Predict protein-protein interaction sites

cb2_ppi2
Videos: YouTube / www.rostlab.org

THANKS:
Tim Karl + Jonas Reeb

Special lectures:
• 12/02 Mikael Boden, Queensland U
• 12/16&18 - Andrea Schafferhans

No lecture:
• Dec 04 Thu (TUM Dies Academicus)

LAST lecture: January 20

Examen: January 22
• Makeup: Apr 14, 2015 - morning/noon

CONTACT: Tanya Goldberg goldberg@rostlab.org
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

B. Causier 2003 (Mass Spectr. Reviews)

©Sven Mika & Burkhard Rost (Columbia New York)
IV.2 protein interactions

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

A - B known experimentally
A' - B' inferred by homology

similarity > X

© Sven Mika & Burkhard Rost (Columbia New York)
Can we transfer binding through homology?

Obviously, otherwise no value in model organisms ...

\[ \text{similarity} > X \]

\[ \text{similarity} > X \]
Much better intra-species

S Mika & B Rost 2006 PLoS Genetics, Vol 2, e29
Much better intra-species

S Mika & B Rost 2006 PLoS Genetics, Vol 2, e29
Much better intra-species

S Mika & B Rost 2006 PLoS Genetics, Vol 2, e29
Much better intra-species

![Graph showing homology-inference for worm (C. Elegans)](image)

- **More similar**
- **Less similar**

S Mika & B Rost 2006 PLoS Genetics, Vol 2, e29
Homology-based inference of PPI: different species

Typical assumption: corresponding pair of proteins/genes in different species.
What about lateral gene transfer?
Homology

© Wikipedia
Genome evolution

duplication

Species A

Species B1

Species B2

Species C1

Species C2
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)
Genome evolution

Orthologs
Paralogs

Species A
Species B1
Species B2
Species C1
duplication
Species C2
Species A
Inter and Intra-species the same?

- Similarity > X
  - Worm
    - A ↔ A'
    - B ↔ B'
  - Human
    - A'' ↔ B''
Much better intra-species

less similar  more similar

© Burkhard Rost

S Mika & B Rost 2006 PLoS Genetics, Vol 2, e29
Much better intra-species

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

Orthologs

Paralogs

Species A

Species B1

Species B2

Species C1

duplication

Species C2
Why?
“Paralogs” conserve interactions
“orthologs” don’t?
Model organisms pose problems for protein-protein interactions
INSERT: measuring the same interaction twice
Measuring the interaction between A-B twice, results in the same interface?
Typical assumption: corresponding pair of proteins/genes in different species.
Not homology-based inference, but details!

A-B\textsuperscript{1} experimental structure 1
A-B\textsuperscript{2} experimental structure 2

identical

interfaces 1 and 2 identical?
Mostly the same but many differ
Many examples for alternative interfaces

A  RAS - SOS

B  Protein Kinase - Cyclin

C  RuBisCO

T Hamp & B Rost 2012 PLoS Comp Biol 8:e1002623
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 (ISIS, 2004)
Predict protein-protein binding partners

Reducing false positives:

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

<table>
<thead>
<tr>
<th></th>
<th>Extra-cellular</th>
<th>Cytoplasm</th>
<th>Organelles</th>
<th>Mitochondria</th>
<th>Nuclear</th>
<th>TM transmembrane</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extra-cellular</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cytoplasm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organelles</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mitochondria</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nuclear</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TM transmembrane</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

Secretory pathway

SVM

Intra-cellular

SVM

Transmembrane

Tatyana Goldberg

Tobias Hamp
Predict protein-protein binding partners

Reducing false positives:

- [x] predict surface residues (PROFacc, 1999)
- [x] predict residues in external interfaces (ISIS, 2004)
- [x] predict residues saturated internally (PROFcon, 2004)
- [x] 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 (ISIS, 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
Predict protein-protein binding partners

Reducing false positives:
- predict surface residues (PROFacc, 1999)
- predict residues in external interfaces
- 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

