title: Prediction of localization
De novo prediction of localization

short title: pp2_localization2

lecture: Protein Prediction 2 - Protein function
TUM Winter 2012/2013
Announcements

Videos: SciVe / www.rostlab.org

THANKS:

Tim Karl + Jonas Reeb

Special lectures:
• Nov 15: Tatyana Goldberg
• Dec 16: Tobias Hamp
• Dec 18: Andrea Schafferhans
• Jan 29: Marco Punta (Pfam)
• Jan 31: Marco De Vivo (ISS Geneva)

No lecture:
• Dec 6 (Dies Academicus)
• Dec 20-Jan 6 (winter break+)

LAST lecture: Feb 5

Examen: Feb 7, 11:00 (likely this room)
• Makeup: may be: Apr 18 - morning

CONTACT: Marlena Drabik assistant@rostlab.org

Monday November 19, 2012
Today: Predict localization 1

☐ LAST WEEKs
  • Protein localization
  • Protein function (Tobias Hamp)

☐ THIS WEEK
  • Prediction of sub-cellular localization, de novo

☐ NEXT WEEK
  • Predicting the effect of mutations
II. De novo 1: Sub-cellular localization
II.6 Predict localization: de novo prediction: overall signal
Localization: simplistic perspective

- EXTRACELLULAR
- NUCLEAR
- CYTOPLASMIC
Localization correlates AA with composition

Do you believe that claim?
Localization correlates 2 AA composition

Residue composition projected onto first two principle components

Localization correlates 2 surface composition!

Localization correlates to surface composition!

Residue composition projected onto first two principle components

Surface “masked” by sugar

Surface composition

Average surface composition

nuclear

cytoplasmic

extracellular


Monday November 19, 2012
Insert: prediction of accessibility
Notation: protein structure 1D, 2D, 3D
Prediction of solvent accessibility
PHDacc: predict solvent accessibility

**PHDacc**

<table>
<thead>
<tr>
<th>local alignment</th>
<th>A A A</th>
<th>A A</th>
<th>L L L</th>
<th>L L L</th>
<th>A A G</th>
<th>C C S</th>
<th>G U V</th>
<th>A A A</th>
</tr>
</thead>
<tbody>
<tr>
<td>adjacent residues</td>
<td>% A A</td>
<td>Length</td>
<td>ΔN-term</td>
<td>ΔC-term</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**input local in sequence**

```
ACLI GSV ins del cons
```

| A C L I G S V | 100 0 0 0 0 0 0 0 0 0 | 1.17 |
| A C L I G S V | 100 0 0 0 0 0 0 33 0 0.42 |
| A C L I G S V | 0 0 100 0 0 0 0 0 33 0.92 |
| A C L I G S V | 0 0 33 66 0 0 0 0 0 0.74 |
| A C L I G S V | 66 0 0 0 33 0 0 0 0 1.17 |
| A C L I G S V | 0 66 0 0 0 33 0 0 0 0.74 |
| A C L I G S V | 0 0 0 33 0 0 66 0 0 0.48 |

**input global in sequence**

- percentage of each amino acid in protein length of protein (\( \leq 60, \leq 120, \leq 240, > 240 \))
- distance: centre, N-term (\( \leq 40, \leq 30, \leq 20, \leq 10 \))
- distance: centre, C-term (\( \leq 40, \leq 30, \leq 20, \leq 10 \))

**output layer**

- first level only

Monday November 19, 2012
Evolution for accessibility prediction

- Detailed prediction problematic
- Significant gain by evolutionary information:
  
in/out with > 75% accuracy!
Number of exposed residues observed (DSSP)

Number of exposed residues predicted (PHDacc)
PROFsec: Evolutionary information + more

B Rost (2001) J Struct Biol 134, 204-18
PROFacc: better for surface

- Detailed prediction still problematic
- More significant gain by evolutionary information:

  in/out with > 78% accuracy!

  surface residues predicted > 80% right
INSERT: using accessibility to predict function
Functionally important residues

ConSurf


© Marco Punta & Yanay Ofran & Burkhard Rost (Columbia New York)
II.7 Predict localization: de novo prediction Rostlab
Localization correlates with surface composition.

