Protein disorder -
IDP: intrinsically disordered proteins

cb1_1d_disorder

Computational Biology 1 - Protein structure (for Informatics) -
TUM summer semester
Videos: YouTube / www.rostlab.org

THANKS:

Special lectures:
- 06/20 Michael Bernhofer
- 07/04 TBC Predrag Radivojac - Indiana Univ.
- 07/06 Yana Bromberg - Rutgers Univ.

No lecture:
- 05/09 no lecture
- 05/23 Student assembly (SVV)
- 05/25 Ascension day
- 06/06 Whitsun holiday
- 06/15 Corpus Christi

LAST lecture: bef: Jul 11
after: Jul 28

Examen: WEDNESDAY(!!) July 12: 18:00-19:30 TBA
- Makeup: TBC: Oct 17 & 19, 2017 - lecture time
Natively Unstructured Disordered proteins
Structure determines function
Central dogma

- *DNA Polymerase*
  - Replication (DNA -> DNA)
- *RNA Polymerase*
  - Transcription (DNA -> RNA)
- *Ribosome*
  - Translation (RNA -> Protein)

**Function**

**Structure**

*dhorspool@en.wikipedia*

*slide: Andrea Schafferhans*
Protein structure determines function


Avner Schlessinger
Proteins are dynamic

Calmodulin

http://molmovdb.org/
Time scales for protein motion

- ligand binding
- catalysis
- folding
- allostERIC regulation
- libration
- sidechain rotation
- vibration

Natively unstructured regions: induced fit

Functionally disordered “state”

- Filters
- Flexible linkers
- Entropic chains: Fly casting again


unstructured
Features of disorder

- Efficient binders (large interface)
- Regulated through post-translational modifications
- Increasing complexity by structural plasticity
- Active in disordered version -> large difference between off and on
Coupled binding and folding

- Fly casting: increase surface to “reach out”
- Initial contacts weak/non-specific
- Folding upon approach of target (like hydrophobic collapse for protein-protein)

Order as scaffold for disorder

LEF-1/Tcf3

Beta Catenin

E-Cadherin (C-terminal domain)

Phosphorylation

WNT pathway

Cell adhesion
Types of natively unstructured regions

Increasing content of stable three-dimensional structure

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

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

- **Linked folded domains** (based on a string)
  - For example, zinc fingers (no DNA)
  - N
  - Finger 3
  - Finger 2
  - Finger 1

- **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
- eIF4E-eIF4G complex


© Burkhard Rost ROSTLAB. TUM
Disorder Data
DisProt

http://www.disprot.org/


© Burkhard Rost

ROSTLAB. TUM
“loopy” disorder
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

J Liu, H Tan & B Rost 2002 J Mol Biol 322:53-64
Length distribution of floppy regions

![Graph showing the length distribution of NORS regions with percentage of proteins and cumulative percentage.]

J Liu, H Tan & B Rost 2002 J Mol Biol 322:53-64
Disorder: simple sequence feature


© Burkhard Rost

Structured

Disordered
Disorder: simple sequence feature

Too easy to be biology

Mean net charge

Mean hydrophobicity

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

Figure from Predrag Radivojac
Indiana Univ
Flexibility of proteins

superposition of 44 hen-white lysozyme structures
Backbone flexibility: B-value

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

2 states: where to threshold?
Backbone flexibility: B-value

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

<table>
<thead>
<tr>
<th>PIDE [%]</th>
<th>All pairs [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[70, 90)</td>
<td>59±7</td>
</tr>
<tr>
<td>[90, 100)</td>
<td>76±4</td>
</tr>
<tr>
<td>100</td>
<td>79±2</td>
</tr>
</tbody>
</table>

Table from Predrag Radivojac, Indiana Univ

B-values imprinted onto sequence
PROFbval - predict residue flexibility

input local in sequence
A C L I G S V ins del cons
100 0 0 0 0 0 0 0 0 1.17
100 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)

local alignment
13 adjacent residues

global statistic
%AA Length ΔN-term ΔC-term
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 H1)

![Graph and diagram showing flexibility and rigidity]

- Red/ - flexible
- Yellow/green - intermediate
- Blue - rigid
PROFbval: predict flexibility/rigidity

© COVER of Proteins

beta-propeller

ras

red=flexible

blue=rigid
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

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
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

implies that many of those considered
1. “false” are “true”,
2. “true” are “false”.

How can a data set with many mistakes be machine learned?
How can a data set with many mistakes be machine learned?

