Prediction of Prot-Prot interactions

pp2_ppi3

Protein Prediction 2 - Protein function
TUM winter 2013
Announcements

Videos: SciVe / www.rostlab.org

THANKS:

Tim Karl + Jona Reeb

Special lectures:

- Oct 29: Tobias Hamp
- Nov 21: Tanya Goldberg
- Nov 28: Arthur Dong
- Dec 03+05: Marco De Vivo/Marco Punta
- Dec 17+19: Andrea Schafferhans

No lecture:

- Oct 10 Thu (Reformation)
- Nov 12 Tue (Student assembly)
- Dec 12 Thu (TUM Dies Academicus)

LAST lecture: Jan 30

Examen: Feb 4 (CompSci) and Feb 6 (CompBio) - likely this room

- Makeup: Apr 9 - morning
IV. (c) Predict protein interactions
IV.6 protein interactions

PPI - predictions
Different interfaces = different physics?

HIV gp120 / CD4 / FAB

Overview: systems
Protein-protein interaction networks

KEGG
tryptophan metabolism

http://www.genome.jp
KEGG metabolism

http://www.genome.jp

Friday January 17, 2014
KEGG
p53 signaling

http://www.genome.jp
Minoru Kanehisa

Director & Prof: Inst Chem Res, Kyoto Univ
Prof: Human Genome Center, Univ Tokyo

© http://kanehisa.kuicr.kyoto-u.ac.jp/People/kanehisa.html

KEGG
Kyoto Encyclopedia of Genes and Genomes
Minoru Kanehisa

Director & Prof: Inst Chem Res, Kyoto Univ
Prof: Human Genome Center, Univ Tokyo

© http://kanehisa.kuicr.kyoto-u.ac.jp/People/kanehisa.html
PPI-prediction

idea 1
Duplication event

Organism P

Organism C
Duplication event

Organism P

change:
one new process:

C1->C2

Organism C
**Duplication event**

Change: one new process:

C1->C2-> ... ->C9

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>C1</th>
<th>C2</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OrgP</strong></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td><strong>OrgC</strong></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>
Duplication event implies what?

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>C1</th>
<th>C2</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OrgP</strong></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td><strong>OrgC</strong></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

Friday January 17, 2014
In silico predictions of P=P interactions

Organism P

change:
one new process:

C1->C2-> ... ->C9
In silico predictions of P=P interactions

change:
one new process:
C1->C2-> ... ->C9

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
<th>C7</th>
<th>C8</th>
<th>C9</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OrgP</strong></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td><strong>OrgC</strong></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
In silico predictions of P=P interactions

(A) PROFILES:
- M Pellegrini, EM Marcotte, MJ Thompson, D Eisenberg and TO Yeates 1999 *PNAS* 96, 4285-4288
PPI-prediction
idea 2
Evolution of domains

Protein A - organism X

Protein B - organism X
Evolution of domains

Protein A - organism X

Protein B - organism X

Protein C - organism Y
In silico predictions of P=P interactions

(B) FUSION:

• T Gaasterland and MA Ragan 1998 *Microb Comp Genomics* 3, 177-192
• EM Marcotte, M Pellegrini, HL Ng, DW Rice, TO Yeates and D Eisenberg 1999 *Science* 285, 751-753
PPI-prediction
idea 3
Correlated mutations/Coevolution

contact in 3D

correlated mutation in 1D


© Debbie Marks - HMS © Burkhard Rost (TUM Munich)
Residue contacts accurately predicted

β2 adrenergic receptor

G-3-P transporter GlpT

doi: 10.1016/j.cell.2012.04.012

© Thomas Hopf - TUM & Debbie Marks - HMS
11 medically important TMH predicted

- **OCTN1**
  - Crohn’s disease,
  - rheumatoid arthritis

- **Adiponectin receptor 1**
  - diabetes,
  - obesity, cancer

- **MT-ND1**
  - LHON, MELAS, Alzheimer, Parkinson

Correlated mutations/Coevolution

Protein A

Protein B

correlated mutation in 1D

contact in 3D


© Debbie Marks - HMS © Burkhard Rost (TUM Munich)
In silico predictions of P=P interactions

(C) CORRELATED MUTATIONS:
- F Pazos and A Valencia 2002 *Proteins* 47, 219-227
PPI-prediction
idea 4
**In silico predictions of P=P interactions**

