Predict PPI / Protein-DNA / GO

pp2_ppi2

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

Let it go. Let it out. Let it all unravel. Let it free and it can be. A path on which to travel.

Monday February 6, 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. (b) Predict protein interactions
IV.6 protein interactions

PPI - predictions
Protein-protein interaction networks

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

<table>
<thead>
<tr>
<th>Genome</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein 1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Protein 2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Protein 3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Protein 4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Protein 1: 10110
Protein 4: 10110

1 and 4 interact
In silico predictions of P=P interactions
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

![Diagram showing Genome A, Protein 1, Protein 2, Genome B, Protein 3]
In silico predictions of P=P interactions
In silico predictions of P=P interactions

(C) CORRELATED MUTATIONS:
- F Pazos and A Valencia 2002 *Proteins* 47, 219-227
Mirror tree: similarity of phylogenetic trees

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
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

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

- Mirror tree performs worse than phylogenetic profiles
In silico predictions of P=P interactions
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
Other sources with evidence for PPI
Features commonly used for PPI prediction

Gene fusion

Homology (interolog)

Domain interaction

Functional similarity

Microarrays

Phylogenetetic profile

Rhodes, et al. (2005) Nature Biotech
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, Columbia Univ  ROSTLAB.
Integrative PPI prediction

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

Monday February 6, 2012
# 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></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)
PPI through array data?
Microarray data

- cDNA microarrays measure gene expression in high-throughput (ht) manner.
High-throughput technologies

- **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.
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
Physical protein–protein interactions predicted from microarrays*

- 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

Microarray samples

Genes

PCA

Genes

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

$F_{AB} = m_A \otimes m_B \oplus r_{AB}$

Outer-product

$F_{AB} = m_A \oplus m_B \oplus r_{AB}$

Concatenation

Ranked by importance

Interaction

Non-interaction

Unknown pair

Kernel function

Classify

Vapnik *Statistical Learning Theory*, 1998

© Ta-Tsen Soong, Columbia Univ
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).
- A small number of PCA components performed better than Bayes (e.g. SVM$_{20}$ > Bayes).
- Performance increases with more input PCA components. Reaches the maximum at ~150 (SVM$_{150}$ > SVM$_{50}$ > SVM$_{20}$).
- SVM provided performance improvement (SVM$_{AllMA}$ > Bayes).

© Ta-Tsen Soong, Columbia Univ

Monday February 6, 2012
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).
- A small number of PCA components performed better than Bayes (e.g. $\text{SVM}_{20} > \text{Bayes}$).
- Performance increases with more input PCA components. Reaches the maximum at ~150 ($\text{SVM}_{150} > \text{SVM}_{50} > \text{SVM}_{20}$).
- SVM provided performance improvement ($\text{SVM}_{\text{AllMA}} > \text{Bayes}$).
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.
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.

© Ta-Tsen Soong, Columbia Univ
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

- **SEC27_YEAST (YGL137W) and GCS1_YEAST (YDL226C)**
  - 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
IV.7 protein interactions

PPI - PiNat
PiNat (Protein Interaction Network analysis tool)

Y Ofран et al. & Rost 2006 Bioinformatics 22:e402-7
Monday February 6, 2012
## Protein-protein interactions across compartments

<table>
<thead>
<tr>
<th></th>
<th>Extracellular</th>
<th>Cyttoplasm</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>Cyttoplasm</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>
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

Monday February 6, 2012
PiNat view of Alzheimers

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

Monday February 6, 2012
PiNat (Protein Interaction Network analysis tool)

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

Monday February 6, 2012
IV.8 protein interactions

Protein-DNA interactions
PPI interfaces use local segments

Datas protein-DNA interaction

- 291 protein-DNA complexes from PDB
- 250 chains bind DNA
- 46,000 residues

- Trevor Siggers / Barry Honig
Impressively accurate

Y Ofран & B Rost (2004) unpublished
Very accurate prediction of DNA binding

Y Ofran & B Rost (2004) *in preparation*
Very accurate prediction of DNA binding

Most predictions are discoveries!

![Graph showing fraction of DNA/RNA binding predictions that include known binding motifs](image)

Y Ofran & B Rost (2004) *in preparation*
Future DNA/RNA-binding

Consolidate Proteomes DNA/RNA DNA-binding and membrane insertion

Experimental verification of new motifs

Discover unknown DNA-binders in regulatory complexes:

- Transcription factor X
- Find all proteins Y implicated with X that:
  - not known to bind DNA/RNA
  - predicted by our method


Monday February 6, 2012
Most predictions new!

Increasing accuracy for subset

Y Ofran & B Rost (2004) unpublished
DNA/RNA motif-discovery engine

Y Ofirn, V Mysore, R Nair & B Rost (2004) *unpublished*
How many known motifs picked up?

Y Ofran & B Rost (2004) unpublished
How many new motifs discovered?

Y Ofran & B Rost (2004) unpublished
How many new motifs discovered?

Y Ofran & B Rost (2004) unpublished
CAFA data

a  Timeline

- January 18, 2011
  Submission deadline

- Prediction Phase
- Target Accumulation Phase

- September 15, 2010
  Sequences released (48,298)

- September 21, 2011
  Target set defined (762)

b  Target Counts

c  Functional Terms

CAFA: P Radivojac et al. & I Friedberg (2012) in submission
CAFA: P Radivojac et al. & I Friedberg (2012) in submission

Monday February 6, 2012
CAFA homology-based inference

CAFA: T Hamp et al. (2012) submitted
## CAFA ranking (homology only)

<table>
<thead>
<tr>
<th></th>
<th>BPO</th>
<th></th>
<th></th>
<th>MFO</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-20</td>
<td>Threshold</td>
<td>Leaf</td>
<td>Top-20</td>
<td>Threshold</td>
<td>Leaf</td>
</tr>
<tr>
<td>Priors</td>
<td>8</td>
<td>8</td>
<td>11</td>
<td>7</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>Priors’</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>BLAST</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>6</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>GOtcha</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>2</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>Student A</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>8</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Student A’</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Student B</td>
<td>11</td>
<td>11</td>
<td>7</td>
<td>11</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>Student B’</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Student C</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>9</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Student C’</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>MetaStudent’</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

CAFA: T Hamp et al. (2012) submitted

Monday February 6, 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
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
## Key Submission Deadlines

<table>
<thead>
<tr>
<th>Event</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>
<tr>
<td>Event</td>
<td>Date</td>
<td></td>
</tr>
<tr>
<td>--------------------------------------------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td>Special Interest Groups</td>
<td>Oct 7 2011</td>
<td></td>
</tr>
<tr>
<td>Special Sessions</td>
<td>Oct 21</td>
<td></td>
</tr>
<tr>
<td>Proceedings Papers</td>
<td>Jan 13 2012</td>
<td></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>