IV: Protein-protein interaction pairs

short title: cb2_ppi_pairs1

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

THANKS:

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 24 (followed by 2 wrap-up sessions)

Examen: Feb 07 10:00-13:00, LMU physics
- Makeup: no retake

CONTACT: teaching@rostlab.org

Dmitrij Nechaev

Let it go. Let it out.
Let it all unravel.
Let it free and let it be.
A path on which to travel.
RECAP
predict interaction sites
Different interfaces = different physics! because they can be predicted

HIV gp120 / CD4 / FAB

Interface types differ in composition

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)

Y Ofran & B Rost 2003 JMB 325:377-87
statistical
significance

NOT

scientific significance
Correlation is not causation

Number of babies born per year

Data from Lower Saxony, Germany

Correlation is not causation

Number of babies born per year

Number of storks per year

Data from Lower Saxony, Germany

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

Accuracy (correctly predicted/predicted)

Coverage (correctly predicted/observed)

Sequence only

Y Ofran & B Rost 2007 Bioinformatics e13-16
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

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

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

  - red: observed
  - purple: predicted

- structure:

enough to publish?
Hot spots reliably predicted from sequence!

hottest of hot = no error!

worst: ~60% right

IV. (c)

Predict protein interactions
protein interaction sites
Protein-nucleotide
PPI interfaces use local segments

IV.8 protein interactions

Protein-DNA interactions
Protein-DNA: Background
DNA helix

- **Major groove**
- **Minor groove**
Protein-DNA binding

Lambda repressor protein interacting with lambda operator

(PDBid 1lmb LJ Beamer & CO Pabo (1992) JMB 227:177-96 )

© Wikipedia
Nucleosome/Histone

DNA (orange) interacting with nucleosome (blue)
Acta Crystallogr D 56:1513-34
© Wikipedia

DNA (orange) interacting with histone H2AFJ
PDBid 1aoi: K Luger et al (1997)
Nature 389:251-60
© Wikipedia
Leucine zipper
PDBid 1fos: JN Glover & SC Harrison
© Wikipedia

Zinc finger
PDBid 1a1l: M Elrod-Erickson et al
© Wikipedia
Census: protein-DNA binding
How many human proteins bind DNA?
How many human proteins bind DNA?

☐ not fully clear
☐ two recent (consistent) views:
  >1,391 transcription factors

~1600
Prediction of protein-DNA interaction from sequence
NLS motifs
Shuttle into the nucleus

Cytoplasm

Nucleus

NLS

M9

Importin

Transportin
Using NLS to bind DNA

## DNA-binding predictions in proteomes

<table>
<thead>
<tr>
<th>Genome</th>
<th>Nprot</th>
<th>Nprot bind-DNA predicted</th>
<th>Nprot bind-DNA known</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>13933</td>
<td>419</td>
<td>141</td>
</tr>
<tr>
<td>Drosophila</td>
<td>14219</td>
<td>300</td>
<td>37</td>
</tr>
<tr>
<td>C. elegans</td>
<td>16232</td>
<td>251</td>
<td>10</td>
</tr>
<tr>
<td>Yeast</td>
<td>6307</td>
<td>67</td>
<td>10</td>
</tr>
<tr>
<td>E. coli</td>
<td>4286</td>
<td>13</td>
<td>3</td>
</tr>
</tbody>
</table>

DNA sites (now history?)
DNA sites: data

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

- Trevor Siggers / Barry Honig

DNAsites: Impressively accurate

Very accurate prediction of DNA binding

what next?
no, we need to check overlap with existing methods and entire genomes
Most predictions are discoveries!

Y Ofran & B Rost (2004) unpublished
Most predictions new!