Rajesh Nair
now: FDA, Washington
Features that discriminate localization

- amino acid
- surface
- sec str
- combined

R Nair & B Rost unpublished

© Rajesh Nair

Monday November 19, 2012
INSERT: concept of neural networks
Simple neural network

\[
\text{out0} = J_{11} \text{in}_1 + J_{12} \text{in}_2
\]

\[
\text{out} = \tanh (\text{out0})
\]
Training a neural network 1

Diagram showing a neural network with inputs and outputs.
Training a neural network 2

Errare = (out net - out want)²
Training a neural network 3

Error

Junctions

Monday November 19, 2012
Training a neural network 3

Round 1
1 1
0 1
0 1
1 1

Round 2
1 0
0 1
0 1
1 2

Error
Junctions

in
-2 -1 1 2
out
-1 1
Training a neural network 3

Round 1

1 1
0 1
0 1
1 1

Round 2

1 0
0 1
0 1
1 1

Error

Junctions

in

out

-2 -1 1 2
Training a neural network 3

Round 1
1
0
0
1

Round 2
1
0
0
1

Round 3
1
0
0
1

Error
Junctions

in
-2
-1
1
2
out
-1
1

Monday November 19, 2012
Neural networks classify points
Neural networks classify points
Neural networks classify points
Neural networks classify points
Neural networks classify points
Simple neural network with hidden layer

$$\text{out}_i = f \left( \sum_j J^2_{ij} \cdot f \left( \sum_k J^1_{jk} \cdot \text{in}_k \right) \right)$$
Principles of neural networks: error

• **output:**

\[
\text{out}_i = \sum_{j=1}^{N_{\text{in}+1}} J_{ij} \text{in}_j
\]

\text{in}_j \text{ value of input unit } j \; ; \; \text{out}_i \text{ value of output unit } i \; ;

\text{J}_{ij} \text{ connection between input unit } j \text{ and output unit } i

• **error:**

\[
E = \sum_{i=1}^{N_{\text{out}}} (\text{out}_i - \text{des}_i)^2
\]

\text{out}_i \text{ value of output unit } i \; ; \; \text{des}_i \text{ secondary structure state observed for central amino acid for output unit } i \text{ (e.g. for a helix: des}_1=1, \text{des}_2=0, \text{des}_3=0)\]

• **free variables:** connections \{ J \}

• **goal:**
  ○ representation of set of examples (training set) for which the mapping input->output is known, i.e., the secondary structure state of the central residue has been observed by the network
Principles of neural networks: training

- training = change of connections \{J\} such that \( E \) decreases

simplest procedure:
- gradient descent

\[
\Delta J_{ij}(t+1) = - \varepsilon \frac{\partial E(t)}{\partial J_{ij}(t)} + \alpha \Delta J_{ij}(t-1)
\]

where \( \frac{\partial E}{\partial J} \) is the derivative of the error with respect to the network connection; \( t \) is the algorithmic time given by the presentation of one example; \( \varepsilon \) determines the step width of the change (learning strength, typically some 0.01); \( \alpha \) gives the contribution of the momentum term (\( \Delta J(t-1) \), typically some 0.2), which permits uphill moves.
Effect of over-training: theory

- Training time
- over-train
Effect of over-training: practice

![Graph showing effect of over-training on correct classifications per example over cycles, with separate lines for ratio of training and testing set ratios.](image)
NN predicts localization

1st layer: feature-to-localization

2nd layer: localization-to-localization

3rd layer: jury

Prediction = max (P_{nuc}, P_{ext}, P_{cyt}, P_{mit}, P_{other})

R Nair and B Rost (2003) *Proteins* 53: 917-30
INSERT: concept of SVM
SVM: non-linear map to feature space

- Map data to a different space (feature space)
- Possibly higher dimension
- Linearly separable in the new space.

