title: IV: Protein protein interactions: sites 1

short title: cb2_ppi1

lecture: Protein Prediction 2 (for Computer Science) - Protein function
TUM winter semester
Announcements

☐ Videos: YouTube / www.rostlab.org

THANKS:

Dmitrij Nechaev

☐ Special lectures:
  • (TBC)

☐ No lecture:
  • 10/30 no lecture
  • 11/01 All Saints
  • 11/13 SVV (student rep)
  • 11/22 Thanksgiving
  • 12/06 Dies Academicus (TUM)
  • 12/20-01/06 - no lecture Xmas+
  • 01/08 may be no lecture?

☐ LAST lecture: Jan 22 (followed by 2 wrap-up sessions)

☐ Examen: Feb 07 10:00-13:00, LMU physics

☐ Makeup: TBC: Apr 23 & Apr 25, 2019 - lecture time

CONTACT: teaching@rostlab.org
CB2 Class 2018/19

CHRIStopher Onucha

TOBIAS Olenyi

Hasan Muhammad X

Reza X Celiv

VAGRAM

OMAR Abibi

JINLONG Ru

SOFIE Kemper

Marla Narazani

GHALIA Rehawi

FRANCE SCO Cogolato

Marco Both

Silvia Severini

Aynesh Sundararas

Muhammad Rafdi

© Burkhard Rost

ROSTLAB. TUM
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
IV.2 protein interactions

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

A - B known experimentally
A' - B' inferred by homology
Homology-based inference for PPI: worm

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

© Wikipedia
**Genome evolution**

Orthologs
Paralogs

Species C1
Species B1
Species B2
Species C2
Species A
duplication
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.

Inter and Intra-species the same?

A

B

A'

B'

A''

B''

Worm

Human

similarity > X

similarity > X

© Sven Mika & Burkhard Rost (Columbia New York)
Worm vs non-worm

Worm (C. Elegans)

Homology-Inference

HSSP

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

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

- More similar
- Less similar

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

<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!**

![Diagram showing the subcellular localization prediction with SVMs and different classes like Cytosol, Golgi, Nucleus, etc.](image)

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

**B Alberts et al. 1994 The Cell Garland**

**Tatyana Goldberg**

**Tobias Hamp**
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, LOCtree, 2004)
- ✗ predict residues in protein-substrate interfaces (active)
- ✔ predict protein domains/improve alignments
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)

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

PPI - data collection
Different interfaces = different physics?


HIV gp120 / CD4 / FAB

At least 6 types of interfaces differ in sequence!

Internal (inter-domain and intra-domain)

External homomers (permanent/transient)

External heteromers (permanent/transient)


Structure: Hendrickson lab
Interface types differ in composition

Amino acid compositions in different types of interfaces

3 interface types:
- internal
- chain-chain
- protein-protein

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
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)

- procedure for P1:
  - randomly draw S
  - repeat R times
  - report pair with minimal JS

- perform procedure for P2 and P3

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

<table>
<thead>
<tr>
<th>Type</th>
<th>Antigen 1</th>
<th>Antigen 2</th>
<th>Antigen 3</th>
<th>Antigen 4</th>
<th>Antigen 5</th>
<th>Antigen 6</th>
</tr>
</thead>
<tbody>
<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-domain</td>
<td>53</td>
<td>812</td>
<td>2</td>
<td>-</td>
<td>128</td>
<td>5</td>
</tr>
<tr>
<td>homooligomer</td>
<td>-</td>
<td>2</td>
<td>925</td>
<td>-</td>
<td>12</td>
<td>61</td>
</tr>
<tr>
<td>homooligomer</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1000</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>heterooligomer</td>
<td>18</td>
<td>130</td>
<td>7</td>
<td>-</td>
<td>811</td>
<td>34</td>
</tr>
<tr>
<td>heterooligomer</td>
<td>-</td>
<td>8</td>
<td>58</td>
<td>-</td>
<td>38</td>
<td>896</td>
</tr>
</tbody>
</table>

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)

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)
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
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

Resolution < 2.5 (1513 proteins)
Resolution < 2 (926 proteins)
Resolution < 1.5 (142 proteins)
B-values imprinted onto sequence
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
PROFbval: predict flexibility/rigidity

© COVER of Proteins

beta-propeller

ras

red=flexible

blue=rigid

A Schlessinger & B Rost 2005 Proteins 61: 115-126
PROFbval somehow separates active sites

A. [Graph showing frequency distribution of active site residues compared to non-active site residues across increasing rigidity and flexibility scales.]