Increasing accuracy for subset

Y Ofran & B Rost (2004) unpublished
now enough?
no, we need to cluster and estimate the number of motifs discovered
DNAsites: next generations
Improve DNAsites

- more important input information
- larger data set
- DNAsites2:

Michael Menden Master Thesis
DNAsites2 improves over DNAsites

DNAsites2:

Michael Menden Master Thesis

significantly better per-residue than existing methods
DNA sites 2: but

- Poor performance on level of per-protein
- NEVER 100% found
Improve DNAsites2

- more important input information
- larger data set
- handle “disastrous” per-protein prediction

DNAsites3:

Peter Hoenigschmid Master Thesis

© Burkhard Rost
SomeNA: no free lunch

SomeNA:

Peter Hoenigschmid Master Thesis

first results:

• better per-residue with leaving negative cases out
• better per-protein with using them

• (i.e. there is no such thing as a free lunch)
SomeNA: performance per-residue

P Hoenigschmid master thesis
SomeNA: performance per-protein

DNA All - Protein Based - Binding

P Hoenigschmid master thesis
Other methods predicting DNA-protein interaction from sequence
Predict DNA-protein binding from sequence

- **DNAsites:**

- **DNAsites2:**
  Michael Menden, Shaila Roessle & B Rost, unpublished

- **SomeNA:**
  P Hoenigschmid & B Rost, unpublished

- **DNABindR:**

- **BindN & BindN+:**

- **DP-Bind:**

- **DBS-Pred:**
  S Ahmad & A Sarai (2005) BMC Bioinformatics 6:33
IV.9 protein interactions

Protein-RNA interactions
RNA hairpin

hairpin loop from mRNA © Wikipedia
50S subunit of ribosome (PDB 3cc2)

23S large ribosomal subunit
G Blaha et al & TA Steitz (2008) JMB
379:505-19
PDBid 3cc2
© Wikipedia
RNA secondary structure

T. thermophila telomerase RNA
Using NLS to bind RNA?

Protein-RNA prediction methods

Recent review:

RR Walia, C Caragea, BA Lewis, F Towfic, M Terribilini, Y El-Manzalawy, D Dobbs & V Honavar (2012)
Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art.
BMC Bioinformatics 13:89
Protein-RNA prediction methods

RR Walia et al. (2012) BMC Bioinformatics 13:8
SomeNA: predicts RNA binding

Peter Hoenigschmid Master Thesis

first results:
• worse than DNA prediction?
• data set very diverse
Protein-RNA prediction: SomeNA

P Hoenigschmid master thesis
Protein-RNA prediction: SomeNA

DNA Only - Residue Based - Binding

RNA All - Residue Based - Binding

P Hoenigschmid master thesis
Protein-RNA prediction: SomeNA

P Hoenigschmid master thesis
Protein-RNA prediction: SomeNA

DNA All - Protein Based - Binding

RNA All - Protein Based - Binding

P Hoenigschmid master thesis
IV.9 protein interactions

PPI - pair predictions from sequence: Challenges
Prediction is the acid test for understanding - what about machine learning?
Machine learning = black magic?
Let neural networks figure it out ...
do NOT choose patterns at random, instead:
EACH part of test exactly once!
WEKA-like cross-validation
3-way cross-validation

Train

Cross-Train

Test
Family clustering

No two from same group in train & test|cross-train
Still not enough: exploit “prerelease” data (latest/hottest)
Prediction of Protein Secondary Structure at Better than 70% Accuracy

Burkhard Rost and Chris Sander

European Molecular Biology Laboratory
Meyerhofstraße 1, D-6900 Heidelberg, Germany

(Received 1 February 1993; accepted 13 April 1993)
Hot spots reliably predicted from sequence!

hottest of hot = no error!

worst: ~60% right

Now enough?

© Wikipedia
## Results from cross-validation

<table>
<thead>
<tr>
<th>Method 1 ({features1}, random forest)</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 2 ({features1}, SVM)</td>
<td>35 %</td>
</tr>
<tr>
<td>Method 3 ({features2}, SVM)</td>
<td>37 %</td>
</tr>
</tbody>
</table>
Results from cross-validation

<table>
<thead>
<tr>
<th>Method</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1 ({features1}, random forest)</td>
<td>34 %</td>
</tr>
<tr>
<td>Method 2 ({features1}, SVM)</td>
<td>35 %</td>
</tr>
<tr>
<td>Method 3 ({features2}, SVM)</td>
<td>37 %</td>
</tr>
</tbody>
</table>

-> Method 3 is best and performs at 37%

CONCLUSION correct?
Importantly missing

☐ Background
  • how good is random?
  • how good are best state-of-the-art methods?
  • tested on same data set?