Original Problem

Transformed Problem

\[ \Phi(\bar{x}) \]


© Rajesh Nair
SVM: non-linear map to feature space

- Find maximal margin hyperplane in feature space
- Maximal margin hyperplane defined by the support vectors

SVM: non-linear map to feature space

- Find maximal margin hyperplane in feature space
- Maximal margin hyperplane defined by the support vectors

Prediction scheme

Protein

NUC

CYT

EXT

MIT

SUM

LOCALIZE

Discrimination algorithm like ‘neural network’ or ‘support vector machine’

© Rajesh Nair
Problems with simple approach

- Parallel architecture
  - Protein trafficking is hierarchical


© Rajesh Nair
Problems with simple approach

- **Parallel architecture**
  - Protein trafficking is hierarchical

- **Only amino acid sequence**
  - Secondary structure/surface
  - Evolution
  - Predictions from high accuracy methods


© Rajesh Nair
Solution: hierarchical SVM

Protein

Secretory Pathway

SVM

Intra-cellular

Cytoplasm

SVM

EXTRA-CELLULAR

ORGANELLE

NUCLEUS

CYTOSOL

MITOCHONDRIA

© Rajesh Nair
Solution: hierarchical SVM

Amino acid composition
+ Residue in secondary structure state
+ N-terminal composition

Protein

Sees only intra-cellular proteins

Secretory Pathway

EXTRA-CELLULAR
ORGANELLE
NUCLEUS
CYTOSOL
MITOCHONDRIA

Cytoplasm

Harder to separate classes as you go down tree
Solution: hierarchical SVM

81/80

Secretory Pathway

SVM

89 ± 2 %

Intra-cellular

Cytoplasm

SVM

EXTRA-CELLULAR

ORGANELLE

NUCLEUS

Cytoplasm

Mitochondria

Observed

Predicted

OK

under

over

Accuracy =

Coverage =

R Nair & B Rost 2005 JMB 348:85-100

© Rajesh Nair

Monday November 19, 2012
Solution: hierarchical SVM

Protein

Secretory Pathway

SVM

Intra-cellular

Cytoplasm

EXTRA-CELLULAR

ORGANELLE

NUCLEUS

78 ± 4 %

Cytoplasm

MITOCHONDRIA

© Rajesh Nair
Monday November 19, 2012
Solution: hierarchical SVM

Protein

SVM 78/78

Secretory Pathway

SVM 83/81

EXTRA-CELLULAR

SVM

ORGANELLE 51/52

NUCLEUS

CYTOSOL

MITOCHONDRIA

63/66 70/67

Intra-cellular

74 6%
## Comparison to other methods

<table>
<thead>
<tr>
<th></th>
<th>LOCtree</th>
<th>TargetP</th>
<th>SubLoc</th>
<th>Psort</th>
<th>NNPSL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Secretory Pathway</strong></td>
<td>Accuracy</td>
<td>87</td>
<td>93</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Coverage</td>
<td>90</td>
<td>73</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Extra-cellular</strong></td>
<td>Accuracy</td>
<td>86</td>
<td>-</td>
<td>73</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>Coverage</td>
<td>93</td>
<td>-</td>
<td>53</td>
<td>32</td>
</tr>
<tr>
<td><strong>Nucleus</strong></td>
<td>Accuracy</td>
<td>77</td>
<td>-</td>
<td>64</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>Coverage</td>
<td>85</td>
<td>-</td>
<td>71</td>
<td>75</td>
</tr>
<tr>
<td><strong>Cytosol</strong></td>
<td>Accuracy</td>
<td>82</td>
<td>-</td>
<td>43</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>Coverage</td>
<td>64</td>
<td>-</td>
<td>56</td>
<td>47</td>
</tr>
<tr>
<td><strong>Mitochondria</strong></td>
<td>Accuracy</td>
<td>73</td>
<td>54</td>
<td>48</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>Coverage</td>
<td>78</td>
<td>75</td>
<td>59</td>
<td>59</td>
</tr>
<tr>
<td><strong>Overall Accuracy</strong></td>
<td></td>
<td>78</td>
<td>-</td>
<td>57</td>
<td>51</td>
</tr>
</tbody>
</table>
Pro/Cons of hierarchical SVM

- Protein
  - SVM
    - Secretry Pathway
    - Intra-cellular
      - SVM
        - EXT
        - ORG
        - NUC
      - SVM
        - Cytoplasm
          - CYT
          - MIT

© Rajesh Nair
Pro/Cons of hierarchical SVM

PRO

better discrimination by focusing only on relevant classes
predicts intermediate classes – not just the final classes
balanced predictions between accuracy and coverage
easy integration of independently developed modules

CON

mistakes at top node cannot be corrected later
no probability distribution for protein
Hierarchical prediction system