(A) **PROFILES:**
- M Pellegrini, EM Marcotte, MJ Thompson, D Eisenberg and TO Yeates *1999 PNAS* 96, 4285-4288

(B) **FUSION:**
- T Gaasterland and MA Ragan *1998 Microb Comp Genomics* 3, 177-192
- EM Marcotte, M Pellegrini, HL Ng, DW Rice, TO Yeates and D Eisenberg *1999 Science* 285, 751-753
turn concept around: NOT pairs of genes/all organism, but: presence/absence in alignment/family
Mirror tree: similarity of phylogenetic trees

Multiple alignment for protein 1
Multiple alignment for protein 2

Same set of species

Distance matrix 1
Distance matrix 2

High $r$ suggests an interaction

Juan et al. (2008). PNAS.
Mirror tree vs. phylogenetic profiles

Mirror tree

Phylogenetic profiles

Mirror tree more sophisticated

F Pazos & A Valencia (2001) Protein Engineering

© Ta-Tsen Soong, Columbia Univ
Mirror tree vs. phylogenetic profiles

Mirror tree

Phylogenetic profiles

- Mirror tree more sophisticated
Mirror tree vs. phylogenetic profiles

Mirror tree performs worse than phylogenetic profiles

F Pazos & A Valencia (2001) Protein Engineering
Mirror tree vs. phylogenetic profiles

Mirror tree performs worse than phylogenetic profiles
PPI-prediction
idea 5
Motifs indicative of interaction?

Protein A - organism X

Protein B - organism X

Proteins C+D - organism X or non-X

means: interact
In silico predictions of P=P interactions

MOTIFS:
• E Sprinzak & H Margalit 2001 *J Mol Biol* 311, 681-692
• SM Gomez & A Rzhetsky 2002 *Pac Symp Biocom* 413-24
In silico predictions of P=P interactions

(A) profiles:
- M Pellegrini, EM Marcotte, MJ Thompson, D Eisenberg and TO Yeates 1999 *PNAS* 96, 4285-4288

(B) fusion:
- T Gaasterland and MA Ragan 1998 *Microb Comp Genomics* 3, 177-192
- EM Marcotte, M Pellegrini, HL Ng, DW Rice, TO Yeates and D Eisenberg 1999 *Science* 285, 751-753

(C) correlated mutations:
- F Pazos and A Valencia 2002 *Proteins* 47, 219-227

(D) sequence signatures

MOTIFS:
- E Sprinzak & H Margalit 2001 *J Mol Biol* 311, 681-692
- SM Gomez & A Rzhetsky 2002 *Pac Symp Biocom* 413-24
Features commonly used for PPI prediction

Gene fusion

Homology (interolog)

Domain interaction

Functional similarity

Microarrays

Phylogenetic profile

Rhodes, et al. (2005) Nature Biotech
PPI-prediction
idea 6
Ta-Tsen Soong

here @ Columbia Univ, Manhattan West Side
now @ Cornell Medical, Manhattan East Side
cDNA microarrays measure gene expression in high-throughput (ht) manner

Cancer cells

Normal cells

RNA isolation

Reverse transcriptase labeling

Hybridization to microarray

mRNA

cDNA

Expression level readout

© Ta-Tsen Soong, Columbia Univ
Yeast two-hybrid system
- Interaction type: transient, binary
- Takes place in the nucleus
- Shortcomings: folding, localization, post-translational modification.

Affinity purification with mass spectrometry (AP-MS)
- Interaction type: protein complex membership
- Takes place in the native cellular environment
- Shortcomings: affinity tag interference, purification, sticky proteins, no details about pairwise binding.
Different interfaces = different physics?

