Announcements PP2

Videos: YouTube / www.rostlab.org

THANKS:

Special lectures:
- (TBC)

No lecture:
- 11/07 Department event
- 11/09 Department event
- 11/16 Department event
- 11/23 Thanksgiving
- 12/19-01/06 - no lecture Xmas

LAST lecture: Jan 18 (followed by 3 x wrap-up sessions)

Examen: TBC: Feb 01 most likely 12:30-14:00
Makeup: TBC: Apr 10 & Apr 12, 2018 - lecture time
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

Correlation is not causation


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

Y Ofran & B Rost 2007 Bioinformatics e13-16
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>:</td>
<td>:</td>
<td>GSAPD NTEKQ CVHIR LMYFW</td>
</tr>
<tr>
<td>G</td>
<td>GGGG</td>
<td>5 . . . . . . . . . . 5 .</td>
</tr>
<tr>
<td>Y</td>
<td>YYYY</td>
<td>. . . . . . . . . . . 2 . 3 .</td>
</tr>
<tr>
<td>I</td>
<td>IIEE</td>
<td>. . . . . . . . . . . .</td>
</tr>
<tr>
<td>Y</td>
<td>YYYY</td>
<td>. . . . . . . . . . . .</td>
</tr>
<tr>
<td>D</td>
<td>DDDD</td>
<td>. . . . . . . . . . . 5 .</td>
</tr>
<tr>
<td>P</td>
<td>PPPP</td>
<td>. . . . . . . . . . . 5 .</td>
</tr>
<tr>
<td>E</td>
<td>AEAA</td>
<td>. . . . . . . . . . . 3 . 2 .</td>
</tr>
<tr>
<td>D</td>
<td>VVEE</td>
<td>. . . . . . . . . . . 1 . 2 . 2 .</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>. . . . . . . . . . . 5 .</td>
</tr>
<tr>
<td>D</td>
<td>DTDD</td>
<td>. . . . . . . . . . . 4 . 1 .</td>
</tr>
<tr>
<td>D</td>
<td>NQNN</td>
<td>. . . . . . . . . . . 1 . 3 . 1 .</td>
</tr>
<tr>
<td>G</td>
<td>GNGG</td>
<td>. . . . . . . . . . . 4 . 1 .</td>
</tr>
<tr>
<td>V</td>
<td>VIVV</td>
<td>. . . . . . . . . . . 4 . 1 .</td>
</tr>
<tr>
<td>N</td>
<td>EPPK</td>
<td>. . . . . . . . . . . 1 . 1 . 2 .</td>
</tr>
<tr>
<td>P</td>
<td>PPPP</td>
<td>. . . . . . . . . . . 5 .</td>
</tr>
<tr>
<td>G</td>
<td>GGGG</td>
<td>. . . . . . . . . . . 5 .</td>
</tr>
<tr>
<td>T</td>
<td>TTTT</td>
<td>. . . . . . . . . . . 5 .</td>
</tr>
<tr>
<td>D</td>
<td>EKSA</td>
<td>. . . . . . . . . . . 1 . 1 . 1 . 1 .</td>
</tr>
<tr>
<td>F</td>
<td>FFFF</td>
<td>. . . . . . . . . . . . . 5 .</td>
</tr>
</tbody>
</table>

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

---


© Burkhard Rost
PP interfaces predicted from sequence

Y Ofran & B Rost 2007 Bioinformatics e13-16
PPI 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

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

- Structure:
enough to publish?
Hot spots reliably predicted from sequence!

IV. (c)
Predict protein interactions
Overview: systems
Protein-protein interaction networks

KEGG: tryptophan metabolism

http://www.genome.jp
KEGG p53 signaling

http://www.genome.jp
IV.9 protein interactions

PPI - pair predictions from co-expression data
Recap: physical protein-protein interactions NOT associations!
Protein-protein interactions = Physical interactions NOT associations

HIV gp120 / CD4 / FAB
Protein association

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

ABCD are associated
Can we predict PPIs from expression data?
cDNA microarrays measure gene expression in high-throughput manner.

