Videos: YouTube / www.rostlab.org
THANKS:

Special lectures:
- (TBC)

No lecture:
- 11/07 Department event
- 11/09 Department event
- 11/16 Department event
- 11/23 Thanksgiving
- 12/19-01/06 - no lecture Xmas+

LAST lecture: Jan 18 (followed by 3 x wrap-up sessions)

Examen:
- TBC: Feb 01 10:15-11:45 lecture room
- Makeup: TBC: Apr 10 & Apr 12, 2018 - lecture time
IV. 
Predict protein interactions
IV.1 protein interactions

Protein-protein interactions (PPI): terminology
Different interfaces = different physics?

HIV gp120 / CD4 / FAB

Protein association

A activates
B activates
C activates
D activates ....

ABCD are associated
Physical interaction NOT association

HIV gp120 / CD4 / FAB

Yeast-2-Hybrid (Y2H) Method

Most common method to obtain binary protein-protein interaction data (Does X bind to Y?)
Original system (GAL4 system) developed by Fields & Song in 1989

Transcription Factor
BD=binding domain
AD=activation domain

BD and AD only function if they are physically linked with each other

© Burkhard Rost

B. Causier 2003 (Mass Spectr. Reviews)
IV.2 protein interactions

PPI - homology-based inference
Homology-based inference of PPI

A-B known experimentally
A'-B' inferred by homology
Much better intra-species

worm (C. Elegans)

Worm
Human

S Mika & B Rost 2006 PLoS Genetics, Vol 2, e29
Homology

© Wikipedia
Genome evolution

Orthologs
Paralogs

Species C1
duplication

Species B1

Species B2

Species C2

Species A
Jumping genes?
Horizontal gene transfer

The sea slug *Elysia chlorotica* incorporates chloroplasts from the algae that it ingests via a process called kleptoplasty. Photosynthesis continues for up to 12 months using genes within the chloroplast, which are directed by algal nuclear genes that were transferred to the nuclei of the slug.

Horizontal gene transfer

© Wikipedia

Figure: Barth F. Smets PhD Thesis
similar to BF Smets & T Barkay Nature Reviews Microbiology 3, 675-678 (September 2005)
Much better intra-species

S Mika & B Rost 2006 PLoS Genetics, Vol 2, e29
Inter and Intra-species the same?

similarity > X

A

B

A'

B'

A''

B''

Worm

Human

© Sven Mika & Burkhard Rost (Columbia New York)
Much better intra-species

S Mika & B Rost 2006 PLoS Genetics, Vol 2, e29
IV.3a protein interactions

PPI de novo?
Can we predict PPIs from sequence alone?
1999: one solution to predict PPI partners

Simple method failed fully to do this, problem: too many false positives
Road to predicting protein-protein partners

Implement simple method to do this failed entirely: too many false positives

Reduce false positives:

- predict surface residues (PROFacc, 1999)
  note: 1/2 of residues -> 1/4 of false positives!
Prediction of solvent accessibility

- 50% of residues somehow accessible to solvent
- 10% not at all
Road to predicting protein-protein partners

- Implement simple method to do this
  failed entirely: too many false positives

- Reduce false positives:
  - predict surface residues (PROFacc, 1999)
    note: 1/2 of residues -> 1/4 of false positives!
Road to predicting protein-protein partners

Implement simple method to do this
failed entirely: too many false positives

Reduce false positives:

- predict surface residues (PROFacc, 1999)
  note: 1/2 of residues -> 1/4 of false positives!

- predict residues in external interfaces (InteractionSites, 2004)
Predict protein-protein binding partners

Reducing false positives:

- predict surface residues (PROFacc, 1999)
- predict residues in external interfaces (InteractionSites, 2004)
- predict residues saturated internally (PROFcon, 2004)
- localization (e.g. only all nuclear, LOCtree, 2004)
Reduction by localization

e.g. nuclear:
=15% nuclear + 15% cytoplasm
=>
• come in with nuclear protein:
  maximally 30% of proteins to test
Predict subcellular localization: **LOCtree 2: 18 classes!**

