% helix or strand over > 70 residues

- machine learning:
  true: all predictions in entire proteomes
  false: the whole PDB

NORSnet captures different aspects

FoldIndex: J Prilusky et al. JL Sussman 2005 *Bioinformatics* 21:3435-8
Natively unstructured ≠ well-structured loops

1pju_A  1aoc_A

1mkf_A  1gx2_A

Unstructured regions promiscuous

Number of interaction partners

Data for C. elegans

1
2

Normalized ratio unstructured/structured

© Burkhard Rost
application: hubs & disorder
Connect micro- and macro-level

**macro level:** networks
UP: more partners

**micro level:** residues
RIGHT: more hotspots
Date- and Party-hubs

□ Hubs: promiscuous proteins

□ Date/Party hubs
Notation introduced by Marc Vidal
JD Han et al. & M Vidal 2004 *Nature* 430:88-93

- **Date hubs** interactions at **different times/same location**?
- **Party hubs** interactions at **same time/different location**
More hotspots -> more party-hub like!

- Macro: more partners
- Micro: more hotspots

Graph showing:
- Non-hubs
- Party hubs
- Date hubs

Y Ofran, A Schlessinger & B Rost submitted - unpublished
Disorder is moldable
Unstructured regions promiscuous

![Bar charts showing the number of interacting partners and normalized values](image)

- **Excess unstructured by NORSnet**
- **Excess unstructured by DISOPRED2**

More unstructured -> more date-hub like!

- **Macro:** more partners
- **Micro:** more hotspots

- **Non-hubs**
- **Party hubs**
- **Date hubs**

Y Ofran, A Schlessinger & B Rost 2007 *PLoS CB*, 3:e140
Examples for Date & Party hubs

FUS3 MAP kinase - date hub (PDB 2b9f)
right complex with MSG5 binding motif (light blue)

ABC10-beta subunit of RNA polymerase - party hub
(PDB 1r9sJ
right: RNA Polymerase II elongation complex (ABC10-beta in red)
Correlation is not causation

Number of babies born per year

Number of storks per year

Data from Lower Saxony, Germany

Yu-An Dong et al & B Rost (2013) Intrinsically disordered proteins are network centers rather than hubs, in submission
Method 3: predict contact-deprived regions
**Ucon:** unstructured regions from contact prediction

![Graph showing PROFcon probability against residue number](image)

Residue number

![Heatmap showing unstructured regions](image)

Residue number

A Schlessinger, M Punta & B Rost 2007 submitted
Myosin serves as a glue

A Schlessinger, M Punta & B Rost 2007 submitted
Many colors of unstructured

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, eIF4E (N terminus is unfolded)

Folding on target binding:
- ACTR-NCBD complex
- Zinc-finger-1-3-DNA complex
- eIF4E-eIF4G complex

Dunker-hypothesis

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

http://www.disprot.org/

Avner Schlessinger
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

<table>
<thead>
<tr>
<th>Method</th>
<th>Area under the curve</th>
</tr>
</thead>
<tbody>
<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>DISOPRED2</td>
<td>0.731</td>
</tr>
<tr>
<td>NORSnet</td>
<td>0.693</td>
</tr>
</tbody>
</table>

**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\(^1\)
  - 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

\(^1\)D 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)

<table>
<thead>
<tr>
<th>Category</th>
<th>MD</th>
<th>IUPred</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virusea</td>
<td>10%</td>
<td>20%</td>
</tr>
<tr>
<td>Eukaryota</td>
<td>7%</td>
<td>43%</td>
</tr>
<tr>
<td>Bacteria</td>
<td>10%</td>
<td>20%</td>
</tr>
<tr>
<td>Archaea</td>
<td>5%</td>
<td>15%</td>
</tr>
</tbody>
</table>

Percentage of proteins with ≥30 consecutive residues

A Schlessinger et al & B Rost 2011 *Curr Opin Struc Biol* 21:412-8
Secondary structure (helix, strand) robust under random mutation, disorder not.
Disorder wonder switch for environment?
Disorder wonder switch for environment?

E Vicedo, A Schlessinger & B Rost 2013 Genome Biology, in preparation
THANKS to Avner Schlessinger

Avner Schlessinger
now: UCSF
Molecular Recognition Element: MoRE

A Keith Dunker
Molecular Recognition Element (MoRE)

A Keith Dunker
Example for MoRE: 4E binding protein

![Graph showing residues and their scores for EF4E Binding Region, PONDR, PHD Helix, and Hydrophobic Moment.]
Dunker et al: Predictors of α-forming MoREs

- Training set: 14 α-MoREs versus 1,200 globular, ordered proteins – both from PDB
- Inputs: short predictions of order, flanked by predictions of disorder using PONDR® VL-XT; flanking regions exhibit absence of hydrophobic clusters, disorder by VL2, low hydrophobic moment values, and GOR I prediction of coil and turns
- Adjust thresholds to reduce false positive error rate on helices from globular proteins while avoiding loss of training set examples

A Keith Dunker
Dunker et al: α-MoRE predictions across 3 kingdoms

% of Proteins with Predicted MoREs

Eukaryotes

Bacteria

Archaea

Predicted MoREs/residue (x10^{-3})

A Keith Dunker
Predicted MoRE of Measles Virus


A Keith Dunker
Measles Virus N and P Proteins


A Keith Dunker
Complex MoRE

Cyclin A

p27kip1

CDK

A Keith Dunker
Dunker’s parallel paradigms

Catalysis

AA seq ➔ 3D structure ➔ function

Signalling

AA seq ➔ disorder (ensemble) ➔ function

A Keith Dunker
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 symposium
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