B Alberts et al. 1994 The Cell Garland

R Nair & B Rost 2005 JMB 348:85-100

Monday November 19, 2012
LOCtree: proteome annotation

- **Yeast**
  - 5 K

- **Arabidopsis Th.**
  - 21 K

- **Human**
  - 31 K

Legend:
- Blue: Extra-cellular
- Cyan: Organelles
- Green: Nuclear
- Red: Cytosol
- Yellow: Mitochondria
- Beige: Chloroplast

---

*R Nair & B Rost 2005 JMB 348:85-100*

Monday November 19, 2012
II.8 Predict localization: de novo prediction some other methods
Publicly available methods

- **NNPSL, SubLoc**
  - Amino acid composition

- **PSORT**
  - Targeting motifs + amino acid composition

- **SignalP, TargetP**
  - N-terminal amino acid sequence

- Most published methods use sequence composition
PSORT
Minoru Kanehisa
Director & Prof: Inst Chem Res, Kyoto Univ
Prof: Human Genome Center, Univ Tokyo
© http://kanehisa.kuicr.kyoto-u.ac.jp/People/kanehisa.html

KEGG
Kyoto Encyclopedia of Genes and Genomes
Minoru Kanehisa

- KEGG - Kyoto
- “Bench 2 Bedside”
- Publications
  - >250 papers (Dec 2011)
  - 2 >1,000 citations (2011/12)
  - 29 > 100 citations (2011/12)
  - H-index: 49 (2011/12)

Minoru Kanehisa
Director & Prof: Inst Chem Res, Kyoto Univ
Prof: Human Genome Center, Univ Tokyo
© http://kanehisa.kuicr.kyoto-u.ac.jp/People/kanehisa.html

Monday November 19, 2012
Kenta Nakai

- predict localization
  - PSORT, WolfPSORT

- Publications
  - >220 papers (Dec 2011)
  - 2 >1,000 citations (2011/12)
  - 18 > 100 citations (2011/12)

Kenta Nakai
Prof: Human Genome Center, Univ Tokyo

© http://akira-pj.lserp.osaka-u.ac.jp/pf_kenta_nakai.html
PSORT - the beginning

Fig. 4. Basic strategy for reasoning of protein localization sites. This is shown schematically in order to clarify the overall organization of rules. The actual path of reasoning does not always follow this tree exactly.

K Nakai and M Kanehisa (1991) *Proteins* 11: 95-110: Fig. 4
Paul Horton

Prof: Comp. Biol. Res. Center (AIST), Univ Tokyo

© http://akira-pj.lserp.osaka-u.ac.jp/pf_kenta_nakai.html
PSORT - the beginning

PSORT


PSORT II

- P Horton and K Nakai (1997) Better prediction of protein cellular localization sites with the k nearest neighbors classifier In 5th ISMB AAAI Press, 147-52
### Table 1. Features detected by PSORT II

