title: Protein disorder - IDP: intrinsically disordered proteins
short title: pp1_1d_disorder
lecture: Protein Prediction 1 (for Computational Biology) - Protein structure
TUM summer semester
Announcements

Videos:

Thank you!

EXERCISES:

Special lectures:
- Mikal Boden-UQ-Brisbane

No lecture:
- 04/26 Security check Rostlab (exercise WILL be)
- 05/01 May Day (also no exercise)
- 05/08 Student representation (SVV) - exercise WILL happen
- 05/10 Ascension Day (also no exercise)
- 05/22 Whitsun holiday (also no exercise)
- 05/31 Corpus Christi (also no exercise)
- 06/14 no lecture (student presentations)
- 06/19 no lecture (but exercise)
- 06/21 no lecture (but exercise)

last lecture: bef: Jul 11

Examen:
- Makeup: Jul 12 18-20:00 (room TBA)
- Makeup: Oct 16 (TBC)

Contact:

© Burkhard Rost

Videos:

YouTube / www.rostlab.org/talks

Contact: pp1ex@rostlab.org

THANKS:

EXERCISES:

Maria Schelling

Dmitrij Nechaev

Let it go. Let it out. Let it all unravel. Let it free and it can be a path on which to travel.
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

1. Replication (DNA -> DNA) by DNA Polymerase
2. Transcription (DNA -> RNA) by RNA Polymerase
3. Translation (RNA -> Protein) by Ribosome

Structure

Function

dhorspool@en.wikipedia
Order
Protein structure determines function

Time scales for protein motion

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

10^{-12} 10^{-9} 10^{-6} 10^{-3} 10^{0} 10^{3} s

Protein motion

PDB 3c4f_A (inactive) -> 3gqi_A (active)

FGFR1 kinase domain

PDB 2gha maltotriose binding protein

UCSF CHIMERA
www.cgl.ucsf.edu/chimera/animations/animations.html
Natively unstructured regions: induced fit

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)

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
Order as scaffold for disorder

LEF-1/Tcf3

Beta Catenin

Phosphorylation

E-Cadherin (C-terminal domain)

WNT pathway

Cell adhesion

Avner Schlessinger
Types of natively unstructured regions

Objective: develop methods without using explicit experimental data to avoid overfitting
“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
NORS: no regular secondary structure

< 5% helix or strand > 70 residues

use to predict in genomes
10% of biomass weird!

- Percentage of proteins
- Percentage of residues

- H sapiens
- M musculus
- S cerevisiae
- D melanogaster
- C elegans
- A thaliana
- N urealyticum
- T pallidum
- T maritima
- S PCC6803
- R prowazekii
- N meningitidis
- M tuberculosis
- M pneumoniae
- M genitalium
- H pylori
- H influenzae
- E coli
- D radiodurans
- C trachomatis
- C pneumoniae
- C jejuni
- B burgdorferi
- B subtilis
- A aeolicus
- P horikoshii
- P abyssi
- M thermoautotrophicum
- M jannaschii
- A fulgidus
- A pernix

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

Percentage of proteins vs. Length of NORS region

Cumulative percentage vs. Length of NORS region

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

where to threshold?
Backbone flexibility: B-value

- resolution < 2.5 (1513 proteins)
- resolution < 2 (926 proteins)
- resolution < 1.5 (142 proteins)

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

Predrag Radivojac

Indiana Univ
B-values imprinted onto sequence
PROFbval - predict residue flexibility