Microarray data

Cancer cells

Normal cells

RNA isolation

Reverse transcriptase labeling

Hybridization to microarray

mRNA

cDNA

Expression level readout

© Ta-Tsen Soong, Columbia Univ

© Burkhard Rost
Predict PPI from co-expression

Evidence that we CAN do it?
Predict PPI from co-expression

- Evidence that we CAN do it?
- Many publications but all target implicitly or explicitly associations, NOT physical transient protein-protein interactions (just a few exceptions)
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.
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 ($r = .57$) (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

Microarray coexpression (Pearson correlation)

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

Association vs. Interaction
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
- cellular processes
- Ribosome biogenesis
- Sulfur amino acid metabolism
- Cell cycle

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 components correspond to distinct biological processes

![PCA diagram](image)

- PCA components correspond to distinct biological processes
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
  - $F_{AD} = m_A \otimes m_B \oplus r_{AB}$ (Outer-product)
  - $F_{AB} = m_A \oplus m_B \otimes r_{AB}$ (Concatenation)

Genes ranked by importance

Kernel function

Classify

- Interaction
- Non-interaction
- Unknown pair

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. $\text{SVM}_{20} > \text{Bayes}$)
- Performance proportional to number of PCA components until plateau $\sim 150$ ($\text{SVM}_{150} > \text{SVM}_{50} > \text{SVM}_{20}$).
- SVM performed best ($\text{SVM}_{\text{ALLMA}} > \text{Bayes}$)

![Graph A and B](image)

**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>0.083</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>0.636</td>
<td>0.157</td>
<td>0.039</td>
</tr>
</tbody>
</table>
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

© Ta-Tsen Soong, Columbia Univ
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 different Purification methods.
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.
Microarray data can predict physical interactions
IV.10 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

Train

Test
3-way cross-validation
Family clustering

No two from same group in train & test|cross-train
Still not enough: exploit “prerelease” data (latest/hottest)
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:

Y Ofran & B Rost 2007 *PLoS CB* 3:e119
Hot spots reliably predicted from sequence!

hottest of hot = no error!

worst: ~60% right

Now enough?
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 1</th>
<th>Method 2</th>
<th>Method 3</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>({features1}, random forest)</td>
<td>({features1}, SVM)</td>
<td>({features2}, SVM)</td>
<td>34 %</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: ±σ, 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?**
All done right.
We can predict.
Do we now understand?
Common misunderstanding: Machine learning = black magic

Misunderstanding caused by lack of careful testing
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
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!**

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

- **SVM**
- **Non membrane**
- **Intra-cellular**
- **Secretory pathway**
- **Transmembrane**

**Eukaryotic Protein Sequence**

- SVM
- SVM
- SVM
- SVM
- SVM
- SVM
- SVM
- SVM
- SVM
- SVM
- SVM
- SVM

- EXT
- GOLGI
- VAC
- CYT
- NUC
- PLAS
- GOLGI
- VAC
- PERO
- MITO
- CHLO
- PLAST

---

**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 \( A \lor B \)

- **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

---


Edward Marcotte
Univ Texas Austin

Y Park
SUNY Buffalo
Reduced performance for new proteins

**PIPE2**

C1 (AB in training)

C3 (AB NOT in training)

**SIGPROD**

C1 (AB in training)

C3 (AB NOT in training)


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

\(\rightarrow\)

not relevant?
Similarity between negatives matters!