<table>
<thead>
<tr>
<th>Feature</th>
<th>Criteria</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-terminal signal peptide</td>
<td>Modified McGeoch's method and the cleavage-site consensus</td>
<td>10, 11</td>
</tr>
<tr>
<td>Mitochondrial-targeting signal</td>
<td>Amino acid composition of the N-terminal 20 residues and some weak cleavage-site consensus</td>
<td>5, 12</td>
</tr>
<tr>
<td>Nuclear-localization signals</td>
<td>Combined score for various empirical rules</td>
<td></td>
</tr>
<tr>
<td>ER-lumen-retention signal</td>
<td>The KDEL-like motif at the C-terminus</td>
<td></td>
</tr>
<tr>
<td>ER-membrane-retention signal</td>
<td>Motifs: XXRR-like (N-terminal) or KKXX-like (C-terminal)</td>
<td></td>
</tr>
<tr>
<td>Peroxisomal-targeting signal</td>
<td>PTS1 motif at the C-terminus and the PTS2 motif</td>
<td></td>
</tr>
<tr>
<td>Vacuolar-targeting signal</td>
<td>[TIK][LP][NKI] motif</td>
<td></td>
</tr>
<tr>
<td>Golgi-transport signal</td>
<td>The YQRL motif (preferentially at the cytoplasmic tail)</td>
<td></td>
</tr>
<tr>
<td>Tyrosine-containing motif</td>
<td>Number of tyrosine residues in the cytoplasmic tail</td>
<td></td>
</tr>
<tr>
<td>Dileucine motif</td>
<td>At the cytoplasmic tail</td>
<td></td>
</tr>
<tr>
<td>Membrane span(s)/topology</td>
<td>Maximum hydrophobicity and the number of predicted spans; charge difference across the most N-terminal transmembrane segment</td>
<td>5, 13, 14</td>
</tr>
<tr>
<td>RNA-binding motif</td>
<td>RNP-1 motif</td>
<td>15</td>
</tr>
<tr>
<td>Actinin-type actin-binding motifs</td>
<td>From PROSITE</td>
<td>15</td>
</tr>
<tr>
<td>Isoprenyl motif</td>
<td>CaaX motif at the C-terminus</td>
<td></td>
</tr>
<tr>
<td>GPI-anchor</td>
<td>Type-1a membrane protein with very short tail</td>
<td></td>
</tr>
<tr>
<td>N-myristoylation motif</td>
<td>At the N-terminus</td>
<td></td>
</tr>
<tr>
<td>DNA-binding motifs</td>
<td>63 motifs from PROSITE</td>
<td>15</td>
</tr>
<tr>
<td>Ribosomal-protein motifs</td>
<td>71 motifs from PROSITE</td>
<td>15</td>
</tr>
<tr>
<td>Prokaryotic DNA-binding motifs</td>
<td>33 motifs from PROSITE</td>
<td>15</td>
</tr>
<tr>
<td>Amino acid composition</td>
<td>Neural network score that discriminates between cytoplasmic and nuclear proteins</td>
<td>3</td>
</tr>
<tr>
<td>Coiled-coil structure</td>
<td>Number of residues in the predicted coiled-coil state</td>
<td>17</td>
</tr>
<tr>
<td>Length</td>
<td>Length of the sequence</td>
<td></td>
</tr>
</tbody>
</table>

K Nakai and P Horton (1999) *Trends Biochem Sci* 24: 34-6: Table 1
PSORT - the beginning

- **PSORT**

- **PSORT II**
  - P Horton and K Nakai (1997) Better prediction of protein cellular localization sites with the k nearest neighbors classifier In 5th ISMB AAAI Press, 147-52

- **Wolf-PSORT**
Wolf-PSORT:

Important extension and update of PSORT-II

Other methods
Publicly available methods

- **NNPSL, SubLoc**
  - Amino acid composition

- **PSORT**
  - Targeting motifs + amino acid composition

- **SignalP, TargetP**
  - N-terminal amino acid sequence

- Most published methods use sequence composition
Sherloc: text mining + prediction


Sherloc: idea

- like LocKey but full abstracts (ALL in SWISS-PROT)
- + de novo prediction methods (MultiLoc et al.)

Sherloc2:
S Briesemeister, T Blum, S Brady, Y Lam, O Kohlbacher and H Shatkay (2009) J Proteome Res 8: 5363-6, Fig. 1
MultiLoc/TargetLoc


A Hoglund, P Donnes, T Blum, HW Adolph and O Kohlbacher (2006) Bioinformatics 22: 1158-65: Fig. 2
MultiLoc 2 = MultiLoc + GO + profiles
coding input

- N-term: special composition
- partition into k-segments of length L (composition for each)
- motifs with gap
- sliding window composition

II.9 predict localization:
more detail: sub-nuclear prediction & models for sorting
Mikael Bodén

- predict localization alignments, motifs (MEME)

- Publications
  - >50 papers (Dec 2011)
  - 10 > 10 citations (2011/12)
  - H-index: 15 (2011/12)

Mikael Bodén
Inf Techn & Electrical Engng and Chemistry & Mol Biosciences Univ Queensland, Australia

© http://itee.uq.edu.au/~mikael

Monday November 19, 2012
Sorting into sub-nuclear compartments
Shuttle into the nucleus

Cytoplasm

Nucleus

Importin

NLS

Transportin

M9
Model for nuclear import

AM Mehdi, MSB Sehgal, B Kobe, TL Bailey, M Boden (2011)
Bioinformatics 27: 1239-46: Fig. 1
A probabilistic model of nuclear import of proteins

Table 3. Accuracy of predicting nuclear import for proteins with less than 30% sequence similarity

