Computational Biology 2 - Protein function:

Predicting protein interaction pairs

cb2_ppi3

Protein Prediction 2 - Protein function
Computational Biology 2 - TUM Winter 2014/15
Videos:  YouTube / www.rostlab.org

THANKS:

Tim Karl + Jonas Reeb

Special lectures:

• 10/28 & 30 - Tobias Hamp
• 11/20 - Tatyana Goldberg
• 12/02 Mikael Boden, Queensland U
• 12/16 & 18 - Andrea Schafferhans

No lecture:

LAST lecture:  January 20
Examen:  January 22

• Makeup:  Apr 14, 2015 - morning/noon
IV. (c) Predict protein interactions
IV.9 protein interactions

PPI - pair predictions

Substantial challenges for machine learning
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
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
Interface types differ in composition

statistical
significance
NOT
scientific significance
Correlation is not causation

Number of babies born per year

Data from Lower Saxony, Germany

Correlation is not causation


Data from Lower Saxony, Germany
Interfaces differ!
What does it mean?

antibody-1  antibody-2  gp120  CD4
Interfaces differ!
What does it mean?
“Acid test”: can we predict?
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:
PD Kwong et al. & WA Hendrickson
Prediction is the acid test for understanding — what about machine learning?
Machine learning = black magic


© Wikipedia

© Wikipedia
Let neural networks figure it out ...

Train

Test
Let neural networks figure it out ...
WEKA-like cross-validation

Train

Test
Enough for simple cross-validation?

© Wikipedia
WEKA-like cross-validation
do NOT choose patterns at random, instead:

EACH part of test exactly once!
Now enough?
## Results from cross-validation

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameters</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>{features1}, random forest</td>
<td>34%</td>
</tr>
<tr>
<td>Method 2</td>
<td>{features1}, SVM</td>
<td>35%</td>
</tr>
<tr>
<td>Method 3</td>
<td>{features2}, SVM</td>
<td>37%</td>
</tr>
</tbody>
</table>
## Results from cross-validation

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

-> Method 3 is best and performs at 37%
## 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>

**CONCLUSION correct?**

-> Method 3 is best and performs at 37%
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>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>32±1%</td>
</tr>
<tr>
<td>Method 1 (features1, random forest)</td>
<td>34±1%</td>
</tr>
<tr>
<td>Method 2 (features1, SVM)</td>
<td>35±1%</td>
</tr>
<tr>
<td>Method 3 (features2, SVM)</td>
<td>37±1%</td>
</tr>
</tbody>
</table>

**CONCLUSION correct?**

-> Method 3 is best and performs at 37%
3-way cross-validation

Train

Cross-Train

Test
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 *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
Hot spots reliably predicted from sequence!

hottest of hot = no error!

worst: ~60% right


© Burkhard Rost
All done right. We can predict. Do we now understand?
Machine learning = black magic

Surface “masked” by sugar

Surface composition

Average surface composition

Taking model apart
Machine learning are black boxes NOT because we cannot write the rules but because we do NOT understand them once written!
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
- predict residues saturated internally
- 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) submitted k-mer profile kernel SVM
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

Y Park & EM Marcotte (2012)
Nature Meth 9: 1134-1136

© Burkhard Rost
ROSTLAB.
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)

**References**

Analysis: T Hamp & B Rost, submitted
Sequence similarity only for PPPIs, i.e. positives enough?
Literature:
nobody ever considered sequence-similarity between negatives
Literature:
nobody ever considered sequence-similarity between negatives

->

not relevant?
Similarity between negatives matters!

“A”

“Pseudo-improvement” through overlap between negatives


T Hamp & B Rost, submitted

Tobias Hamp
we HAVE to also consider negative PPIs
Redundancy reduction for PPI networks
Non-redundant set

Redundant

Non-redundant

Protein

Positive training PPI

T Hamp & B Rost, submitted
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
May be using all data helps, may be not

more good data helps
more good data hurts


Tobias Hamp

T Hamp & B Rost, submitted
C1 = AB in training

More good data helps
More good data hurts

Tobias Hamp


T Hamp & B Rost, submitted
May be using all data helps, may be not

more good data helps

more good data hurts

Cross-validation challenge squared for PPIs
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

Tobias Hamp

T Hamp & B Rost, to be submitted

V. Predict variant effects toward Personalized health
Personalized health: harnessing the power of diversity
Learning from the mutability landscape

Many variants between us have effect

Predicted impact of mutation (SNAP score)

Cumulative fraction

Y Mahlich, M Hecht et al in submission
Lecture plan (CB2 function)

- 01: 2014/10/07: no lecture
- 02: 2014/10/09: welcome: who we are
- 03: 2014/10/14: no lecture (prof sick)
- 04: 2014/10/16: no lecture (prof sick)
- 05: 2014/10/21: Personalized medicine - predict effects of SNPs
- 06: 2014/10/23: Intro - function 1: concepts / homology
- 08: 2014/10/30: Tobias Hamp: Homology-based prediction of function 2
- 09: 2014/11/04: no lecture: SVV (student reps)
- 10: 2014/11/06: Intro - function 2: homology
- 14: 2014/11/20: Localization 2 - Tanya Goldberg
- 16: 2014/11/27: Protein-protein interaction 1
- 17: 2014/12/02: Protein-protein interaction 2
- 18: 2014/12/04: no lecture: Dies Academicus
- 19: 2014/12/09: Mikael Boden - Simple motifs
- 20: 2014/12/11: Mikael Boden - Complex motifs
- 21: 2014/12/16: Andrea Schafferhans: 3D function prediction
- 23-26: no lectures - winter break (2014/12/24 - 2015/01/06)
- 27: 2015/01/08: Punta - Pfam
- 28: 2015/01/13: Protein-protein interactions 3 - tripping in machine learning
- 29: 2015/01/15: WRAP up 1
- 30: 2015/01/20: WRAP up 2 / SNP effect
- 31: 2015/01/22: examen

2015