"Pseudo-improvement" through overlap between negatives

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

- Protein
- Sequence similar
- Training positive (interaction)
- Training negative (non-interacting)


Tobias Hamp
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)
Seemingly inverts previous results

Using all good data bad for C3

results for C3 (AB not in training)


Cross-validation challenge squared for PPIs
PPI pair prediction: Challenge 3 – how to do it best?
Two approaches for method
Approach 1: Sliding window/sections
answers: where A and B interact and how

input: feature of particular region/residue
output: 1 if two stretches close in space
Approach 2: Entire object - answers only whether A and B interact or not

input: feature of entire protein
output: 1 if two proteins interact
Step 1: binary decision: which method?
Step 1: binary decision
Profile-kernel SVM

T Hamp & B Rost (2015) Bioinformatics 31:1945-50
### Profile kernel example

<table>
<thead>
<tr>
<th>Prot1</th>
<th>CATLGLLGLVAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
</tr>
<tr>
<td>R</td>
<td>3</td>
</tr>
<tr>
<td>T</td>
<td>1</td>
</tr>
<tr>
<td>O</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>4</td>
</tr>
<tr>
<td>T</td>
<td>3</td>
</tr>
<tr>
<td>O</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Prot2</th>
<th>CARRGLLWAVAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>R</td>
<td>4</td>
</tr>
<tr>
<td>T</td>
<td>2</td>
</tr>
<tr>
<td>O</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
</tr>
<tr>
<td>T</td>
<td>1</td>
</tr>
<tr>
<td>O</td>
<td>4</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Examples**

- k-mer length = 3, conservation threshold = 5

\[
\Phi(\text{Prot1}) = (0, \ldots, 1, 1, 1, 0, \ldots, 0, \ldots, 0)
\]

\[
\Phi(\text{Prot2}) = (0, \ldots, 0, 1, 1, 0, \ldots, 0, \ldots, 0)
\]

\[
K(\text{Prot1}, \text{Prot2}) = \Phi(\text{Prot1}) \cdot \Phi(\text{Prot2}) = 32
\]
How many k-mer motifs are relevant to predict one protein? (here localisation)
## Recap: Experimental NLS: positive charges