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (MCC)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mouse</td>
<td>Yeast</td>
<td></td>
</tr>
<tr>
<td>Combined model</td>
<td>0.50</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td>cNLS Mapper</td>
<td>0.28</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>NLStradamus</td>
<td>0.29</td>
<td>0.19</td>
<td></td>
</tr>
</tbody>
</table>
Bayesian network for sub-nuclear prediction
Sub-nuclear: Bayesian network

ADC Bauer, K Willadsen, FA Buske, KA Le Cao, TL Bailey, G Dellaire, M Boden (2011) Bioinformatics 27: i7-i14: Fig. 1
# Sorting the nuclear proteome

**Table 1.** Cross-validated prediction accuracy on proteins with known compartment associations

<table>
<thead>
<tr>
<th>Compartment</th>
<th>Proteins</th>
<th>AUC50 (SD)</th>
<th>AUC (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cajal body</td>
<td>51</td>
<td>0.22 (0.02)</td>
<td>0.60 (0.03)</td>
</tr>
<tr>
<td>Chromatin</td>
<td>323</td>
<td>0.17 (0.02)</td>
<td>0.71 (0.01)</td>
</tr>
<tr>
<td>Nuclear lamina</td>
<td>77</td>
<td>0.17 (0.04)</td>
<td>0.70 (0.01)</td>
</tr>
<tr>
<td>Nuclear pore</td>
<td>51</td>
<td>0.41 (0.07)</td>
<td>0.79 (0.05)</td>
</tr>
<tr>
<td>Nuclear speckle</td>
<td>404</td>
<td>0.24 (0.01)</td>
<td>0.71 (0.01)</td>
</tr>
<tr>
<td>Nucleolus</td>
<td>596</td>
<td>0.14 (0.01)</td>
<td>0.60 (0.01)</td>
</tr>
<tr>
<td>Perinucleolar</td>
<td>24</td>
<td>0.41 (0.09)</td>
<td>0.80 (0.05)</td>
</tr>
<tr>
<td>PML body</td>
<td>91</td>
<td>0.23 (0.06)</td>
<td>0.77 (0.03)</td>
</tr>
<tr>
<td>Mean (compartment)</td>
<td></td>
<td>0.25</td>
<td>0.71</td>
</tr>
</tbody>
</table>

ADC Bauer, K Willadsen, FA Buske, KA Le Cao, TL Bailey, G Dellaire, M Boden (2011) *Bioinformatics* 27: i7-i14: Table 2
**Sub-nuclear compartments: new annotations**

### Sorting the nuclear proteome

**Table 3.**
Predicted compartment associations for 2281 unannotated nuclear proteins

<table>
<thead>
<tr>
<th>Compartment</th>
<th>Additional proteins</th>
<th>Probability threshold</th>
<th>FDR at threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cajal body</td>
<td>23</td>
<td>0.24</td>
<td>0.76</td>
</tr>
<tr>
<td>Chromatin</td>
<td>509</td>
<td>0.43</td>
<td>0.52</td>
</tr>
<tr>
<td>Nuclear lamina</td>
<td>17</td>
<td>0.38</td>
<td>0.68</td>
</tr>
<tr>
<td>Nuclear pore</td>
<td>12</td>
<td>0.31</td>
<td>0.43</td>
</tr>
<tr>
<td>Nuclear speckle</td>
<td>229</td>
<td>0.41</td>
<td>0.45</td>
</tr>
<tr>
<td>Nucleolus</td>
<td>1266</td>
<td>0.44</td>
<td>0.47</td>
</tr>
<tr>
<td>Perinucleolar</td>
<td>1</td>
<td>0.29</td>
<td>0.58</td>
</tr>
<tr>
<td>PML body</td>
<td>96</td>
<td>0.34</td>
<td>0.64</td>
</tr>
</tbody>
</table>

ADC Bauer, K Willadsen, FA Buske, KA Le Cao, TL Bailey, G Dellaire, M Boden *(2011)* *Bioinformatics* 27: i7-i14: Table 3
### Sub-nuclear compartments: TF enrichment

#### Sorting the nuclear proteome

**Table 4.** Transcription factor enrichment in compartments, with significant over-representation within compartments marked.