HIV gp120 / CD4 / FAB

**Microarrays**

- **Large amount of data available**
  - Human: ~137,000 samples in GEO microarray database (Barrett, T. et al. 2007. NAR)
  - 18 organisms with > 1000 samples

- **mRNA level correlates with protein abundance (r= .57)** (Ghaemmaghami, et al. 2003. Nature)

- **PPI prediction from microarrays**
  - Correlation of expression patterns
    - Stable, permanent protein complexes
    - Transient, direct, physical PPIs
  - **Difficult to predict physical PPIs from microarray data**

R Jansen et al. & M Gerstein (2002) Genome Research

Microarray coexpression (Pearson correlation)
Experiments

☐ Yeast S. cerevisiae

☐ Interactions:
  • 5299 interactions from DIP (Salwinski, et al. 2004. NAR)

☐ Microarrays:
  • 349 microarrays from GEO database (Barrett, et al. 2007. NAR)
  • Remove noise and extract underlying biological processes

☐ Compare our protocol with correlation-based predictions
  • Cross validation
  • Genome wide analysis
Physical protein–protein interactions predicted from microarrays*

Microarray expression reveals functional associations

Association vs. Interaction

7 physical PPI: AB, BC, CD, DE, DF, EF, FG
7*6/2=21 associations
Microarray expression reveals functional associations

Most associated proteins are not in direct physical contact.

Our goal: predict physical interactions from microarray data

Two components of method

- PCA to group the microarray experiments (noise reduction)
- SVM to separate association and physical interaction
Step 1: PCA noise reduction

- **Remove noise and recover underlying biological processes**
  - **Principal Component Analysis (PCA)**
    - Statistical technique (projection method)
      - Liebermeister (2002) Bioinformatics
    - PCA components correspond to distinct biological processes
  - PCA component, *expression mode*, eigenarray
    - Ranked by importance (eigenvalue)
Step 2: SVM physical vs associate

Learn PPIs from PCA components with SVM

- Top N PCA components
- Protein features: $m_A$, $m_B$
- Protein pairwise features:
  - Outer-product: $F_{AB} = m_A \otimes m_B \oplus r_{AB}$
  - Concatenation: $F_{AB} = m_A \oplus m_B \oplus r_{AB}$
- Ranked by importance
- Kernel function
- Classify

Vapnik Statistical Learning Theory, 1998
SVM provided better prediction than correlation

- Implemented the correlation-based method as a Bayes model
- Bayes (correlation) performed slightly better than random (green vs. diagonal)
- Using fewer PCA components performed better than Bayes (e.g. SVM_{20} > Bayes)
- Performance proportional to number of PCA components until plateau ~150 (SVM_{150} > SVM_{50} > SVM_{20}).
- SVM performed best (SVM_{AllMA} > Bayes)

Table 1. AUC for inferring interactions

<table>
<thead>
<tr>
<th>Classifier</th>
<th>AUC (all)</th>
<th>AUC (FPR &lt; 0.1)</th>
<th>AUC (FPR &lt; 0.01)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM20</td>
<td>0.748</td>
<td>0.241</td>
<td>0.052</td>
</tr>
<tr>
<td>SVM50</td>
<td>0.765</td>
<td>0.277</td>
<td>0.063</td>
</tr>
<tr>
<td>SVM100</td>
<td>0.768</td>
<td>0.290</td>
<td>0.067</td>
</tr>
<tr>
<td>SVM150</td>
<td>0.766</td>
<td>0.289</td>
<td>0.079</td>
</tr>
<tr>
<td>SVM200</td>
<td>0.766</td>
<td>0.286</td>
<td>0.076</td>
</tr>
<tr>
<td>SVM250</td>
<td>0.758</td>
<td>0.278</td>
<td>0.074</td>
</tr>
<tr>
<td>SVM_{AllMA}</td>
<td>0.719</td>
<td>0.220</td>
<td>0.047</td>
</tr>
<tr>
<td>Bayesian model</td>
<td>0.630</td>
<td>0.157</td>
<td>0.039</td>
</tr>
</tbody>
</table>
SVM provided better prediction than correlation

- Implemented the correlation-based method as a Bayes model
- Bayes (correlation) performed slightly better than random (green vs. diagonal)
- Using fewer PCA components performed better than Bayes (e.g. SVM$_{20}$ > Bayes)
- Performance proportional to number of PCA components until plateau ~150 (SVM$_{150}$ > SVM$_{50}$ > SVM$_{20}$).
- SVM performed best (SVM$_{AllMA}$ > Bayes)
PCA components improve SVM

- Compared SVM performance with increasing PCA components (red) to using randomly selected microarrays (green) as input.
- PCA components provide a more distinct representation of gene activity.
Prediction score indicative of network distance


Dist(A,B)=1
Dist(A,C)=2
Predicted interaction score for all protein pairs in the DIP network and plotted against network distance.