<table>
<thead>
<tr>
<th>NLS</th>
<th>Protein</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>RKRKK</td>
<td>YstDNAPolalpha</td>
<td>Hsieh et al., 1998</td>
</tr>
<tr>
<td>RKRRR</td>
<td>Amida</td>
<td>Irie et al., 2000</td>
</tr>
<tr>
<td>KKKKKRRRKEK</td>
<td>LEF-1</td>
<td>Prieve et al., 1998</td>
</tr>
<tr>
<td>KKKKRRSREK</td>
<td>TCF-1</td>
<td>Prieve et al., 1998</td>
</tr>
<tr>
<td>RQARRNRRRRWR</td>
<td>HIV-1 Rev</td>
<td>Truant et al., 1999</td>
</tr>
<tr>
<td>RMWKKK</td>
<td>PDX-1</td>
<td>Moede et al., 1999</td>
</tr>
<tr>
<td>PKKKRRKVK</td>
<td>SV40 LrgT</td>
<td>Kalderon et al., 1984</td>
</tr>
<tr>
<td>PRRR</td>
<td>SRY</td>
<td>Sudbeck and Scherer, 1997</td>
</tr>
<tr>
<td>KKRRSA</td>
<td>H2B</td>
<td>Moreland et al., 1987</td>
</tr>
<tr>
<td>KA</td>
<td>v-Rel</td>
<td>Gilmore and Temin, 1988</td>
</tr>
<tr>
<td>RR</td>
<td>Amida</td>
<td>Irie et al., 2000</td>
</tr>
<tr>
<td>M</td>
<td>RanBP3</td>
<td>Welch et al., 1999</td>
</tr>
<tr>
<td>PKKR</td>
<td>Pho4p</td>
<td>Welch et al., 1999</td>
</tr>
<tr>
<td>PYLNKRKGKP</td>
<td>p53-NLS1</td>
<td>Liang and Clarke, 1999</td>
</tr>
<tr>
<td>KRx{7.9}FPKKKP</td>
<td>Hum-Ku70</td>
<td>Koike et al., 1999</td>
</tr>
<tr>
<td>KVTKRKHDEGSGSKRPK</td>
<td>GAL4</td>
<td>Chan et al., 1998</td>
</tr>
<tr>
<td>RLKKLCSDKx{19}KTKR</td>
<td>TCPTP</td>
<td>Chan et al., 1998</td>
</tr>
<tr>
<td>RKKIREDRX{18}RRKR</td>
<td>BDV-P</td>
<td>Schwemmle et al., 1999</td>
</tr>
<tr>
<td>RRERx{4}RPRKIPR</td>
<td>act/inh betaA</td>
<td>Blauer et al., 1999</td>
</tr>
<tr>
<td>KKKKKEEGEGKK</td>
<td>BDV-P</td>
<td>Shoya et al., 1998</td>
</tr>
<tr>
<td>PRPRKIPR</td>
<td>BDV-P</td>
<td>Shoya et al., 1998</td>
</tr>
<tr>
<td>PPRYIPQLPASAPT</td>
<td>TR2</td>
<td>Yu et al., 1998</td>
</tr>
<tr>
<td>KDCVINKHHHRNRCYCRLQR</td>
<td>PolyomaVP1</td>
<td>Chang et al., 1992</td>
</tr>
<tr>
<td>APKRKS6GSVKC</td>
<td>HIV-1 Tat</td>
<td>Truant et al., 1999</td>
</tr>
<tr>
<td>RKKRRQRRR</td>
<td>Rex</td>
<td>Palmeri and Malim, 1999</td>
</tr>
<tr>
<td>MPKTTRRPRRSQRKRPT</td>
<td>SRY</td>
<td>Sudbeck and Scherer, 1997</td>
</tr>
<tr>
<td>KRPMNAFIVWSRQRKK</td>
<td>SOX9</td>
<td>Sudbeck and Scherer, 1997</td>
</tr>
<tr>
<td>KRPMNAFMVWAQAARRK</td>
<td>NS5A</td>
<td>Ide et al., 1996</td>
</tr>
<tr>
<td>PPRKRTVV</td>
<td>DNAse EBV</td>
<td>Liu et al., 1998</td>
</tr>
<tr>
<td>YKRPCKRSFIRFI</td>
<td>DNAse EBV</td>
<td>Lyons et al., 1987</td>
</tr>
<tr>
<td>LKDVRRRRGLGPH</td>
<td>AdenovE1a</td>
<td>Bouvier and Baldacci, 1995</td>
</tr>
<tr>
<td>KRRPRP</td>
<td>hVDR</td>
<td>Vihinen-Ranta et al., 1997</td>
</tr>
<tr>
<td>RRSMKRK</td>
<td>CPV capsid</td>
<td>Kaneko et al., 1997</td>
</tr>
<tr>
<td>PAKRARRGYK</td>
<td>hGlu.cort.</td>
<td>Kaneko et al., 1997</td>
</tr>
<tr>
<td>RKCLQAGMNLEARKTKK</td>
<td>CFOS</td>
<td>Kaneko et al., 1997</td>
</tr>
<tr>
<td>RRERNKMAAACRNRKK</td>
<td>CJUN</td>
<td>Kaneko et al., 1997</td>
</tr>
<tr>
<td>KRMKNIAASKCRKRKL</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How many k-mer motifs are relevant to predict one protein? (here localization)
Many many many ‘motifs’ contribute!

![Box-and-whisker plot](image)

<table>
<thead>
<tr>
<th>Domain</th>
<th>k-mer length</th>
<th>localization</th>
<th>Avg # k-mers at X% of total decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eukaryota</td>
<td>6</td>
<td>cytoplasm</td>
<td>257,483 148,616 41,804</td>
</tr>
<tr>
<td>Eukaryota</td>
<td>6</td>
<td>nucleus</td>
<td>302,701 166,312 42,208</td>
</tr>
</tbody>
</table>
Many many many ‘motifs’ contribute!

Figure 3.1: Number of k-mers involved in the localization prediction of each dataset visualized as box-and-whisker plots. The number after each localization name is the k-mer length of this dataset. The box reaches from the $Q_1$ quartile of the data to the $Q_3$ quartile, the median is marked red. The $Q_1$ quartile is the middle value between the smallest data point and the median, $Q_2$ is the median and $Q_3$ is the middle value between the largest data point and the median. The lower whisker ranges to the smallest point within $Q_1$, $Q_3$ is the upper whisker.