**Local alignment**
- **13 adjacent residues:**
  - A
  - C
  - L
  - I
  - G
  - S

**Global statistics**
- Length
- ΔN-term
- ΔC-term

**Input local in sequence**

<table>
<thead>
<tr>
<th>Intron</th>
<th>cons</th>
<th>del</th>
<th>ins</th>
<th>input</th>
<th>target</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td></td>
<td></td>
<td></td>
<td>100</td>
<td>0.17</td>
</tr>
<tr>
<td>A.A.</td>
<td></td>
<td></td>
<td></td>
<td>100</td>
<td>0.42</td>
</tr>
<tr>
<td>LLL</td>
<td></td>
<td></td>
<td></td>
<td>0 0</td>
<td>0.92</td>
</tr>
<tr>
<td>LLI</td>
<td></td>
<td></td>
<td></td>
<td>0 0</td>
<td>0.74</td>
</tr>
<tr>
<td>AAG</td>
<td></td>
<td></td>
<td></td>
<td>66</td>
<td>1.17</td>
</tr>
<tr>
<td>CCS</td>
<td></td>
<td></td>
<td></td>
<td>0 66</td>
<td>0.74</td>
</tr>
<tr>
<td>GWV</td>
<td></td>
<td></td>
<td></td>
<td>0 0</td>
<td>0.48</td>
</tr>
</tbody>
</table>

**Input global in sequence**

- Percentage of each amino acid in protein length of protein
  - (≤60, ≤120, ≤240, >240)
- Distance: center, N-term
  - (≤40, ≤20, ≤10)
- Distance: center, C-term
  - (≤40, ≤30, ≤20, ≤10)

**Output layer**
- **Rigid**
- **Flexible**
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

red = flexible
blue = rigid

beta-propeller

ras

Avner Schlessinger
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

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

NORSnet captures different aspects

FoldIndex: J Prilusky et al. JL Sussman 2005 *Bioinformatics* 21:3435-8

© Burkhard Rost
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
Many colors of unstructured
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 (based on c-DNA): 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

A Keith Dunker
# Experimental data for disorder

<table>
<thead>
<tr>
<th>Technique</th>
<th>Specific experiment</th>
<th>Information provided</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Circular Dichroism (CD)</strong></td>
<td>Far-ultraviolet (far-UV; 190-240 nm) CD</td>
<td>Secondary structure content, conformational changes.</td>
</tr>
<tr>
<td></td>
<td>Near-UV (250-350 nm)</td>
<td>Low intensity is correlated with disorder.</td>
</tr>
<tr>
<td><strong>Nuclear Magnetic Resonance (NMR)</strong></td>
<td>Chemical shift</td>
<td>Secondary structures and non native secondary structures</td>
</tr>
<tr>
<td></td>
<td>Nuclear Overhauser effect (NOE)</td>
<td>Long-range NOE- distances between residues that are far apart in sequence but close in space Mid-range NOE- secondary structure elements.</td>
</tr>
<tr>
<td></td>
<td>residual dipolar couplings (RDC)</td>
<td>The relative orientation of folded domains;</td>
</tr>
<tr>
<td></td>
<td>Exchange of amide protons</td>
<td>The overall protein’s stability and hydrogen bonds.</td>
</tr>
<tr>
<td></td>
<td>pulsed-field-gradient NMR (PFG-NMR)</td>
<td>Proteins shape and compactness.</td>
</tr>
<tr>
<td><strong>Limited proteolysis</strong></td>
<td></td>
<td>Solvent accessibility, flexibility.</td>
</tr>
<tr>
<td><strong>Small-angle X-ray solution scattering (SAXS)</strong></td>
<td></td>
<td>Proteins shape and compactness.</td>
</tr>
<tr>
<td><strong>Fluorescence techniques</strong></td>
<td>Intrinsic fluorescence of proteins</td>
<td>Compactness and conformational changes</td>
</tr>
<tr>
<td></td>
<td>Dynamic quenching of fluorescence</td>
<td>Solvent accessibility of defined groups and compactness of the protein</td>
</tr>
<tr>
<td></td>
<td>Fluorescence polarization and anisotropy</td>
<td>Mobility and compactness of native and molten globule states</td>
</tr>
<tr>
<td></td>
<td>Fluorescence resonance energy transfer (FRET)</td>
<td>Distances between atoms. hydrodynamic properties.</td>
</tr>
<tr>
<td></td>
<td>ANS fluorescence</td>
<td>Stability and transient intermediates in protein folding pathways</td>
</tr>
<tr>
<td><strong>Infra red (IR) spectroscopy</strong></td>
<td>Fourier-transform infrared spectroscopy (FTIR)</td>
<td>Bonds vibration frequencies, Secondary structure content. Investigation of protein denaturation and renaturation processes.</td>
</tr>
</tbody>
</table>
DisProt