☐ Error estimates: ±x, e.g. rule-of-thumb
  standard error = σ/√N
### Results from cross-validation

<table>
<thead>
<tr>
<th>Method</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>32±1</td>
</tr>
<tr>
<td>Method 1</td>
<td>34±1</td>
</tr>
<tr>
<td>Method 2</td>
<td>35±1</td>
</tr>
<tr>
<td>Method 3</td>
<td>37±1</td>
</tr>
</tbody>
</table>

-> Method 3 is best and performs at 37%

**CONCLUSION correct?**
xxbr

CONTINUE HERE
All done right. We can predict. Do we now understand?
Machine learning = black magic?

Machine learning
NOT black magic

Misunderstanding caused by lack of careful testing
Strength of prediction reflects reliability?
Machine learning are black boxes NOT because we cannot write the rules but because we do NOT understand them once written!
Machine-learning solution match the complexity of the problem
Challenges for predicting pairs of protein-protein interactions (PPIs)
Now predict protein-protein interactions?

HIV gp120 / CD4 / FAB

Predict PPI A-B through some sequence-derived features from PredictProtein
PPI partners
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, 2005-7)
- predict residues in protein-substrate interfaces (active)
- predict protein domains/improve alignments (2005-2008)

Put it all together & predict binding partners
Predict subcellular localization: **LOCtree 2: 18 classes!**

**Eukaryotic Protein Sequence**

- **SVM**
  - **Non membrane**
    - **SVM**
      - **Intra-cellular**
        - **SVM**
          - **ER**
            - **EXT**
            - **GOLGI**
          - **SVM**
            - **VAC**
            - **CYT**
            - **NUC**
            - **PLAS**
            - **GOLGI**
          - **SVM**
            - **NUC**
            - **PERO**
            - **MITO**
            - **CHLO**
            - **PLAST**
    - **SVM**
      - **SECRETORY PATHWAY**
      - **SVM**
        - **ER**
        - **NUC**
        - **CHLO**
        - **MITO**
      - **SVM**
        - **VAC**
  - **SVM**
    - **Intra-cellular**
      - **SVM**
        - **Ext**
        - **GOLGI**
      - **SVM**
        - **VAC**
        - **CYT**
        - **NUC**
        - **PLAS**
        - **GOLGI**
      - **SVM**
        - **NUC**
        - **PERO**
        - **MITO**
        - **CHLO**
        - **PLAST**
      - **SVM**
        - **VAC**

**T Goldberg, T Hamp & B Rost (2012) Bioinformatics 28:i458-i465**

**Tatyana Goldberg**

**Tobias Hamp**
PPI challenge machine learning MUCH more
PPI pair prediction:
Challenge 1 – redundancy reduction
Family clustering

No two from same group in train & test|cross-train
Family clustering

No two from same group in train & test|cross-train
Family clustering

No two from same group in train & test|cross-train
PPI sampling needs to consider proteins

- **Case 1**: both used before: i.e. training contained $A \land B$ 
  NOT interaction $AB$

- **Case 2**: either used for training 
  i.e. train on $A \mid B$

- **Case 3**: neither $A$ nor $B$ used before

---

Yungki Park 
SUNY Buffalo

Edward Marcotte 
Univ Texas Austin

**Y Park & EM Marcotte** (2012) 
*Nature Meth* 9: 1134-1136
PPI sampling needs to consider proteins

- **case 1**: both used before:
  i.e. training contained $A \land B$
  NOT interaction $AB$

- **case 2**: either used for training
  i.e. train on $A \mid B$

- **case 3**: neither $A$ nor $B$ used before

---


Yungki Park
SUNY Buffalo

Edward Marcotte
Univ Texas Austin

© Burkhard Rost
ROSTLAB. TUM
Reduced performance for new proteins


Sequence similarity only for PPiS, i.e. positives enough?
Literature: nobody ever considered sequence-similarity between negatives

→

not relevant?
Similarity between negatives matters!