3.1.2 Supporting to Non-Supporting k-mer Ratio

After determining how many k-mers contribute to a localization prediction the next step was to determine how much the individual k-mers contributed and how large the variation between the contributions was. As mentioned before in the Introduction, on each SVM level in the decision tree the k-mers are weighted according to their support for or against the location(s) that are classified by this level (see Figure 1.6). Therefore the number of k-mers actually supporting the final location is of interest. A simple assumption about the ratio between the number of supporting and the total number of k-mers would be that there are more k-mers supporting the final localization than there are k-mers which challenge this decision. As presented in Figures 3.6a to 3.2c this seems to hold true for most localizations across all three domains of life. In fact in almost all observed localizations there seem to be slightly more k-mers supporting a localization than there are non-supporting k-mers. The only exceptions to this are the plasma membrane in archaea and the outer membrane in bacteria, both of these classes contain very few proteins, therefore this might be an artifact created by the small sample size.
PPI from sequence through SVM profile kernel

- PPI-PK
- PPI-PK+expression
- SIGPROD
- PIPE2

C1 proteins have known PPIs

C3 not PPI known

Tobias Hamp

Note: profile kernel extracts thousands of patterns for two regions/domains/proteins
IV.11 protein interactions

PPI - pair predictions from other sources
PPI-prediction

idea 1: duplication
Duplication event

Organism P

Organism C
Duplication event

Organism P

change:
one new process:

C1—>C2
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>OrgP</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>
| OrgC  | x | x | x | x  | x  | x | x | x | x | x | x
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>OrgP</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>
| OrgC | x | x | x | x  | x  | x | x | x | x | x | x
**In silico predictions of P=P interactions**

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>OrgP</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></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>OrgC</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: domain copy
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
Correlated mutations/Coevolution

contact in 3D

correlated mutation in 1D

A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T K L C L T A K K E G P K D
A T K L T L T A K K E G P K D
A T K L T L T G A K K E G G C D
A T W L T L T A K K V G P C D
A T W L T L T A K K V G P C D


© Debbie Marks - HMS
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


© Thomas Hopf - TUM & Debbie Marks - HMS
Correlated mutations/Coevolution

Protein A

Protein B

correlated mutation in 1D

contact in 3D


© Debbie Marks - HMS
**In silico predictions of P=P interactions**

(C) CORRELATED MUTATIONS:
- F Pazos and A Valencia *2002 Proteins* 47, 219-227
Evolutionary couplings predict PPIs

Evolutionary couplings predict PPIs

Blinded prediction of inter-protein contacts in complexes with known 3D structure

CyoA – CyoB
EnvZ – OmpR (homolog)
MoaD – MoaE
FimC – FimD
BluC – BluF
BluC – BluD
DhaK – DhaL
CarB – CarA
GcsT – GcsH
RS3 – RS14

PPI-prediction idea 4: mirror tree
**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

Juan et al. (2008). PNAS.
© Ta-Tsen Soong, Columbia Univ
Mirror tree vs. phylogenetic profiles

Mirror tree performs worse than phylogenetic profiles

F Pazos & A Valencia (2001) Protein Engineering

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

Mirror tree more sophisticated

Phylogenetic profiles
PPI-prediction
idea 5: motifs
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:
arrays
(see above)
PPI-prediction
idea 7:
structure
Different interfaces = different physics?

HIV gp120 / CD4 / FAB

Using structure to predict PPI

Fig. 1
Fig. 3

PKD1 - PKCe
Template: E2D3 (blue) - ubiquitin (red)

EF1δ - VHL

Fig. 3
Structure-based prediction (PrePPI) good?

