Predict protein-protein interactions

pp2_ppi1

Protein Prediction 2 - Protein function
TUM Winter 2011/2012
Announcements

Videos: SciVe / www.rostlab.org

THANKS:
  Tim Karl + Julia Gerke

Special lectures:
  • Jan 25: Marco De Vivo (ISS Geneva)
  • Jan 27: Marco Punta (Pfam)

NO lectures (not final):

LAST lecture: Feb 3
Examen: Feb 8, 12:00 (likely this room)
  • Makeup: likely: Apr 19 - morning

CONTACT: Marlena Drabik assistant@rostlab.org

Monday January 23, 2012
Today: Secondary structure prediction 1

LAST YEAR
• Predicting effects of change

THIS WEEK
• Predicting effects of change
• Protein-protein interactions

NEXT WEEK
• Marco Punta (Pfam, Sanger, Cambridgeshire): Families
• Marco DeVito (Geneva, ISS): Drug design

2 WEEKs from now
• Protein-protein interactions
• Protein-DNA interactions
IV.
Predict protein interactions
IV.1 protein interactions
Protein-protein interactions (PPI): terminology
Different interfaces = different physics?

Protein association

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

ABCD are associated
Physical interaction NOT association

HIV gp120 / CD4 / FAB


Monday January 23, 2012
IV.2 protein interactions

PPI de novo?
1999: Want to predict protein-protein 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!
  - 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)
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-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 (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

Put it all together and predict binding partners
IV.3 protein interactions

PPI - data collection
Different interfaces = different physics?

HIV gp120 / CD4


Structure: Hendrickson lab
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: composition

Interface types differ in composition

Are these differences statistically significant?
Are these differences statistically significant?

- Chi-square test:
  - known problem: small data sets
  - here millions of points
Are these differences statistically significant?

**Chi-square test:**

- known problem: small data sets
- here millions of points

**all differences < 10^{-300}**

-> SIGNIFICANT
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 ...

Y Ofiran & B Rost 2005 submitted
Find-self test (statistical significance)

- Procedure for P1:
  - Randomly draw S
  - Randomly draw S
  - Repeat R times
  - Report pair with minimal JS
  - Perform procedure for P2 and P3

- Data set 1
- Data set 2
- Data set 3

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

<table>
<thead>
<tr>
<th></th>
<th>Internal</th>
<th>Domain-domain</th>
<th>Hom-obliger</th>
<th>Hom-oligomer</th>
<th>Hetero-obliger</th>
<th>Hetero-oligomer</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>Hom-obliger</td>
<td>-</td>
<td>2</td>
<td>925</td>
<td>-</td>
<td>12</td>
<td>61</td>
</tr>
<tr>
<td>Hom-oligomer</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1000</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Hetero-obliger</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.4 protein interactions
PPI - homology-based inference
Physical interaction **NOT** association

HIV gp120 / CD4 / FAB


Monday January 23, 2012
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
Can we transfer binding through homology?

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

A

B

\[ \text{similarity} > X \]

A'

B'

\[ \text{similarity} > X \]

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

©Sven Mika & Burkhard Rost (Columbia New York)
Inter and Intra-species the same?

similarity > X

Worm
Inter and Intra-species the same?

similarity > X

Worm

Human

©Sven Mika & Burkhard Rost (Columbia New York)
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

**worm (C. Elegans)**

- ▲ worm vs worm
- □ worm vs non-worm

**Fruitfly (Drosophila Melanogaster)**

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

---


Monday January 23, 2012
Why?
Why?
“Paralogs” conserve interactions
“orthologs” don’t?
Model organisms pose problems for protein-protein interactions
IV.5 protein interactions

PPI predict binding sites
Different interfaces = different physics?

HIV gp120 / CI

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

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)

How to collect data?
How to collect data?
PDB: numbers

- 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?
Remove redundancy

CD-Hit / UniqueProt, e.g. 70% PIDE
Histogram of protein family sizes
Histogram of protein family sizes

#

?
Histogram of protein family sizes

#
Zipf law / Power law

Wikipedia:
George Kingsley Zipf (Jan 2, 1902-Sep 25, 1950; Harvard)
Size of protein families (assumption)

Wikipedia:
George Kingsley Zipf (Jan 2, 1902-Sep 25, 1950; Harvard)

PDB

“nature”
Size of protein families: assumed relation

“nature”

PDB
Size of protein families: assumed relation

“nature”

PDB
Problems with redundancy reduction

- What approximates “nature”?

