title: Localization 4
short title: cb2_localization4
lecture: Protein Prediction 2 (for Computer Science) - Protein function
TUM winter semester
Announcements

Videos: YouTube / www.rostlab.org

THANKS:

Tobias Olenyi

Special lectures:

- 11/21 THU TBC: Konstantin Weissenow deep learning binding sites
- 12/12 THU TBC: Maria Littmann & Jia Jun Qiu: Protein-DNA-RNA-ligands binding

No lecture:

- 10/31 THU All Saints
- 11/12 TUE SVV (student rep)
- 11/28 THU Thanksgiving
- 12/05 THU TUM Dies Academicus
- 12/19-01/07 - no lecture Xmas+

LAST lecture: Jan 21 (followed by 2 wrap-up sessions)

Examen: Jan 30 11:00-13:00 Room LMU Physics HS019

Makeup: NONE (emergency: Apr 21 & Apr 23, 2020 lecture time)

CONTACT: TBC teaching@rostlab.org
Recap
Protein function
Cellular Compartmentalization

Prokaryotic Cell

Eukaryotic Cell

Common compartment

Common physiological function

K Rogers (2011) Britannica
Protein Function: Examples

- **Nucleosome:** DNA Maintenance
- **Ribosome:** Translation
- **Collagen:** Structural Support
- **Crystalline:** Capturing Light
- **ATP Synthase:** Molecular Motor
- **G Protein:** Signaling & Transport Across Cell Membranes
Sequence to Function Gap

- UniProt: 181M (10^6) protein sequences (2018_11)
- Swiss-Prot: 561K (10^3) curated entries (2019_11)

Need for reliable automatic predictions of protein function from amino acid sequence alone

MTSHSYKDRLGFDPNGEQ
PGSNNSMKRSSSRQTHH
HQSYYHATTSSQSPARISV
SPGGNNGTLEYQQVQREN
NW…

Predictor → Protein Function

Describing Protein Function

Gene Ontology (GO)
Describing Protein Function

The Gene Ontology

Gene Ontology (GO)

- Cellular Component
  - 3461 terms
- Molecular Function
  - 10543 terms
- Biological Process

Cellular Component = localization = WHERE

Stats: Nov 2013 26116 terms

Describing Protein Function

The Gene Ontology (GO)

- Cellular Component: 3461 terms
- Molecular Function: 10543 terms
- Biological Process: 26116 terms

Stats: Nov 2013

Molecular Function = e.g. activity = HOW

simulations: courtesy of Marco Punta (ICR London) & Marco de Vivo (ISS Geneva)

Describing Protein Function

Gene Ontology (GO)

Cellular Component
- 3461 terms

Molecular Function
- 10543 terms

Biological Process
- 26116 terms

Stats: Nov 2013

Illustration: © The Nobel Committee for Physiology or Medicine, 2019.
Illustrator: Mattias Karlén

Biological Process
= pathway
= WITH WHO/WHEN

Recap
Sub-cellular localization: application
Drug targets tend to be found in membranes, cytoplasm or are extra-cellular!
## Mis-localized -> disease

<table>
<thead>
<tr>
<th>Protein</th>
<th>Disease</th>
<th>Mechanism</th>
<th>Mislocalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRY</td>
<td>Swyer syndrome</td>
<td>Mutation of NLS</td>
<td>Loss of nuclear localization</td>
</tr>
<tr>
<td>SHOX</td>
<td>Léri–Weill dyschondrodystostosis</td>
<td>Mutation of NLS</td>
<td>Cytoplasmic retention</td>
</tr>
<tr>
<td>TRPS1</td>
<td>TRPS</td>
<td>Mutation of NLS</td>
<td>Loss of nuclear localization</td>
</tr>
<tr>
<td>ARX</td>
<td>XLAG</td>
<td>Mutation of NLS</td>
<td>Loss of nuclear localization</td>
</tr>
<tr>
<td>FOXP2</td>
<td>Speech–language disorder</td>
<td>Mutation of NLS</td>
<td>Loss of nuclear localization</td>
</tr>
<tr>
<td>AIRE</td>
<td>APECED</td>
<td>Mutation of ZFD</td>
<td>Cytoplasmic retention</td>
</tr>
<tr>
<td>RPS19</td>
<td>Diamond–Blackfan anemia</td>
<td>Mutation of NoS</td>
<td>Loss of nucleolar localization</td>
</tr>
<tr>
<td>AGT</td>
<td>Primary hyperoxaluria type 1</td>
<td>Polymorphism and/or mutation</td>
<td>Mitochondrial mislocalization</td>
</tr>
<tr>
<td>hsMOK2</td>
<td>Laminopathy</td>
<td>Mutation of lamin A/C</td>
<td>Formation of nuclear aggregates</td>
</tr>
<tr>
<td>SHOC2</td>
<td>Noonan-like syndrome</td>
<td>Acquired N-myristoylation</td>
<td>Mislocalization to the plasma membrane</td>
</tr>
<tr>
<td>Rhodopsin</td>
<td>Retinitis pigmentosa</td>
<td>Mutations</td>
<td>ER retention</td>
</tr>
<tr>
<td>AVPR2</td>
<td>Nephrogenic diabetes insipidus</td>
<td>Mutations</td>
<td>ER retention</td>
</tr>
<tr>
<td>ATP7B</td>
<td>Wilson disease</td>
<td>H1069Q mutation</td>
<td>ER retention</td>
</tr>
<tr>
<td>ABCA1</td>
<td>Tangier disease</td>
<td>Mutations</td>
<td>Loss of plasma membrane localization</td>
</tr>
<tr>
<td>Tau</td>
<td>Neurodegenerative diseases</td>
<td>Hyperphosphorylation</td>
<td>Mislocalization to dendritic spines</td>
</tr>
<tr>
<td>TARDBP</td>
<td>ALS and FTLD</td>
<td>Unknown</td>
<td>Cytoplasmic mislocalization</td>
</tr>
<tr>
<td>FUS</td>
<td>FTLD</td>
<td>Mutations</td>
<td>Cytoplasmic mislocalization</td>
</tr>
<tr>
<td>FOXO</td>
<td>Various types of cancer</td>
<td>Post-translational modifications</td>
<td>Cytoplasmic mislocalization</td>
</tr>
<tr>
<td>p53</td>
<td>Various types of cancer</td>
<td>Mutations, post-translational</td>
<td>Cytoplasm</td>
</tr>
</tbody>
</table>
Mislocalization of PKC => Cytokinesis Failure