<table>
<thead>
<tr>
<th>Compartment</th>
<th>RIKEN TFs</th>
<th></th>
<th></th>
<th>DBD TFs</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count</td>
<td>%</td>
<td></td>
<td>Count</td>
<td>%</td>
<td></td>
</tr>
<tr>
<td>Cajal body</td>
<td>10</td>
<td>19.6</td>
<td></td>
<td>4</td>
<td>7.8</td>
<td></td>
</tr>
<tr>
<td>Chromatin</td>
<td>100</td>
<td>31.0*</td>
<td></td>
<td>47</td>
<td>14.6*</td>
<td></td>
</tr>
<tr>
<td>Nuclear lamina</td>
<td>4</td>
<td>5.2</td>
<td></td>
<td>4</td>
<td>5.2</td>
<td></td>
</tr>
<tr>
<td>Nuclear pore</td>
<td>0</td>
<td>0.0</td>
<td></td>
<td>0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Nuclear speckle</td>
<td>54</td>
<td>13.4</td>
<td></td>
<td>19</td>
<td>4.7</td>
<td></td>
</tr>
<tr>
<td>Nucleolus</td>
<td>71</td>
<td>11.9</td>
<td></td>
<td>18</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>Perinucleolar</td>
<td>6</td>
<td>25.0</td>
<td></td>
<td>0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>PML body</td>
<td>27</td>
<td>29.7*</td>
<td></td>
<td>17</td>
<td>18.7*</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>213</td>
<td>16.9</td>
<td></td>
<td>95</td>
<td>7.4</td>
<td></td>
</tr>
</tbody>
</table>

*P<0.05.

ADC Bauer, K Willadsen, FA Buske, KA Le Cao, TL Bailey, G Dellaire, M Boden *(2011)* **Bioinformatics** 27: i7-i14: Table 4
Fig. 2

ADC Bauer, K Willadsen, FA Bus Bailey, G Dellaire, M Boden (2011) Bioinformatics 27: i7-i14: Fig. 2
II.10 use predicted localization:

PPI - PiNat
PiNat (Protein Interaction Network analysis tool)

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

<table>
<thead>
<tr>
<th></th>
<th>Extracellular</th>
<th>Cytoplasm</th>
<th>Organelles</th>
<th>Mitochondria</th>
<th>Nuclear</th>
<th>TM transmembrane</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extracellular</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cytoplasm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organelles</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mitochondria</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nuclear</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TM transmembrane</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**PiNat** (Protein Interaction Network analysis tool)

Y Ofran et al. & Rost *Bioinformatics* 22:e402-7

Monday November 19, 2012
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 November 19, 2012
II.11 Predict localization: prediction systems
Surface “masked” by sugar

Post translational modification

- N-terminal signal peptide cleavage
- Proteolytic cleavage, proteosome cleavage
- Phosphorylation
- Lipid modification
- $N$- and $O$-glycosylations
List of services:
http://us.expasy.org/tools/
Lecture plan (PP2 function)

01: 2012/10/16: no lecture
02: 2012/10/18: welcome: who we are
03: 2012/10/23: individualized medicine
04: 2012/10/25: Intro - function 1: concepts
05: 2012/10/30: Tatyana Goldberg: localization
06: 2012/11/01: no lecture: All Saints
07: 2012/11/06: Intro - function 2: homology
08: 2012/11/08: Intro - function 3: motifs
09: 2012/11/13: Localization 1
11: 2012/11/20: Localization 2
13: 2012/11/27: Localization 4
14: 2012/11/29: SNP effect 1
15: 2012/12/04: SNP effect 2
16: 2012/12/06: no lecture: Dies Academicus
17: 2012/12/11: SNP effect 3
18: 2012/12/13: Protein-protein interaction 1
19: 2012/12/18: Andrea Schafferhans: 3D function prediction
20: 2012/12/20: no lecture
21-24: no lectures - winter break (2012/12/23 - 2013/01/06)
25: 2013/01/08: Protein-protein interaction 2
26: 2013/01/10: Protein-protein interaction 3
27: 2013/01/15: Protein-DNA interaction 1
28: 2013/01/17: Protein-DNA/RNA interaction 2
29: 2013/01/22: Andrea Schafferhans: Docking
30: 2013/01/24: networks
31: 2013/01/29: Marco Punta (Pfam)
32: 2013/01/31: Marco De Vivo (ISS Geneva)
33: 2013/02/05: wrap-up
34: 2013/02/07: examen