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-domain</th>
<th>homo-obligomer</th>
<th>homo-oligomer</th>
<th>hetero-obligomer</th>
<th>hetero-oligomer</th>
</tr>
</thead>
<tbody>
<tr>
<td>gp120</td>
<td>909</td>
<td>65</td>
<td>3</td>
<td>-</td>
<td>21</td>
<td>2</td>
</tr>
<tr>
<td>CD4</td>
<td>53</td>
<td>812</td>
<td>2</td>
<td>-</td>
<td>128</td>
<td>5</td>
</tr>
<tr>
<td>antibody-1</td>
<td>-</td>
<td>2</td>
<td>925</td>
<td>-</td>
<td>12</td>
<td>61</td>
</tr>
<tr>
<td>antibody-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1000</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>gp120</td>
<td>18</td>
<td>130</td>
<td>7</td>
<td>-</td>
<td>811</td>
<td>34</td>
</tr>
<tr>
<td>CD4</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

Develop method

1. PDB->Unique
2. parse heavy atoms <6.5 Ångstrøm (0.65 nm)
Different interfaces = different physics

HIV gp120 / CD4 / FAB

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

Machine learning
how to choose the input features?
ask your friend
(ideally in the group)
Strength of prediction reflects reliability?

0.9 (strong)
0.6 (weak)
0.1
0.4
More complex system to predict structure

Sequence → **PSI-BLAST** → **Filter**

**PROFsec**

**PROFacc**

1999

**PHD->PROF**

SPLIT again

JURY over 20

secondary structure
solvent accessibility
Alignment information

<table>
<thead>
<tr>
<th>Protein</th>
<th>Alignments</th>
<th>profile table</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>GGGG</td>
<td>5 ... 5 ... 5 ... 5 ...</td>
</tr>
<tr>
<td>Y</td>
<td>YYYY</td>
<td>2 ... 3 ... 5 ... 5 ...</td>
</tr>
<tr>
<td>I</td>
<td>IIEE</td>
<td>12221111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111...</td>
</tr>
</tbody>
</table>
Few features

- Profile
- Predicted 1D structure
  - Secondary structure
  - Solvent accessibility
  - Membrane regions
  - Disorder
- Predicted aspects of function
Publish not to perish
-
tie it up
Are we there yet?
Let neural networks figure it out ...
Cross-validation: how?

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

Y Ofran & B Rost 2007 Bioinformatics e13-16
Strength of prediction reflects reliability?

<table>
<thead>
<tr>
<th>Strength</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>strong</td>
<td>0.9</td>
</tr>
<tr>
<td>weak</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
</tr>
</tbody>
</table>
PP interfaces predicted from sequence

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

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

Successful prediction: skp1-skp2

Uniquitin ligase skp1-skp2 complex

Green: 2 correctly predicted residues
(pocket binding TRP109 of SKP-2 F-box protein)

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

Prediction system

Level 1:
Neural networks
input: alignment profile/predicted secondary structure + accessibility (PROF)/predicted sequence complexity/overall features (protein length, amino acid composition, asf.)
output: 2 units: is or is not P=P