“Pseudo-improvement” through overlap between negatives

A

△Precision (%) vs. Recall (%)

we HAVE to also consider negative PPiS
Redundancy reduction for PPI networks

protein
sequence similar

training positive (interaction)
training negative (non-interacting)

Non-redundant set

Redundant

Non-redundant

○ Protein

Positive training PPI

PPI pair prediction:

Challenge 2 - data
We do not have enough experimental data

-> take all we have?
How much data is needed?
double data: improves -> need more than half
Using more data helps C1 results for C1 (AB in training)


Using all good data bad for C3

results for C3 (AB not in training)

Cross-validation challenge squared for PPPIs
PPI pair prediction:

Challenge 3 - how to do it best?
Profile-kernel SVM
PPI from sequence through SVM profile kernel

C1 proteins have known PPIs

C3 not PPI known


Tasmanian 2015

T Hamp & B Rost (2015) Bioinformatics 31:1945-50
IV.11 protein interactions

PPI - pair predictions from expression arrays
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

Microarray data

- RNA isolation
- Reverse transcriptase labeling
- Hybridization to microarray
- Expression level readout

© Ta-Tsen Soong, Columbia Univ
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.
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 ($r = 0.57$) (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
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

7 physical PPI:
AB, BC, CD, DE, DF, EF,
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: noise reduction
Observations and hypotheses

UNOBSERVED
- Ribosome biogenesis
- Sulfur amino-acid metabolism
- Cell cycle

Genes

OBSERVED
by microarray measurement

Heat shock condition

IS Lee et al. (2003) Genome Biology
Step 1: PCA noise reduction

- Remove noise and recover underlying biological processes
  - Principal Component Analysis (PCA)
    - Statistical technique (projection method)
      - Liebermeister (2002) Bioinformatics
  
  PCA component, *expression mode*, eigenarray

- PCA components correspond to distinct biological processes

© Ta-Tsen Soong, Columbia Univ
Step 2: Machine learning to separate association from physical interaction
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

- Interaction
- Non-interaction
- Unknown pair

Vapnik *Statistical Learning Theory*, 1998
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\textsubscript{20} > Bayes)
Performance proportional to number of PCA components until plateau \( \sim 150 \) (SVM\textsubscript{150} > SVM\textsubscript{50} > SVM\textsubscript{20}).
SVM performed best (SVM\textsubscript{ALLMA} > Bayes)

Table 1. AUC for inferring interactions\textsuperscript{a}

<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>SVM\textsubscript{20}</td>
<td>0.748</td>
<td>0.241</td>
<td>0.052</td>
</tr>
<tr>
<td>SVM\textsubscript{50}</td>
<td>0.765</td>
<td>0.277</td>
<td>0.063</td>
</tr>
<tr>
<td>SVM\textsubscript{100}</td>
<td>0.768</td>
<td>0.290</td>
<td>0.067</td>
</tr>
<tr>
<td>SVM\textsubscript{150}</td>
<td>0.766</td>
<td>0.289</td>
<td>0.079</td>
</tr>
<tr>
<td>SVM\textsubscript{200}</td>
<td>0.766</td>
<td>0.286</td>
<td>0.076</td>
</tr>
<tr>
<td>SVM\textsubscript{250}</td>
<td>0.758</td>
<td>0.278</td>
<td>0.074</td>
</tr>
<tr>
<td>SVM\textsubscript{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>

\textsuperscript{a} SVM provided better prediction than correlation

© Ta-Tsen Soong, Columbia Univ
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.

A. FPR<0.05

B. Entire ROC
Prediction score indicative of network distance


Dist(A,B) = 1
Dist(A,C) = 2

GO semantic similarity

Network distance

© Ta-Tsen Soong, Columbia Univ
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.