*T Goldberg, T Hamp & B Rost* (2012) submitted k-mer profile kernel SVM

**Eukaryotic Protein Sequence**

- **SVM**
  - **Non membrane**
    - **SVM**
      - Intra-cellular
        - **SVM**
          - **ER**
            - **EXT**
          - **GOLGI**
          - **VAC**
          - **CYT**
          - **NUC**
          - **PERO**
          - **MITO**
          - **PLAST**
          - **CHLO**
      - **SECRETORY PATHWAY**
        - **SVM**
          - **PLAS**
          - **GOLGI**
          - **VAC**
          - **CHLO**
          - **MITO**
  - **Transmembrane**
    - **SVM**
      - Intra-cellular

**Tatyana Goldberg**

**Tobias Hamp**

**B Alberts et al. 1994**

*The Cell* Garland
Predict protein-protein binding partners

Reducing false positives:

- predict surface residues (PROFacc, 1999)
- predict residues in external interfaces (InteractionSites, 2004)
- predict residues saturated internally (PROFcon, 2004)
- localization (e.g. only all nuclear, LOCtree, 2004)
- predict residues in protein-substrate interfaces (active)
Predict protein-protein binding partners

Reducing false positives:

- predict surface residues (PROFacc, 1999)
- predict residues in external interfaces (InteractionSites, 2004)
- predict residues saturated internally (PROFcon, 2004)
- localization (e.g. only all nuclear, LOCreer, 2004)
- predict residues in protein-substrate interfaces (active)
- predict protein domains/improve alignments
Predict protein-protein binding partners

Reducing false positives:
- predict surface residues (PROFacc, 1999)
- predict residues in external interfaces (Interaction Sites, 2004)
- predict residues saturated internally (PROFcon, 2004)
- localization (e.g. only all nuclear, LOCtree, 2004)
- predict residues in protein-substrate interfaces (active)
- predict protein domains/improve alignments

Put it all together and predict binding partners
IV.4 protein interactions

PPI - data collection
Different interfaces = different physics?

At least 6 types of interfaces differ in sequence!

Internal (inter-domain and intra-domain)
External homomers (permanent/transient)
External heteromers (permanent/transient)

Interface types differ in composition

Interface types differ in composition

They obviously differ!
But, are these differences meaningful?
How to answer the meaningful question?
Are these differences statistically significant?

Chi-square test:

- known problem: small data sets
- here millions of points
- another not so well known problem: too large->problem

Y Ofran & B Rost 2005 unpublished
Are these differences statistically significant?

- Chi-square test:
  - known problem: small data sets
  - here millions of points
  - another not so well known problem: too large->problem

- All differences < 10^{-300}
  -> SIGNIFICANT
Are these differences statistically significant?

- Chi-square test:
  - known problem: small data sets
  - here millions of points
  - another not so well known problem: too large→problem

- all differences < 10^{-300}
  - → SIGNIFICANT

- ... unfortunately also:
  - proteins [a-b] vs [c-d]
  - 1 vs 2 authors
  - random subsets ...
Find-self test (statistical significance)

Y Ofran & B Rost 2005 unpublished
Find-self test on six types of interfaces

<table>
<thead>
<tr>
<th></th>
<th>internal</th>
<th>domain</th>
<th>homo-</th>
<th>homo-</th>
<th>hetero-</th>
<th>hetero-</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>obligomer</td>
<td>oligomer</td>
<td>obligomer</td>
<td>oligomer</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>internal</td>
<td>909</td>
<td>65</td>
<td>3</td>
<td>-</td>
<td>21</td>
<td>2</td>
</tr>
<tr>
<td>domain</td>
<td>53</td>
<td>812</td>
<td>2</td>
<td>-</td>
<td>128</td>
<td>5</td>
</tr>
<tr>
<td>homo-obligomer</td>
<td>-</td>
<td>2</td>
<td>925</td>
<td>-</td>
<td>12</td>
<td>61</td>
</tr>
<tr>
<td>homo-oligomer</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1000</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>hetero-obligomer</td>
<td>18</td>
<td>130</td>
<td>7</td>
<td>-</td>
<td>811</td>
<td>34</td>
</tr>
<tr>
<td>hetero-oligomer</td>
<td>-</td>
<td>8</td>
<td>58</td>
<td>-</td>
<td>38</td>
<td>896</td>
</tr>
</tbody>
</table>

IV.5 protein interactions

PPI predict binding sites
Different interfaces, different physics!

