title:  IV Protein interactions - sites 2

short title:  pp2_ppi_sites2

lecture:  Protein Prediction 2 (for Computational Biology) - Protein function
TUM winter semester
Videos: YouTube / www.rostlab.org

THANKS:

Special lectures:
- 11/29 Jonas Reeb - CAFA
- 12/01 Tatyana Goldberg - Localization
- 01/10 Jana Schmidt

No lecture:
- 11/01 All Saints
- 11/15 SVV Student representation
- 11/24 Thanksgiving
- 12/20-01/06 - no lecture Xmas

LAST lecture: January 19 (followed by 3 x wrap-up sessions)

Examen: February 07 time TBA
- Makeup: TBC: Apr 25 & Apr 26, 2017 - lecture time
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

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

Yeast-2-Hybrid (Y2H) Method

©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
Can we transfer binding through homology?

- Obviously, otherwise no value in model organisms...

\[
\text{similarity} \; > \; X
\]

\[
\text{similarity} \; > \; X
\]

\[
\text{similarity} \; > \; X
\]

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

More similar  

Worm  

Human

Less similar

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)

Worm  
Human

less similar  more similar

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

worm (C. Elegans)

More similar

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

![Diagram showing A, B, A', B' with similarity > X and A-B known experimentally, A'-B' inferred by homology.

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

© Wikipedia
Genome evolution
Genome evolution

Orthologs

Paralogs

Species C1

duplication

Species B1

Species B2

Species A

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

similarity > X

A

A'

A''

B

B'

B''

Worm

Human

© Sven Mika & Burkhard Rost (Columbia New York)
worm (C. Elegans)

Homology-Inference

HSSP

worm vs non-worm

Worm

Human

less similar

more similar

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

**Fruitfly (Drosophila Melanogaster)**

- ▲ drome vs drome
- □ drome vs non-drome

**worm (C. Elegans)**

- ▲ worm vs worm
- □ worm vs non-worm

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

Orthologs
Paralogs

Species C1

duplication

Species B1
Species B2

Species C2

Species A
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?
NOT homology-based

\[ \text{similarity} > X \]

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
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

₁ IMPLEMENT SIMPLE METHOD TO DO THIS
failed entirely: too many false positives

₂ REDUCE FALSE POSITIVES:

₂₁ 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!**

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

Eukaryotic Protein Sequence

Non membrane

Transmembrane

Intra-cellular

Secretory pathway

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 (IntSites, 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
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)

- 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)

extract interactions how?
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
CD-Hit / UniqueProt, e.g. 70% PIDE
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)

- procedure for P1:
  - randomly draw S
  - random draw
  - JS?
  - 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>Green</th>
<th>Yellow</th>
<th>Blue</th>
<th>Red</th>
<th>Pink</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal</td>
<td>909</td>
<td>65</td>
<td>3</td>
<td>-</td>
<td>21</td>
</tr>
<tr>
<td>Domain-domain</td>
<td>53</td>
<td>812</td>
<td>2</td>
<td>-</td>
<td>128</td>
</tr>
<tr>
<td>Homo-obligomer</td>
<td>-</td>
<td>2</td>
<td>925</td>
<td>-</td>
<td>12</td>
</tr>
<tr>
<td>Homo-oligomer</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1000</td>
<td>-</td>
</tr>
<tr>
<td>Hetero-obligomer</td>
<td>18</td>
<td>130</td>
<td>7</td>
<td>-</td>
<td>811</td>
</tr>
<tr>
<td>Hetero-oligomer</td>
<td>-</td>
<td>8</td>
<td>58</td>
<td>-</td>
<td>38</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)

Different interfaces = different physics?

HIV gp120 / CD4 / FAB

Develop method

- 1. PDB->Unique
- 2. parse heavy atoms <6.5 Ångstrøm (0.65 nm)
Each atom in the PDB belongs to a residue
Each residue belongs to a chain
Chains may have “breaks”
Map PDB to annotations

BLAST 10-10, >90% PIDE over >90% of length

chain A maps to SPa, chain B maps to SPb
if (SPa=SPb) assume A and B from same protein
else A and B from two different proteins
Develop method

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

1. parse heavy atoms <6.5 Ångström (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?

<table>
<thead>
<tr>
<th></th>
<th>strong</th>
<th>weak</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.9</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.4</td>
</tr>
</tbody>
</table>
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>G</td>
<td>GGGG</td>
<td>5 . . . . . . . .  .  5  .</td>
</tr>
<tr>
<td>Y</td>
<td>YYYY</td>
<td>. . . . . . . . . .  3  .</td>
</tr>
<tr>
<td>I</td>
<td>IIEE</td>
<td>. . . . . . . .  2 . . . .  .</td>
</tr>
<tr>
<td>Y</td>
<td>YYYY</td>
<td>. . . . . . . . . .  5 .</td>
</tr>
<tr>
<td>D</td>
<td>DDDD</td>
<td>. . . . . . . .  5 . . . .</td>
</tr>
<tr>
<td>P</td>
<td>PPPP</td>
<td>.  .  .  .  .  .  .  .  .</td>
</tr>
<tr>
<td>E</td>
<td>AEAA</td>
<td>.  .  3 . . . . . . . .  .</td>
</tr>
<tr>
<td>D</td>
<td>VVEE</td>
<td>. . . . . . . .  1 . . . .</td>
</tr>
<tr>
<td>G</td>
<td>GGGG</td>
<td>5 .  .  .  .  .  .  .</td>
</tr>
<tr>
<td>D</td>
<td>DDDD</td>
<td>. . . . . . . .  5 . . . .</td>
</tr>
<tr>
<td>P</td>
<td>PPPP</td>
<td>. . . . . . . . . . . . . .  .</td>
</tr>
<tr>
<td>D</td>
<td>DTDD</td>
<td>. . . . . . . .  4 . . . .</td>
</tr>
<tr>
<td>D</td>
<td>NQNN</td>
<td>. . . . . . . .  1 . .  1</td>
</tr>
<tr>
<td>G</td>
<td>GNGG</td>
<td>4 . . . . . . . .  .</td>
</tr>
<tr>
<td>V</td>
<td>VIVV</td>
<td>. . . . . . . .  .  .  4 . . . .</td>
</tr>
<tr>
<td>N</td>
<td>EPKK</td>
<td>. . . . . . . .  1 . . . . . . . .  .</td>
</tr>
<tr>
<td>P</td>
<td>PPPP</td>
<td>.  .  .  .  .  .  .  .  .</td>
</tr>
<tr>
<td>G</td>
<td>GGGG</td>
<td>5 . . . . . . . .  .  .</td>
</tr>
<tr>
<td>T</td>
<td>TTTT</td>
<td>.  .  .  .  .  .  .  .  .</td>
</tr>
<tr>
<td>D</td>
<td>EKSA</td>
<td>.  .  1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>F</td>
<td>FFFF</td>
<td>.  .  .  .  .  .  .  .  .</td>
</tr>
</tbody>
</table>