Fig. 2
IV.12 protein interactions

Back to sequences ...
Profile-kernel SVM
PPI partners
PPI-residue pair prediction: cross-validation

PPI-residue pair prediction: new proteins


IV.13 protein interactions

PPI - pair predictions

data perspective: integration
Integrating diverse data types

Gene fusion

Homology

Microarray

Functional similarity

Mirror tree

SVM-based protocol

Phylogenetic profiles

Conserved coexpression

Sequence domain

Subcellular localization

Text mining

Integration (naïve Bayes)

Ta-Tsen Soong & B Rost, unpublished

© Ta-Tsen Soong, Columbia Univ
Integrative PPI prediction

R Nair & B Rost (2005) LocTree. JMB
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
## 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)
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.14 protein interactions

PPI – PiNat
PiNat (Protein Interaction Network analysis tool)
<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><strong>Extra-cellular</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cytoplasm</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Organelles</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mitochondria</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Nuclear</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TM transmembrane</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
PiNat (Protein Interaction Network analysis tool)

Y Ofiran, G Yachdav, E Mozes, T Soong, R Nair, B Rost al.
2006 Bioinformatics 15:22 e402-7
PiNat view of Alzheimers

Q9P2H0

ADD

Y Ofran 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

Q9P2H0

ADD

© Burkhard Rost

ROSTLAB.

178/248
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**

extracellular → hsl(239,100%,40%)
cytoplasm → 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
IV.15 protein interactions
Protein–nucleotide
PPI interfaces use local segments

IV.8 protein interactions

Protein-DNA interactions
Protein-DNA: Background
DNA helix

major groove

minor groove

dye Hoechst33258

© Wikipedia

© Burkhard Rost
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
Protein-DNA binding (gal4): units

transcriptional activator GAL4 structure + DNA

Protein-DNA binding (gal4): surface

transcriptional activator GAL4
structure + DNA

Protein-DNA binding (gal4): domains

transcriptional activator GAL4
structure + DNA


green: 2 x leuzine zipper, red: 2 x zinc-finger binding
Protein-DNA binding (gal4): domains

transcriptional activator GAL4 structure + DNA

Protein-DNA binding (gal4): flexibility

transcriptional activator GAL4 structure + DNA


B-values ~ width
Census: protein-DNA binding
How many human proteins bind DNA?
How many human proteins bind DNA?

- not fully clear
- one extreme: >1,391 transcription factors
  e.g. JM Vaquerizas et al & NM Luscombe
  Nature Reviews Genetics 10, 252-63
From 3D to data for protein-DNA prediction methods
Protein-DNA interactions in 3D

Tomas Norambuena & Francisco Melo
(2010)
Protein-DNA interactions: 3 modes

A. interaction orthogonal
B. parallel to double-helix
C. both

Protein-DNA interactions: web interface

Prediction of protein-DNA interaction from 3D
Prediction of DNA-binding from 3D

- Use motifs (hth), e.g.

- Use charges/dipole moments, e.g.

- Use 3D + conservation
PatchFinder: concept

- **input:**
  - 3D structure or homology-based model
  - alignment

- **extraction:**
  - surface patches
  - evolutionary conservation
  - 3D-features: electrostatic potential, sec str cont, patch size, dipole moment, spatial asymmetry of residues, hydrogen donors/acceptors on surface

- **ML tool:** Random forest

G Nimrod et al & N Ben-Tal (2009) JMB, 387:1040-53 (Fig. 1)
PatchFinder: accuracy/coverage

G Nimrod et al & N Ben-Tal (2009) JMB, 387:1040-53 (Fig. 1 - Fig. 4)
PatchFinder: 3D change upon binding

HhaI DNA methyltransferase

apo: green
bound: blue
DNA: orange

PatchFinder:
12 binding residues for apo
15 binding residues in bound

G Nimrod et al & N Ben-Tal (2009) JMB, 387:1040-53 (Fig. 3)
PatchFinder: most valuable information

G Nimrod et al & N Ben-Tal (2009) JMB, 387:1040-53 (Fig. 5)
PatchFinder: “blind” prediction

Prediction

3D alignment of 2 proteins:

green: Q81BA8_BACCR: predicted to bind DNA
blue: Q82ZI8_ENTFA (belongs to Pfam DNA alkylation repair enzymes)
note: 17% PIDE