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>Ianluca 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>
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>

**MD:** A Schlessinger, M Punta, G Yachdav, L Kajan & B Rost (2009)

*PLoS One 4:e4433*
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

---

Max transcription factor (1an2)

Capsid protein from cricket paralysis virus (1b35_C)
Christian Schaefer
Regular secondary structure sustains random mutations

C Schaefer & B Rost 2010 Bioinformatics 26:625-31
Astonishing mutations: helix<->strand

C Schaefer & B Rost 2010 Bioinformatics 26:625-31
Long disorder disappears

C Schaefer & B Rost 2010 Bioinformatics 26:625-31
Secondary structure (helix, strand) robust under random mutation, disorder not
Disorder stepping stone to increasing complexity?
Eukaryotes dominate disorder (4-10x)

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)
Example for MoRE: 4E binding protein

Residue

Relative Score

0 20 40 60 80 100

EF4E Binding Region  PONDR  PHD Helix  Hydrophobic Moment

A Keith Dunker
Dunker et al: Predictors of $\alpha$-forming MoREs

- Training set: 14 $\alpha$-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**
  - % of Proteins: 90%
  - MoREs/Residue: 0.3

- **Bacteria**
  - % of Proteins: 30%
  - MoREs/Residue: 0.1

- **Archaea**
  - % of Proteins: 10%
  - MoREs/Residue: 0.05

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

A Keith Dunker
Predicted MoRE of Measles Virus


A Keith Dunker
Complex MoRE

Cyclin A

CDK

p27kip1

A Keith Dunker
Dunker’s parallel paradigms

**Catalysis**

AA seq $\rightarrow$ 3D structure $\rightarrow$ function

**Signalling**

AA seq $\rightarrow$ disorder (ensemble) $\rightarrow$ function

A Keith Dunker
Eukaryotes dominate disorder (4-10x)

Prediction method: MD IUPred

<table>
<thead>
<tr>
<th>Domain</th>
<th>MD</th>
<th>IUPred</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virusea</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eukaryota</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacteria</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Archaea</td>
<td>10</td>
<td>20</td>
</tr>
</tbody>
</table>

Percentage of proteins with \(\geq 30\) consecutive residues

36-43% Eukaryotes vs 7-13% Bacteria and Archaea

A Schlessinger et al & B Rost 2011 *Curr Opin Struc Biol* 21:412-8
Eukaryotes dominate disorder (4-10x)

A Schlessinger et al & B Rost 2011 Curr Opin Struc Biol 21:412-8
Proteome disorder content more similar to habitat than family

Esmeralda Vicedo

Avner Schlessinger

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
Quick reaction to heat stress: avoid disorder
Quick reaction to heat stress: avoid disorder

Dark proteome

Sean O’Donoghue
CSIRO & Garvan Inst Sydney
Structural genomics: light the dark
$ 1,000,000,000

spent by PSI/NIGMS/NIH, USA
on Structural Genomics
Structural genomics: find common units

3D comparisons         yield         3D modules

- SCOP
  http://scop.mrc-lmb.cam.ac.uk/scop/
  A Murzin et al. 1995 JMB:247, 536-540

- CATH
  http://www.cathdb.info/
  AL Cuff et al. 2009 NAR 37:D310-314;
  CA Orengo et al. 1997 Structure 15:1093-1108

- COPS
  http://cops.services.came.sbg.ac.at
  SJ Suhrer et al. 2009 NAR 37:W539-W44

Multiple 3D alignment identifies consensus secondary structure
SG target selection: goal

- Identify all important families without 3D structure knowledge

All we need is that you choose the targets that work!