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

---


© Burkhard Rost ROSTLAB
What it takes to realize this as a server

PDB   SWISS-PROT   TrEMBL   HSSP

Sequence → PSI-BLAST → Filter
Much more complex system for function

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB</td>
<td>UCSD</td>
<td>daily</td>
</tr>
<tr>
<td>SWISS-PROT</td>
<td>Geneva</td>
<td>weekly</td>
</tr>
<tr>
<td>TrEMBL</td>
<td>EBI Hinxton UK</td>
<td>weekly</td>
</tr>
<tr>
<td>HSSP</td>
<td>EBI</td>
<td>monthly</td>
</tr>
<tr>
<td>FSSP</td>
<td>EBI</td>
<td>monthly</td>
</tr>
<tr>
<td>DIP</td>
<td>UCLA</td>
<td>manually</td>
</tr>
<tr>
<td>Pfam</td>
<td>Sanger Hinxton UK</td>
<td>manually</td>
</tr>
<tr>
<td>EVAdb</td>
<td>internal</td>
<td>daily</td>
</tr>
<tr>
<td>BIG</td>
<td>internal</td>
<td>weekly</td>
</tr>
<tr>
<td>PEP</td>
<td>internal</td>
<td>weekly</td>
</tr>
</tbody>
</table>

ISIS 2004
Few features

- Profile
- predicted 1D structure
  - secondary structure
  - solvent accessibility
  - membrane regions
  - disorder
- predicted aspects of function
Let neural networks figure it out ...
Random split not enough
avoid overlap training/cross-training vs. testing
Performance in predicting PPI sites
Method: neural network
Strength of prediction reflects reliability?

- **Strong**: P (in P=P) with 0.9
- **Weak**: N (not in P=P) with 0.1

- **Strong**: 0.6
- **Weak**: 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

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</td>
<td>GGGG</td>
<td>5 : : : : :</td>
</tr>
<tr>
<td>Y</td>
<td>YYYY</td>
<td>2 : : : : :</td>
</tr>
<tr>
<td>D</td>
<td>DDDD</td>
<td>: : : : : : : :</td>
</tr>
<tr>
<td>D</td>
<td>DDDD</td>
<td>: : : : : : : :</td>
</tr>
<tr>
<td>D</td>
<td>EKSA</td>
<td>: : : : : : : :</td>
</tr>
</tbody>
</table>

 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
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?

![Graph showing accuracy vs. coverage](image)

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

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?

Number of PPIs of one protein

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)
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
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

![Bar chart showing NORSnet](chart.png)

- **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)

Y Ofran, A Schlessinger & B Rost submitted
Lecture plan (PP2 function)

01: 2016/10/18: no lecture (makeup examen; PP last year)
02: 2016/10/20: no lecture (makeup examen; PP last year)
03: 2016/10/25: Welcome: who we are
04: 2016/10/27: Predicting effects of sequence variants (chalk talk)
05: 2016/11/01: No lecture (All Saints)
06: 2016/11/03: Intro function 1: concept of protein function
07: 2016/11/08: Intro function 2: homology-based inference of function
  CAFA homology-based prediction of function
08: 2016/11/10: Compute chemistry - enzymatic activity (slides from Janet Thornton)
09: 2016/11/15: No lecture: SVV (student reps)
10: 2016/11/17: Localization 1 (chalk)
12: 2016/11/24: No lecture: Thanksgiving
14: 2016/12/01: Tatyana Goldberg - Localization 4 PPI 1 - Protein-protein interactions - overview
15: 2016/12/06: Machine Learning
16: 2016/12/08: Localization 3 / PPI 1 - interaction sites
17: 2016/12/13: PPI 2 - interaction sites
18: 2016/12/15: PPI 3 - pairs
19: 2016/12/20: No lecture
20: 2016/12/22: No lecture
21-24: no lectures - winter break (2016/12/24 - 2017/01/06)
25: 2017/01/10: Jana Schmidt: Data mining in health care
28: 2017/01/12: SNP effect 1 - chalk
29: 2017/01/17: SNP effect 2
30: 2017/01/19: SNP effect 3
31: 2017/01/24: WRAP up 1
32: 2017/01/26: WRAP up 2
33: 2017/01/31: WRAP up 3
34: 2017/02/07: CONFIRMED examen (11:30-13:00)
35: 2017/02/09: No lecture

2017