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

weak

0.9
0.1
0.6
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>: : : :</td>
<td>GSAPD NTEKQ CVHIR LMYFW</td>
<td></td>
</tr>
<tr>
<td>G G G G</td>
<td>5 . . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>Y Y Y Y</td>
<td>2 . . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>I I E E</td>
<td>3 . . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>Y Y Y Y</td>
<td>5 . . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>D D D D</td>
<td>5 . . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>P P P P</td>
<td>5 . . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>E A E A</td>
<td>3 . . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>D V V E</td>
<td>1 . . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>G G G G</td>
<td>5 . . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>D D D D</td>
<td>5 . . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>P P P P</td>
<td>5 . . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>D D T D</td>
<td>4 . . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>D N Q N N</td>
<td>1 . . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>G G N G</td>
<td>4 . . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>V V V V</td>
<td>4 . . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>N E P K</td>
<td>1 . . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>P P P P</td>
<td>5 . . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>G G G G</td>
<td>5 . . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>T T T T</td>
<td>. . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>D E K S A</td>
<td>. . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>F F F F</td>
<td>. . . . . . . . . . .</td>
<td></td>
</tr>
<tr>
<td>: : : :</td>
<td>. . . . . . . . . . .</td>
<td></td>
</tr>
</tbody>
</table>

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

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?

PDB-PP

All proteins

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?

\[ P (\text{in } P=P) \]
\[ N (\text{not in } P=P) \]

\begin{align*}
\text{strong} & : 0.9 \\
\text{weak} & : 0.6
\end{align*}

\begin{align*}
\text{weak} & : 0.4
\end{align*}
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

![Graph showing accuracy vs coverage]

- IntSites
- Sequence only
- Random

Y Ofran & B Rost 2007 Bioinformatics e13-16

© Burkhard Rost
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:
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!

- micro: more hotspots
- macro: more partners

Graphs showing:
- 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

**Graph: NORSnet**
- **Non-hubs**
- **Party hubs**
- **Date hubs**
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)
preliminary Lecture plan (CB2 function)

01: 2018/10/16: No lecture (makeup examen; PP last year)
02: 2018/10/18: No lecture (makeup)
03: 2018/10/23: Welcome: who we are
04: 2018/10/25: Intro function 1: concept of protein function
05: 2018/10/30: No lecture
06: 2018/11/01: No lecture (holiday, All Saints)
07: 2018/11/06: Intro function 2: homology-based inference
08: 2018/11/08: Localization 1 (chalk talk)
09: 2018/11/13: No lecture (SVV)
10: 2018/11/15: Localization 2 (homology, motifs)
12: 2018/11/22: No lecture (Thanksgiving)
14: 2018/11/29: Localization 5
15: 2018/12/04: Localization 6
16: 2018/12/06: No lecture (Dies Academicus)
17: 2018/12/11: PPI 1 - sites / pairing (chalk)
18: 2018/12/13: PPI 2 - sites / PPI pairing (chalk)
19: 2018/12/18: PPI 3 - sites / DNA / RNA (Jia Jun Qiu)
20: 2018/12/20: No lecture
21-24: no lectures - winter break (2018/12/24 - 2019/01/06)
25: 2019/01/08: PPI 4 - PPI pairing
26: 2019/01/10: PPI 5 - PPI pairing 2
27: 2019/01/15: SNP effect 1 (chalk talk)
28: 2019/01/17: SNP effect 2
29: 2019/01/22: SNP effect 3
30: 2019/01/24: WRAP up 1
31: 2019/01/29: WRAP up 2
32: 2019/01/31: ?
33: 2019/02/05: ?
34: 2019/02/07: Examen (10:00-13:00, lecture room LMU physics)