G Nimrod et al & N Ben-Tal (2009) JMB, 387:1040-53 (Fig. 7)
Prediction of protein-DNA interaction from sequence
NLS motifs
Shuttle into the nucleus

NLS

M9

Importin

Transportin

Cytoplasm

Nucleus
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?)
DNAsites: 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

Michael Menden
DNAsites2 improves over DNAsites

DNAsites2:

Michael Menden Master Thesis

significantly better per-residue than existing methods
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
SomeNA: no free lunch

SomeNA:

Peter Hoenigschmidt 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

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 eta l (2012) BMC Bioinformatics 13:8
SomeNA: predicts RNA binding

SomeNA:

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

P Hoenigschmid master thesis
<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>01: 2017/10/17</td>
<td>No lecture (makeup examen; PP last year)</td>
</tr>
<tr>
<td>02: 2017/10/19</td>
<td>No lecture (sick)</td>
</tr>
<tr>
<td>03: 2017/10/04</td>
<td>No lecture (sick)</td>
</tr>
<tr>
<td>04: 2017/10/26</td>
<td>Welcome: who we are</td>
</tr>
<tr>
<td>05: 2017/10/31</td>
<td>No lecture (SVV)</td>
</tr>
<tr>
<td>06: 2017/11/02</td>
<td>Intro function 1: concept of protein function</td>
</tr>
<tr>
<td>07: 2017/11/07</td>
<td>No lecture (Department event)</td>
</tr>
<tr>
<td>08: 2017/11/09</td>
<td>No lecture (Department event)</td>
</tr>
<tr>
<td>09: 2017/11/14</td>
<td>Localization 1</td>
</tr>
<tr>
<td>10: 2017/11/16</td>
<td>No lecture (Department event)</td>
</tr>
<tr>
<td>11: 2017/11/21</td>
<td>Localization 2</td>
</tr>
<tr>
<td>12: 2017/11/23</td>
<td>No lecture (Thanksgiving)</td>
</tr>
<tr>
<td>14: 2017/11/30</td>
<td>Localization 4</td>
</tr>
<tr>
<td>15: 2017/12/05</td>
<td>PPI 1 - interaction sites (chalk)</td>
</tr>
<tr>
<td>16: 2017/12/07</td>
<td>No lecture (Dies Academicus)</td>
</tr>
<tr>
<td>17: 2017/12/12</td>
<td>PPI 2 - sites</td>
</tr>
<tr>
<td>18: 2017/12/14</td>
<td>PPI 3 - sites</td>
</tr>
<tr>
<td>19: 2017/12/19</td>
<td>No lecture</td>
</tr>
<tr>
<td>20: 2017/12/21</td>
<td>No lecture</td>
</tr>
<tr>
<td>21-24:</td>
<td>no lectures - winter break (2017/12/24 - 2018/01/06)</td>
</tr>
<tr>
<td>25: 2018/01/09</td>
<td>PPI 4 - PPI pairing</td>
</tr>
<tr>
<td>28: 2018/01/11</td>
<td>PPI 5 - PPI pairing</td>
</tr>
<tr>
<td>29: 2018/01/16</td>
<td>SNP effect 1 (chalk)</td>
</tr>
<tr>
<td>30: 2018/01/18</td>
<td>SNP effect 2</td>
</tr>
<tr>
<td>31: 2018/01/23</td>
<td>WRAP up 1</td>
</tr>
<tr>
<td>32: 2018/01/25</td>
<td>WRAP up 2</td>
</tr>
<tr>
<td>33: 2018/01/30</td>
<td>WRAP up 3</td>
</tr>
<tr>
<td>34: 2018/02/01</td>
<td>examen (11:30-13:00)</td>
</tr>
<tr>
<td>35: 2018/02/06</td>
<td>No lecture</td>
</tr>
<tr>
<td>36: 2018/02/08</td>
<td>No lecture</td>
</tr>
</tbody>
</table>