At least 6 types of interfaces differ in sequence!

Internal (inter-domain and intra-domain)
External homomers (permanent/transient)
External heteromers (permanent/transient)

Molecules of experimentally determined structure (3D co-ordinates)

www.pdb.org
- check out: Molecule of the Month

Stat 2010/04:
- ~65,000 structures
- 60K proteins
- 2K DNA/RNA
- 3K complexes
- 56K X-ray
- 8K NMR
- 0.3K Electron microscopy
extract interactions how?
Different interfaces = different physics?

HIV gp120 / CD4 / FAB

Different interfaces = different physics?

HIV gp120 / CD4 / FAB

1. PDB->Unique
2. parse heavy atoms <6.5 Ångstrøm (0.65 nm)
Develop method

1. parse heavy atoms <6.5 Ångstrøm (0.65 nm)
2. map chains to SWISS-PROT, distinguish transient protein-protein interactions from others
3. PDB sub(PP)->Unique

NOW we have a data set and can apply machine learning
PPI interfaces use local segments

INSERT:
Problem with choosing thresholds:
protein flexibility prediction
PROFbval
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

![Graph showing frequency of normalized B-values across different resolutions.](image)

A Schlessinger & B Rost *Proteins* 2005: 115-126
B-values imprinted onto sequence

A Schlessinger & B Rost 2005 *Proteins* 61: 115-126
PROFBval reliability correlates with accuracy

non-strict mode
1 if $B_{\text{norm}} \geq 0.03$
0 else

strict mode
1 if $B_{\text{norm}} \geq -0.30$
0 else

A Schlessinger & B Rost 2005 *Proteins* 61: 115-126
PROFbval: predict flexibility/rigidity

© COVER of Proteins

red = flexible
blue = rigid

beta-propeller

ras
PROFbval somehow separates active sites

A Schlessinger & B Rost unpublished
Backbone flexibility: B-value

A Schlessinger & B Rost 2005 *Proteins* 61: 115-126
Machine learning
how to choose the input features?
ask your friend
(ideally in the group)
Strength of prediction reflects reliability?

Strong: 0.9
Weak: 0.6

More complex system to predict structure

Sequence → PSI-BLAST → Filter

PROFsec → PROFacc

1999
## Alignment information

<table>
<thead>
<tr>
<th>Protein</th>
<th>Alignments</th>
<th>Profile table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y Y Y Y</td>
<td>I I E E</td>
<td>2 : 3 : : : :</td>
</tr>
<tr>
<td>D V V E</td>
<td>, 1 2 2 : : : :</td>
<td></td>
</tr>
<tr>
<td>D D T D</td>
<td>D N Q N N</td>
<td>: : : : 1 3 : : :</td>
</tr>
<tr>
<td>N E P K</td>
<td>N E P K</td>
<td>: 1 1 1 2 : : : :</td>
</tr>
<tr>
<td>F F F F</td>
<td>F F F F</td>
<td>: : : : : : : :</td>
</tr>
</tbody>
</table>

Corresponds to the 21*3 bits coding for the profile of one residue

B Rost & C Sander (1993) PNAS 90:7558-62
Few features

- Profile
- predicted 1D structure
  - secondary structure
  - solvent accessibility
  - membrane regions
  - disorder
- predicted aspects of function
Performance in predicting type of PPI interface
Are we there yet?
Let neural networks figure it out ...
Cross-validation: how?