Wayne Hendrickson
Columbia Univ, NYC
Structural genomics: problems & solutions
Structural residue coverage (anything)

Organism

Archaebacteria
- A pernix
- A fulgidus
- M janaschii
- M thermoautotrophicum
- P abyssii
- P honkoehii

Prokaryotes
- A aeolicus
- B subtilis
- B burgdorferi
- C jejuni
- C pneumoniae
- C trachomatis
- D radiodurans
- E coli
- H influenzae
- H pylori
- M genitalium
- M pneumoniae
- M tuberculosis
- N meningitidis
- R prowazekii
- S PCC6803
- T maritima
- T pallidum

Eukaryotes
- S cerevisiae
- C elegans
- D melanogaster
- H sapiens (SPTrEml)

ALL

Percentage of residues in entire organism

Jinfeng Liu
SG contributes significantly to novel!

Every 4th novel protein from 4 US centers

PDB chains
Novel proteins

Percentage of all structures in year

2001 2002 2003 2004 2005


Jinfeng Liu
PSI record & attrition in structure determination

Date: 2000/07 - 2016/09

B Rost unpublished; source: TargetTrack @ PDB John Westbrook & Helen Berman
How many clusters?

16,000  MANY MORE

Prokaryotes only suffices?

YES  Only for now
We really cannot only do prokaryotes!
NYCOMPS structures sample diversity of life

More clones -> more success

© Edda Kloppmann
Close to known structure -> more success

![Graph showing the relationship between PIDE (Percentage pairwise sequence identity) and Success in PDB/clone]
Homologue structure of the SLAC1 anion channel for closing stomata in leaves

Yu-hang Chen1,2, Lei Hu3, Marco Punta2,4, Renato Bruni2, Brandon Hillerich2, Brian Kloss2, Burkhard Rost1,2,4, James Love2, Steven A. Siegelbaum3,5,6 & Wayne A. Hendrickson1,2,6,7

Wayne Hendrickson
Columbia Univ, NYC

Bacterium reveals how plants work

A peep through anion channels

The crystal structure of a protein channel provides clues about the mechanisms that control the closure of pores found in the epidermis of plant leaves. Excitingly, the protein channel folds in a way never seen before. See ARTICLE P.1074
NYCOMPS pipeline stages

Selected
Cloned
Expressed
Purified
Crystallize
DiffQual
Diffraction
in PDB

0 3100 6200 9300 12400 15500

NYCOMPS

ALL 9 PSI membrane

Thanks & Cheers

Wayne Hendrickson

Renato Bruni
Brian Kloss

Gunnar von Heijne
Stockholm Univ

Michael Wiener
U of Virginia

Larry Shapiro

Tingting Yang
Qun Liu
Filippo Mancia
Ming Zhou

Jinfeng Liu

Marco Punta
Edda Kloppmann

Reinhard Grisshamer
NIH-NINDS
Conclusion

15 years of PSI SG

~6,000 new structures

many more needed

and to do that is tough
Dark proteome

Authors

N Perdigao et al. & SI O’Donoghue (2015) PNAS 112:15898-903
Dark and non-dark

1BMF
BOVINE MITOCHONDRIAL F1-ATPASE
“Explanations” for darkness: incomplete

Dataset: dark proteins

N Perdigao et al. & SI O’Donoghue (2015) *PNAS* 112:15898-903, Fig. 3

©2015 by National Academy of Sciences
Darkness mostly NOT explained by common suspects

N Perdigao et al. & SI O’Donoghue (2015) PNAS 112:15898-903, Fig. 3
**Differences (dark-light): dark smaller families**