- How far to move “is” toward “should be”?
Develop method

1. PDB->Unique
Different interfaces = different physics

HIV gp120 / CD4 / FAB

Different interfaces = different physics?

HIV gp120 / CD4 / FAB


Monday January 23, 2012
Develop method

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


HIV gp120 / CD4 / FAB
PDB chains

- Each atom in the PDB belongs to a residue
- Each residue belongs to a chain
- Chains may have “breaks”
Different interfaces = different physics

HIV gp120 / CD4 / FAB

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
Remove redundancy for transient PP

PDB-PP

CD-Hit / UniqueProt, e.g. 70% PIDE
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
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?
Strength of prediction reflects reliability?

- **Strong:**
  - $P \text{ (in } P=P)$: 0.9
  - $N \text{ (not in } P=P)$: 0.1

- **Weak:**
  - $P \text{ (in } P=P)$: 0.6
  - $N \text{ (not in } P=P)$: 0.4
More complex system to predict structure

Sequence → PSI-BLAST → Filter

PROFsec → PROFacc

1999
Much more complex system for function
Much more complex system for function

PDB  SWISS-PROT  TrEMBL  HSSP  DIP  Pfam  EVAdb  BIG PEP

Sequence → PSI-BLAST

CHOP  PROFtmb  PHDhtm  LOC0
PROFsec  SEG  GLOBE
PROFacc  ISIS

Filter

2004
Much more complex system for function
Much more complex system for function

Sequence → PSI-BLAST → Filter

CHOP

PROFsec  SEG  GLOBE

PROFacc  PHDhtm  LOCo

ISIS

2004
Much more complex system for function

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Institution</th>
<th>Frequency</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
Publish not to perish
-
tie it up
Are we there yet?
Let neural networks figure it out ...
Let neural networks figure it out ...
Cross-validation: how?

CD-Hit / UniqueProt, e.g. 70% PIDE

PDB-PP
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?

Strong: 0.9
Weak: 0.6
PP interfaces predicted from sequence

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

Y Ofran & B Rost 2007 Bioinformatics e13-16
PPI hot spots?
residues that are essential for protein-protein interactions

- residue in the interface
- 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)

red: observed

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

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%

Functionally important residues - interactions sites


© Marco Punta & Yanay Ofran & Burkhard Rost (Columbia New York)
Find non-homologous competitive binder

![Diagram of molecular structures and annotations](image)

- **Find motif in family of known function**: USE as motif
- **Search for motif in new protein**: Transfer annotation of function
IV.5 protein interactions

PPI - hubs
Connect micro- and macro-level

macro level: networks
UP: more partners

micro level: residues
RIGHT: more hotspots
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!

- Non-hubs
- Party hubs
- Date hubs

macro: more partners
micro: more hotspots

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

- **macro:** more partners
- **micro:** more hotspots

- **Non-hubs**
- **Party hubs**
- **Date hubs**
More hotspots -> more party-hub like!

Y Ofran, A Schlessinger & B Rost submitted

Monday January 23, 2012
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

Monday January 23, 2012
More unstructured -> more date-hub like!

macro: more partners
micro: more hotspots

NORSnet

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)