SVM score is significantly more correlated with network distance than Bayes is (p<<.05).

Potential use of SVM score to help functional prediction in a network context.
Predictions confirmed by experimental annotations

- SVM in general have more predictions confirmed by BioGRID*.
- SVM also predicted other types of interactions (e.g. genetic)
- Big difference between two Affinity Purification methods.
Promising predictions by the SVM

- 8% of top predictions share specific Gene Ontology annotations suggesting biologically plausible interactions, while only 2% are expected by chance.

- Examples from literature:
  - **POB3_YEAST (YML06W) and CTK3_YEAST (YML11W)**
    - Both interact with RNA pol II and are involved in chromatin modulated transcription functions
    - Suggested role in regulation of FACT via the Ctk kinase complex
  
  - **SEC27_YEAST (YGL137W) and GCS1_YEAST (YDL226C)**
    - Implicated through E-MAP experiments
    - Sec27p is a coatamer subunit and is known to bind the di-lysine motif critical to retrograde transport of proteins from the Golgi to the ER.
    - Gcs1p contains the di-lysine motif and also acts as a mediator in the secretory pathway, suggesting a plausible interaction.
A→B→C→D : 
6 possible, 3 true
Microarray data can predict physical interactions

A->B->C->D : 6 possible, 3 true

T-t Soong, K Wrzeszczynski & B Rost 2008 Bioinformatics: 2608-14
PPI-prediction data perspective: integration
Integrating diverse data types

Gene fusion
Homology
Microarray
Functional similarity
Mirror tree

Integration
(naïve Bayes)

SVM-based protocol
Phylogenetic profiles
Conserved coexpression
Sequence domain
Subcellular localization

Text mining

Ta-Tsen Soong & B Rost, unpublished
© Ta-Tsen Soong, Thesis Defense (2009), Columbia Univ. © Burkhard Rost (TUM Munich)
Integrative PPI prediction

YEAST, FPR< .01

Area under ROC

R Nair & B Rost (2005) LocTree. JMB

© Ta-Tsen Soong, Columbia Univ

Friday January 17, 2014
Integrative PPI prediction

YEAST, FPR < 0.01

Area under ROC

HUMAN, FPR < 0.01

Area under ROC

R Nair & B Rost (2005) LocTree. JMB
Integrative PPI prediction

☐ all better than random (0.005)
☐ combination best
☐ major contributions: GO, Text mining, SVM
☐ at low FPR: homology, gene fusion, domain interaction

R Nair & B Rost (2005) LocTree. JMB

© Ta-Tsen Soong, Columbia Univ
# Data coverage

<table>
<thead>
<tr>
<th>Feature</th>
<th>Human</th>
<th>Yeast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence</td>
<td>20288 (100.0%)</td>
<td>6522 (100.0%)</td>
</tr>
<tr>
<td>Gene Ontology</td>
<td>7186 (35.4%)</td>
<td>3733 (57.2%)</td>
</tr>
<tr>
<td>Microarray $^1$</td>
<td>16433 (81.0%)</td>
<td>5823 (89.3%)</td>
</tr>
<tr>
<td>PFam domain</td>
<td>15956 (78.6%)</td>
<td>4363 (66.9%)</td>
</tr>
<tr>
<td>Subcellular localization $^2$-Pred.</td>
<td>19881 (98.0%)</td>
<td>6514 (99.9%)</td>
</tr>
<tr>
<td>Exp.</td>
<td>6639 (32.7%)</td>
<td>3506 (53.8%)</td>
</tr>
<tr>
<td>Text-mining $^3$</td>
<td>6061 (29.9%)</td>
<td>2401 (36.9%)</td>
</tr>
</tbody>
</table>

1 GPL570 for human, GPL90 for yeast
2 Predictions made with LocTree (Nair and Rost, 2005). Experimental annotations taken from SWISS-PROT
3 Annotations taken from the GeneWays database (Rzhetsky, et al. 2004)
Microarray CAN be used to predict physical PPIs

- SVM-based method significantly improves prediction of physical protein-protein interactions from microarrays

- Improvement originates from both sources:
  1. PCA component extraction
  2. SVM machine learning

- Prediction score reflects network distance and seems helpful for predicting function (GO-terms) in a network context

- Genome-wide predictions provide interactions worthy of biochemical validation.
Conserved mRNA coexpression

Similar in overall performance

- Yeast: AUC ~ 0.63
- Human: AUC ~ 0.60

Improvement for top predictions

1792 (27%) yeast and 13515 (67%) human proteins have orthologs with expression data in other organisms.
PPI-prediction

idea 7
Different interfaces = different physics?