- **Dark proteins**
- **Overlap (dark/non-dark)**
- **Non-dark proteins**

**3D structures known for most very large protein families**

---

N Perdigao et al. & SI O’Donoghue (2015) *PNAS* 112:15898-903, Fig. 5

Some types of function more dark

- under-represented in dark proteins
- over-represented in dark proteins
- larger boxes: more significant

N Perdigao et al. & SI O’Donoghue (2015) *PNAS* 112:15898-903, Fig. 6
Darkness is functionally crucial

Density

Average RNA expression

0.6

0.4

0.2

0.0

1

10

100

1000

work in submission

dark proteins

overlap (dark/non-dark)

non-dark proteins

analysis: Andrea Schafferhans (HSWT Freising & TUM Munich) & Sean O’Donoghue (CSIRO & Garvan, Sydney)
Darkness is special

Machine Learning

1. Split subsample of 20% seq. identity data in training and test set

<table>
<thead>
<tr>
<th></th>
<th>train</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>dark proteins</td>
<td>1,814</td>
<td>171</td>
</tr>
<tr>
<td>non-dark</td>
<td>983</td>
<td>80</td>
</tr>
<tr>
<td>all</td>
<td>2797</td>
<td>251</td>
</tr>
</tbody>
</table>

2. Construction of kernel matrices > normalization
3. Grid search in combination with 5-fold cross validation on training set (WEKA)
   L: k-mer lengths      Y: conservation threshold

![Graph showing performance metrics with dark and non-dark proteins](image)

work in submission

Tatyana Goldberg  Jonas Reeb
More THANKS

Thanks for crucial SLIDES

Andrea Schafferhans-Fuhrmann
HSWT Weihenstephan & TUM

Thanks for work

Nelson Perdigao
Andrea Schafferhans

Sean O’Donoghue
CSIRO & Garvan Inst Sydney

THANKS for slides to:

Marco Punta
ICR London

A Keith Dunker
Indiana Univ

Avner Schlessinger
Mount Sinai, NYC

Predrag Radivojac
Indiana Univ
www.rostlab.org
THANK YOU

Dmitrij Nechaev

Maria Schelling
Thank YOU

© Michael Leunig

© Burkhard Rost

ROSTLAB. TUM
Lecture plan (PP1 structure/comp biol)

- 01: 04/10 Tue: No lecture
- 02: 04/12 Thu: No lecture
- 03: 04/17 Tue: No lecture
- 04: 04/19 Thu: Intro 1: organization of lecture: intro into cells & biology
- 05: 04/24 Tue: Intro 2: amino acids, protein structure (comparison), domains
- 06: 04/26 Thu: No lecture
- 07: 05/01 Tue: SKIP: May Day
- 08: 05/03 Thu: Alignment 1
- 09: 05/08 Tue: SKIP: Student Representation (SVV)
- 10: 05/10 Thu: SKIP: Ascension Day
- 11: 05/15 Tue: Alignment 2
- 12: 05/17 Thu: Comparative modeling & exp structure determination & secondary structure assignment
- 13: 05/22 Tue: SKIP: Whitsun holiday
- 14: 05/24 Thu: Comp modeling 2 & 1D: Secondary structure prediction 1
- 15: 05/29 Tue: 1D: Secondary structure prediction 1
- 16: 05/31 Thu: SKIP: Corpus Christi
- 17: 06/05 Tue: 1D: Secondary structure prediction 2
- 18: 06/07 Thu: 1D: Transmembrane structure prediction 1
- 19: 06/12 Tue: 1D: Transmembrane / Solvent accessibility / Disorder / 2D / 3D
- 20: 06/14 Thu: No lecture (exercise presentations)
- 21: 06/19 Tue: No lecture (but exercises)
- 22: 06/21 Thu: No lecture (but exercises)
- 23: 06/26 Tue: recap 1
- 24: 06/28 Thu: recap 2
- 25: 07/03 Tue: TBA
- 26: 07/05 Thu: TBA
- 27: 07/10 Tue: TBA
- 28: 07/12 Thu: TBA