ISIS unstructured

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

01: 2011/10/19: welcome: who we are
02: 2011/10/21: individualized medicine
03: 2011/10/26: Intro - function 1: concepts
04: 2011/10/28: ?
05: 2011/11/02: FVV (Student plenary meeting)
06: 2011/11/04: ?
07: 2011/11/09: Intro - function 2: homology
09: 2011/11/16: Andrea Schafferhans: Docking
10: 2011/11/18: Andrea Schafferhans: 3D function prediction
11: 2011/11/23: Localization 1
13: 2011/11/30: Marc Offman: Flexibility 1
14: 2011/12/02: Marc Offman: Flexibility 2
15: 2011/12/07: Bioinfo & Industry + Localization 3
16: 2011/12/09: Localization 3
17: 2011/12/14: skip
18: 2011/12/16: Localization 4: Tatyana Goldberg
19-20: no lectures (2011/12/21 - 2011/12/23)
21-24: no lectures - winter break (2011/12/21 - 2012/01/06)
25: 2012/01/11: SNP effect 1
26: 2012/01/13: SNP effect 2
27: 2012/01/18: SNP effect 3 / Protein-protein interaction 1
28: 2012/01/20: Protein-protein interaction 2
29: 2012/01/25: Marco De Vivo (ISS Geneva)
30: 2012/01/27: Marco Punta (Pfam)
31: 2012/02/01: Protein-DNA interaction 1
32: 2012/02/03: Protein-DNA interaction 2
ISMB 2012 Long Beach Jul 15-17

Honorary Chair
Sydney Brenner
UCSD, USA

ISCB Conferences Director
Steven Leard
Marketwhys Corp.

SC Co-Chairs
Terry Gaasterland
UCSD, USA

Rick Lathrop
UC Irvine, USA

Burkhard Rost
TUM Munich, Germany
& Columbia Univ USA

© Burkhard Rost (ISCB President)
### Key Submission Deadlines

<table>
<thead>
<tr>
<th>Event</th>
<th>Deadline</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Special Interest Groups</td>
<td>Oct 7</td>
<td>2011</td>
</tr>
<tr>
<td>Special Sessions</td>
<td>Oct 21</td>
<td></td>
</tr>
<tr>
<td>Proceedings Papers</td>
<td>Jan 13</td>
<td>2012</td>
</tr>
<tr>
<td>Workshops</td>
<td>Feb 10</td>
<td></td>
</tr>
<tr>
<td>Highlights Papers</td>
<td>Mar 2</td>
<td></td>
</tr>
<tr>
<td>Late Breaking</td>
<td>Mar 9</td>
<td></td>
</tr>
<tr>
<td>Posters</td>
<td>Mar 16</td>
<td></td>
</tr>
<tr>
<td>Travel Fellowship Applications</td>
<td>Apr 13</td>
<td></td>
</tr>
<tr>
<td>Technology Track</td>
<td>Apr 20</td>
<td></td>
</tr>
<tr>
<td>Late Posters</td>
<td>Apr 20</td>
<td></td>
</tr>
<tr>
<td>Student Council Symposium</td>
<td>May 4</td>
<td></td>
</tr>
</tbody>
</table>
## Key Submission Deadlines

<table>
<thead>
<tr>
<th>Category</th>
<th>Date</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Special Interest Groups</td>
<td>Oct 7</td>
<td>2011</td>
</tr>
<tr>
<td>Special Sessions</td>
<td>Oct 21</td>
<td></td>
</tr>
<tr>
<td>Proceedings Papers</td>
<td>Jan 13</td>
<td>2012</td>
</tr>
<tr>
<td>Workshops</td>
<td>Feb 10</td>
<td></td>
</tr>
<tr>
<td>Highlights Papers</td>
<td>Mar 2</td>
<td></td>
</tr>
<tr>
<td>Late Breaking</td>
<td>Mar 9</td>
<td></td>
</tr>
<tr>
<td>Posters</td>
<td>Mar 16</td>
<td></td>
</tr>
<tr>
<td>Travel Fellowship Applications</td>
<td>Apr 13</td>
<td></td>
</tr>
<tr>
<td>Technology Track</td>
<td>Apr 20</td>
<td></td>
</tr>
<tr>
<td>Late Posters</td>
<td>Apr 20</td>
<td></td>
</tr>
<tr>
<td>Student Council Symposium</td>
<td>May 4</td>
<td></td>
</tr>
</tbody>
</table>

**Google “ismb 2012” for details**
Scalalife Winterschool on Molecular Modelling
Feb. 27 – Mar. 02

Day 1: Lifescience Suite: Schrodinger Software
Day 2: HPC tools and resources for life sciences - LRZ
Day 3: Discussion in small groups and trip to Zugspitze
Day 4: GROMACS
Day 5: DALTON / DISCRETE / Visiting SuperMUC

APPLY @ http://www.lrz.de/services/schulung/kursanmeldung/