HIV gp120 / CD4 / FAB

Using structure to predict PPI

Fig. 1
Figure 3. Models for the PPI formed between (A) PKD1 and PKCε, and (B) EF1δ and VHL using homology models and remote structural relationships. The same template complex of ubiquitin-conjugating enzyme E2D3 and ubiquitin (PDB code: 2fuh A and B chain, shown in blue and red respectively) was used in both cases. The structures of the PH domain of PKD1 and the GNE domain of EF1δ (shown in green and purple) are homology models from ModBase; the structure of a C1 domain of PKCε (yellow) is a homology model from SkyBase; the structure of VHL (cyan) is from PDB (1lm8 V chain). In each case, the relevant homology models are structurally superimposed on one of the two templates in the E2-ubiquitin complex.

Fig. 3
Figure 2. ROC curve (A) and Venn diagram (B) for PrePPI predictions and high-throughput (HT) experiments for yeast. HT experiments are labeled with the first author of the relevant publication (Table S4). The number of interactions in each set is given after the set label in the Venn diagram.


Fig. 2
IV.7 protein interactions

PPI - PiNat
PiNat (Protein Interaction Network analysis tool)

Y Ofран et al. & Rost 2006 Bioinformatics 22:e402-7
## Protein-protein interactions across compartments

<table>
<thead>
<tr>
<th></th>
<th>Extracellular</th>
<th>Cytosplasm</th>
<th>Organelles</th>
<th>Mitochondria</th>
<th>Nuclear</th>
<th>TM transmembrane</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extracellular</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>

© Burkhard Rost (TUM Munich)
PiNat (Protein Interaction Network analysis tool)

Y Ofran G Yachdav, E Mozes, T Soong, R Nair, B Rost al.
2006 Bioinformatics 15:22 e402-7

© Burkhard Rost (TUM Munich)
PiNat view of Alzheimers

Y Ofrran G Yachdav, E Mozes, T Soong, R Nair, B Rost al.
2006 Bioinformatics 15:22 e402-7
PiNat (Protein Interaction Network analysis tool)

Y Ofran G Yachdav, E Mozes, T Soong, R Nair, B Rost al.
2006 Bioinformatics 15:22 e402-7
Lecture plan (PP2 function)

01: 2013/10/15: no lecture
02: 2013/10/17: welcome: who we are
03: 2013/10/22: Intro - function 1: concepts
04: 2013/10/24: Intro - function 2: homology
05: 2013/10/29: Tobias Hamp: Homology-based prediction of function
06: 2013/10/31: no lecture Reformation
07: 2013/11/05: Tobias Hamp: Homology-based prediction of function 2
08: 2013/11/07: Intro - function 3: motifs
09: 2013/11/12: no lecture: SVV (student reps)
10: 2013/11/14: Localization 1
11: 2013/11/19: Localization 2
12: 2013/11/21: Localization 3 - Tatyana Goldberg
14: 2013/11/28: Localization 4 - Tatyana Goldberg
15: 2013/12/03: Marco De Vivo (ISS Genoa) - Drug Design
16: 2013/12/05: Marco Punta (Pfam, EBI) - Pfam
17: 2013/12/10: Exercise/Presentations
18: 2013/12/12: no lecture: Dies Academicus / Protein-protein interaction 1
19: 2013/12/17: Andrea Schafferhans: 3D function prediction
20: 2013/12/19: Andrea Schafferhans: Docking
21-24: no lectures - winter break (2013/12/23 - 2014/01/06)
25: 2014/01/07: Protein-protein interaction 1
26: 2014/01/09: Protein-protein interaction 2
27: 2014/01/14: Protein-protein interaction 3 / Protein-DNA/RNA interaction
28: 2014/01/16: Protein-DNA/RNA interaction 2
29: 2014/01/21: SNP effect 1
30: 2014/01/23: SNP effect 2
31: 2014/01/28: SNP effect 3
32: 2014/01/30: wrap-up
33: 2014/02/04: examen
34: 2014/02/06: no lecture