All proteins

PDB-PP

CD-Hit / UniqueProt, e.g. 70% PIDE
Random split not enough
avoid overlap training/cross-training vs. testing
Now, are we there yet?
PP interfaces predicted from sequence
Strength of prediction reflects reliability?

strong
0.9
0.1
weak
0.6
0.4
PP interfaces predicted from sequence

Y Ofran & B Rost 2007 Bioinformatics e13-16
PP interfaces predicted from sequence

Accuracy:
>94% for 1 in 10
>70% for 2 in 10

PPI interfaces use local segments

PP interfaces predicted from sequence

Y Ofran & B Rost 2007 Bioinformatics e13-16
corresponds to the the 21*3 bits coding for the profile of one residue
PP interfaces predicted from sequence

Y Ofran & B Rost 2007 Bioinformatics e13-16

© Burkhard Rost

ROSTLAB. TUM
PPI hot spots?
Interaction HOT SPOTS

☐ residues that are essential for protein-protein interactions

☐ operational:
  • 1. residue in the interface
  • 2. mutation of the residue knocks out interaction
PP interfaces predicted from sequence

Very strong = hot spots ?

Y Ofran & B Rost 2007 Bioinformatics e13-16
Prediction of *hot spots* for CD4

- alanine scan for V1 domain of CD4 (bound to gp120)
  - A Ashkenazi et al. & DJ Capon 1990 PNAS 87: 7150

- structure:

- red: observed
- purple: predicted

enough to publish?
Hot spots reliably predicted from sequence!

hottest of hot = no error!

worst: ~60% right

What makes it work?

- Evolutionary information:
  - Optimally choosing profile
  - Explicitly using conserved residues

- (Predicted) 1D Structure
  important: good prediction + used correctly
  - Surface residues
  - Secondary structure

- Mark low-complexity and *sticky*

- Filtering “isolated predictions”
Hot spots prediction requires full information

- Sequence + Evolution + Exp. structure: 89
- Sequence + Evolution + Pred. structure: 85
- Evolution only: 36
- Sequence only: 35
- Hydrophobic Moment: 12

IV.6 protein interactions

PPI - hubs
Network level distribution of PPiS
Will all proteins have a similar number of interactions on average, or will have some more than others?
Which curve do you expect for PPIs per protein?

Histogram (number of proteins with that number of PPIs)

Number of PPIs of one protein
Which distribution do you expect?
If you plotted the histogram of settlement sizes, how would that look?
How to answer the question?
Sizes of metropolitan areas in the USA

Zipf’s law

\[ y = \frac{1}{x} \]
Curve for PPIs per protein trivially Zipf!

Histogram

(number of proteins with that number of PPIs)

Number of PPIs of one protein
Pick points at random: what will remain?
Half a Zipf is a Zipf
we have a method that predicts the number of interactions per protein: run for all
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

Non-hubs
Party hubs
Date hubs

Y Ofran, A Schlessinger & B Rost submitted
More unstructured -> more date-hub like!

macro: more partners

micro: more hotspots

- Non-hubs
- Party hubs
- Date hubs

Y Ofran, A Schlessinger & B Rost submitted
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)
Lecture plan (CB2 function)

01: 2017/10/17: No lecture (makeup examen; PP last year)
02: 2017/10/19: No lecture (sick)
03: 2017/10/04: No lecture (sick)
04: 2017/10/26: Welcome: who we are
05: 2017/10/31: No lecture (SVV)
06: 2017/11/02: Intro function 1: concept of protein function
07: 2017/11/07: No lecture (Department event)
08: 2017/11/09: No lecture (Department event)
09: 2017/11/14: Localization 1
10: 2017/11/16: No lecture (Department event)
12: 2017/11/23: No lecture (Thanksgiving)
14: 2017/11/30: Localization 4
15: 2017/12/05: PPI 2 - interaction sites (chalk)
16: 2017/12/07: No lecture (*Dies Academicus*)
17: 2017/12/12: PPI 2 - sites
18: 2017/12/14: PPI 3 - pairs
19: 2017/12/19: No lecture
20: 2017/12/21: No lecture
21-24: no lectures - winter break (2017/12/24 - 2018/01/06)
25: 2018/01/09: PPI 4 - PPI pairs 2
28: 2018/01/11: SNP effect 1 (chalk)
29: 2018/01/16: SNP effect 2
30: 2018/01/18: SNP effect 3
31: 2018/01/23: WRAP up 1
32: 2018/01/25: WRAP up 2
33: 2018/01/30: WRAP up 3
34: 2018/02/06: to be CONFIRMED examen (10:00-11:30)
35: 2018/02/08: No lecture

2018