Level 2:
Neural networks using input from previous level

Level 3:
simple clustering
PPI interfaces use local segments

PP interfaces predicted from sequence

Y Ofran & B Rost 2007 Bioinformatics e13-16
Alignment information

<table>
<thead>
<tr>
<th>Protein</th>
<th>Alignments</th>
<th>profile table</th>
</tr>
</thead>
<tbody>
<tr>
<td>G G G G</td>
<td>G G G G</td>
<td>5 \ldots \ldots \ldots \ldots : 5</td>
</tr>
<tr>
<td>Y Y Y Y</td>
<td>I I E E</td>
<td>\ldots : 2 : 3 : \ldots</td>
</tr>
<tr>
<td>I I E E</td>
<td>Y Y Y Y</td>
<td>\ldots \ldots \ldots \ldots : 5</td>
</tr>
<tr>
<td>D D D D</td>
<td>P P P P</td>
<td>\ldots : 5 : \ldots \ldots \ldots</td>
</tr>
<tr>
<td>P P P P</td>
<td>A E A A</td>
<td>\ldots : 3 : 2 : \ldots</td>
</tr>
<tr>
<td>D V V E</td>
<td>G G G G</td>
<td>\ldots : 1 : 2 : 2 : \ldots</td>
</tr>
<tr>
<td>G G G G</td>
<td>D D D D</td>
<td>\ldots : 5 : \ldots \ldots \ldots</td>
</tr>
<tr>
<td>D D D D</td>
<td>P P P P</td>
<td>\ldots : 5 : \ldots \ldots \ldots</td>
</tr>
<tr>
<td>P P P P</td>
<td>D T D D</td>
<td>\ldots : 4 : 1 : \ldots</td>
</tr>
<tr>
<td>D N Q N N</td>
<td>G N G G</td>
<td>\ldots : 1 : 3 : 1 : \ldots</td>
</tr>
<tr>
<td>G N G G</td>
<td>V I V V</td>
<td>\ldots \ldots \ldots \ldots : 4 : 1 : \ldots</td>
</tr>
<tr>
<td>V I V V</td>
<td>N E P K</td>
<td>\ldots : 1 : 1 : 1 : \ldots</td>
</tr>
<tr>
<td>N E P K</td>
<td>P P P P</td>
<td>\ldots : 5 : \ldots \ldots \ldots</td>
</tr>
<tr>
<td>P P P P</td>
<td>G G G G</td>
<td>\ldots : 5 : \ldots \ldots \ldots</td>
</tr>
<tr>
<td>G G G G</td>
<td>T T T T</td>
<td>\ldots : 5 : \ldots \ldots \ldots</td>
</tr>
<tr>
<td>D E K S A</td>
<td>F F F F</td>
<td>\ldots : 1 : 1 : 1 : \ldots</td>
</tr>
<tr>
<td>F F F F</td>
<td>: : : :</td>
<td>\ldots \ldots \ldots \ldots : 5</td>
</tr>
</tbody>
</table>

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

B Rost & C Sander (1993) PNAS 90:7558-62
PP interfaces predicted from sequence

Y Ofran & B Rost 2007 Bioinformatics e13-16
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)

red: observed

purple: predicted

(Y Ofran & B Rost (2006) ISIS *submitted*)

• structure:


(Y Ofran & B Rost 2007 *PLoS CB* 3:e119)
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

Functionally important residues - interactions sites

..LNDRA.

..LNDRA.

..---P--.


© Marco Punta & Yanay Ofran & Burkhard Rost (Columbia New York)
Find non-homologous competitive binder
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 distribution do you expect?
What do you expect of the following?

Histogram
(number of proteins with that number of PPIs)

Number of PPIs of one protein
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

\[ y = \frac{1}{x} \]
What do you expect of the following?

Number of PPIs of one protein

Histogram

(number of proteins with that number of PPIs)

0 25 50 75 100

0 25 50 75 100

Number of PPIs of one protein
Pick points at random: then what?
Half a Zipf is a Zipf
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

NORSnet

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)

ISIS unstructured

ABC10-beta subunit of RNA polymerase - party hub
(PDB 1r9sJ
right: RNA Polymerase II elongation complex (ABC10-beta in red)
Lecture plan (PP2 function)

- 01: 2014/10/07: no lecture
- 02: 2014/10/09: welcome: who we are
- 03: 2014/10/14: no lecture (prof sick)
- 04: 2014/10/16: no lecture (prof sick)
- 05: 2014/10/21: no lecture (make-up examen; PP last year)
- 06: 2014/10/23: Intro - function 1: concepts / homology
- 08: 2014/10/30: Tobias Hamp: Homology-based prediction of function 2
- 09: 2014/11/04: no lecture: SVV (student reps)
- 10: 2014/11/06: Intro - function 3: motifs
- 12: 2014/11/13: Localization 1
- 14: 2014/11/20: Localization 3 / Protein-protein interaction 1
- 17: 2014/12/02: SNP effect 1
- 18: 2014/12/04: no lecture: Dies Academicus
- 19: 2014/12/09: SNP effect 2
- 20: 2014/12/11: SNP effect 3 / Marco De Vivo (ISS Genoa) - Drug Design
- 21: 2014/12/16: Andrea Schafferhans: 3D function prediction
- 23-26: no lectures - winter break (2014/12/24 - 2015/01/06)
- 27: 2015/01/08: Punta - Pfam
- 28: 2015/01/13: Marco De Vivo (ISS Genoa) - Drug Design
- 29: 2015/01/15: GO enrichment
- 30: 2015/01/20: WRAP up !Protein-DNA/RNA interaction 2
- 31: 2015/01/22: examen

© Burkhard Rost ROSTLAB. TUM