© Ta-Tsen Soong, Columbia Univ
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 coatmer 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.
Microarray data can predict physical interactions

TT Soong, KO Wrzeszczynski & B Rost (2008) Bioinformatics 15:2608-14
IV.12 protein interactions

PPI - pair predictions
data perspective: integration
Integrating diverse data types

Homology

Microarray

Functional similarity

Mirror tree

Gene fusion

SVM-based protocol

Phylogenetic profiles

Conserved coexpression

Sequence domain

Subcellular localization

Integration (naïve Bayes)

Text mining

Ta-Tsen Soong & B Rost, unpublished

© Ta-Tsen Soong, Columbia Univ
Integrative PPI prediction

![Graph showing Area under ROC for different methods]

- R Nair & B Rost (2005) LocTree. JMB

© Ta-Tsen Soong, Columbia Univ
Integrative PPI prediction
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
## 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&lt;sup&gt;1&lt;/sup&gt;</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&lt;sup&gt;2&lt;/sup&gt; - Pred.</td>
<td>19881 (98.0%)</td>
<td>6514 (99.9%)</td>
</tr>
<tr>
<td></td>
<td>Exp.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6639 (32.7%)</td>
<td>3506 (53.8%)</td>
</tr>
<tr>
<td>Text-mining&lt;sup&gt;3&lt;/sup&gt;</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.
IV.13 protein interactions

PPI - PiNat
PiNat (Protein Interaction Network analysis tool)

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

<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>
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
PiNat view of Alzheimers

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

Q9P2H0

ADD
PiNat (Protein Interaction Network analysis tool)

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

Christian Dallago
TUM/Harvard Univ

http://cell.dallago.us (hopefully soon: http://cell.map )
For suggestions https://github.com/sacdallago/cellmap/issues

Christian Dallago
Summary

Many existing tools visualize protein-protein interaction (PPI) networks. CellMap adds important novelty as a prototype visualizing the PPI network in the context of subcellular localization, i.e. the location in a cell in which a PPI happens. Users can upload images of cells and define areas of interest against which the PPIs for selected proteins are displayed. The visualizer and server are written in JavaScript, making CellMap easy to customize and to extend by researchers and developers, and we offer a running version with human protein data from [1] and [2].
A selected set of proteins as coloured dots, placed in annotated localizations as per [1]. The protein-centric PPI network drawn by protein with UniProt ID Q13671 and its interacting partners, and scores on lines that indicate likelihood of interaction according to [2].
**CELLMAP: WHAT MAKES IT UNIQUE**

**Localization name determines color**

\[ f: \text{hash(s)} \rightarrow \text{hsl}(239,100\%, 40\%) \]

- extracellular \rightarrow \text{hsl}(239,100\%, 40\%)
- cytoplasm \rightarrow \text{hsl}(14,100\%, 40\%)

A hashing function projects string names of protein localizations into color space. This allows the tool to be extensible and provides deterministic coloring.

**Visualize Protein-Protein Networks constrained by Localization**

View protein-protein interaction networks constrained by protein localization. A label is displayed on the line connecting two proteins indicating the likelihood of interaction.

**Use your own images and define regions**

Upload images of cells and start defining regions of interest (organelles, etc.). Reduce clutter by defining only interesting regions and visualize sub-PPIs for proteins of interest in regions of interest.
CELLMAP: BETTER FOR HAIRBALLS

**CELLNetVis [3]**
- Visualizes PPI networks constrained by protein localization
- Uses force-based layout
- Allows exploration of interactions
- Fixed cell structure
- Doesn’t come with out of the box data

**Cytoscape [4]**
- Quickly renders graphs using different layouts
- Subject to hairball effect
- No localization constraint
- Hairball effect
- Must download application on computer

**Cytoscape.JS [4]**
- Allows visualization in browser
- Allows application of different layouts
- No localization constraint
- Hairball effect
- Big networks render slowly with certain layouts
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: No lecture
26: 2019/01/10: PPI 4 - sites: DNA / RNA (Jia Jun Qiu) + PPI pairing 1
27: 2019/01/15: PPI 5 - PPI pairing 2
28: 2019/01/17: SNV effect 1 (chalk talk)
29: 2019/01/22: SNV effect 2
30: 2019/01/24: SNV effect 3
31: 2019/01/29: WRAP up 1
32: 2019/01/31: WRAP up 2
33: 2019/02/05: ?
34: 2019/02/07: Examen (10:00-13:00, lecture room LMU physics)