TRICK: only SIGNAL is consistent
Natively unstructured ≠ well-structured loops

1pju_A  1aoc_A

1mkf_A  1gx2_A

Method 3: predict contact-deprived regions
Ucon: unstructured regions from contact prediction
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)

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

Increasing content of stable three-dimensional structure

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

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

**Linked folded domains** (loops 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
- eIF4E-eIF4G complex

Natura Reviews | Molecular Cell Biology

Dunker-hypothesis

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

\( \frac{\text{Disorder} - \text{Order}}{\text{Order}} \)

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

A Keith Dunker

© Burkhard Rost

ROSTLAB.
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

**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>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>
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)

**NORSnet** (loopy)

Max transcription factor (1an2)

Capsid protein from cricket paralysis virus (1b35_C)
Secondary structure (helix, strand) robust under random mutation, disorder not
Disorder stepping stone to increasing complexity?
Secondary structure (helix, strand) robust under random mutation, disorder not
Eukaryotes dominate disorder (4-10x)

Prediction method: MD IUPred

Percentage of proteins with ≥30 consecutive residues

Virusea
Eukaryota
Bacteria
Archaea

36-43%
7-13%

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

Cyclin A

CDK

p27kip1

A Keith Dunker
Proteome disorder content more similar to habitat than family
Yeast quick reaction to heat stress?

Orna Dahan
Yitzhak Pilpel
Quick reaction to heat stress: duplicate chromosome

Chromosomal duplication is a **transient**
evolutionary solution to stress

AH Yona et al & Y Pilpel & O Dahan
Quick reaction to heat stress: avoid disorder
Quick reaction to heat stress: avoid disorder

Watchya: science is communication
Correlation is not causation

Number of babies born per year

Data from Lower Saxony, Germany

Correlation is not causation

Number of babies born per year

Number of storks per year

Data from Lower Saxony, Germany

“Next Generation Biotechnologies and Society: New insights”

Dr. Erik Bongcam-Rudloff
SLU-Global Bioinformatics Centre
Uppsala, Sweden
Observing a phenomenon that is in some way interesting or puzzling.

Making a guess as to the explanation of the phenomenon.

Devising a test to show how likely this explanation is to be true or false.

Carrying out the test, and, on the basis of the results, deciding whether the explanation is a good one or not. In the latter case, a new explanation will (with luck) 'spring to mind' as a result of the first test.
The Observed phenomenon
Selection of test times
But was is the real event?
Sometimes you could be lucky

Positive

“Positive” results are used “negative” rejected
Why?
Only positive results are publishable
THANKS for slides to:

Marco Punta
ICR London

Avner Schlessinger
Mount Sinai, NYC

A Keith Dunker
Indiana Univ

Predrag Radivojac
Indiana Univ
THANK YOU

Dmitrij Nechaev

Jonas Reeb

Lothar Richter

Michael Bernhofer
THANK YOU
Lecture plan (CB1 structure)

- 01: 04/25 Tue: no lecture
- 02: 04/27 Thu: no lecture
- 03: 05/02 Tue: Intro 1: organization of lecture: intro into cells & biology
- 04: 05/04 Thu: Intro 2: amino acids, protein structure (comparison), domains
- 05: 05/09 Tue: No lecture
- 06: 05/11 Thu: Alignment 1
- 07: 05/16 Tue: Alignment 2
- 08: 05/18 Thu: Comparative modeling & exp structure determination & secondary structure assignment
- 09: 05/23 Tue: SKIP: student assembly (SVV)
- 10: 05/25 Thu: SKIP: Ascension Day
- 11: 05/30 Tue: 1D: Secondary structure prediction 1
- 12: 06/01 Thu: 1D: Secondary structure prediction 2
- 13: 06/06 Tue: SKIP: Whitsun holiday (06/03-06)
- 14: 06/08 Thu: 1D: Secondary structure prediction 3
- 15: 06/13 Tue: 1D: Transmembrane structure prediction 1
- 16: 06/15 Thu: SKIP: Corpus Christi
- 17: 06/20 Tue: 1D: Transmembrane structure prediction 2 / Solvent accessibility prediction
- 17: 06/20 Tue: 1D: Transmembrane structure prediction 2 / Solvent accessibility prediction
- 18: 06/22 Thu: 1D: Transmembrane structure prediction 3 / Solvent accessibility prediction
- 19: 06/27 Tue: 1D: Disorder prediction
- 20: 06/29 Thu: Recap 1
- 21: 07/04 Tue: Predrag Radivojac - Indiana Univ.
- 22: 07/06 Thu: Yana Bromberg - Rutgers Univ.
- 23: 07/11 Tue: Recap 2
- 24: 07/12 Thu: examen
- 24: 07/12 Thu: examen
- 25: 07/13 Tue: TBA
- 26: 07/18 Thu: TBA
- 27: 07/20 Tue: TBA
- 28: 07/22 Thu: TBA
- 29: 07/25 Thu: TBA