Control

Control

Microtubules
Nuclei

© http://www.jbc.org/content/279/6.cover-expansion
D Chen, AC Newton, et al. (2004) JBC
Why predict subcellular localization?

- Knowledge of subcellular localization of a protein provides information about its functional role.
- Improved target identification during the drug discovery process.
- Aberrant protein subcellular location has been observed in the cells of several diseases.
- Can help validate or analyze protein-protein interactions.
Prediction of subcellular localization
Recap: zip-codes
Localization Predictions Indispensable

High-throughput methods
- cost money
- not accurate
- not complete

SWISS-PROT: *Saccharomyces cerevisiae*
- 6,621 manually annotated entries
- 4,935 have localization information (Nov 2014)
Localization is ideal for Predictions
Trafficking in the Cell

Key:
- Orange: gated transport
- Green: transmembrane transport
- Blue: vesicular transport

Diagram showing cell organelles and pathways for different types of transport:
- Cytosol
- Nucleus
- Mitochondria
- Peroxisome
- Plastids
- ER
- Golgi
- Lysosome
- Endosome
- Secretory vesicles
- Cell surface

B Alberts, et al. (1994) *The Cell*
A majority of signal peptides are still unknown

For signal patches the situation is even worse
Protein Trafficking via Sorting Signals

1999 Nobel Prize in Physiology/Medicine given to Günter Blobel

„for the discovery that proteins have intrinsic signals that govern their transport and localization in the cell“

=> „address tag“ or „zip code“

http://www.time.com

© Tatyana Goldberg
© Burkhard Rost
© ROSTLAB.TUM
Predicting Localization

1) Sorting signals
   (signalP, chloroP)

2) Homology-based inference
   (R Nair and B Rost, 2002)

3) Text-based analysis
   (LOCkey)

4) De novo
   (CELO v.2.5)

5) Hybrid approaches
   (LOCtree, MultiLoc2, WoLF PSORT)
III.11 LocTree2
De novo prediction

LocTree2 predicts localization for all domains of life
Tatyana Goldberg¹,*,†, Tobias Hamp¹,† and Burkhard Rost¹,²
¹TUM, Bioinformatik-I12, Boltzmannstrasse 3, Gaaching 85748, Germany and ²New York Consortium on Membrane Protein Structure (NYCOMPS) and Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA
New method: LocTree2

LocTree2 predicts localization for all domains of life
Tatyana Goldberg¹,*,+ , Tobias Hamp¹,+ and Burkhard Rost¹,2
¹TUM, Bioinformatik-I/12, Informatik, Boltzmannstrasse 3, Garching 85748, Germany and ²New York Consortium on Membrane Protein Structure (NYCOMPS) and Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA

Eukaryota

Bacteria

Archaea
Novel method

Data Set Preparation

Training the Prediction Method

Testing

Comparison to External Tools

© Tatyana Goldberg
© Burkhard Rost
ROSTLAB.TUM

5-fold cross-validation

10-fold cross-validation

Archaea: 3 classes, 59 sequences
Bacteria: 6 classes, 479 sequences
Eukaryota: 18 classes, 1682 sequences

SVMs

Kernel Selection

Multiclass Classifier Selection

SVMs Parameter Optimization

© Tatyana Goldberg
© Burkhard Rost
ROSTLAB.TUM
LocTree - data
Novel method

Data Set Preparation

- 5-fold cross-validation
- Training the Prediction Method
- Testing
- Comparison to External Tools

Archaea: 3 classes, 59 sequences
Bacteria: 6 classes, 479 sequences
Eukaryota: 18 classes, 1682 sequences

SVMs
- Kernel Selection
- Multiclass Classifier Selection
- SVMs Parameter Optimization

© Tatyana Goldberg
© Burkhard Rost
© ROSTLAB
Data Set for Development

<table>
<thead>
<tr>
<th>ID</th>
<th>TMM18_HUMAN</th>
<th>Reviewed; 140 AA.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>Q96B42; D6W4X9; Q8N5H2; Q9NTH3;</td>
<td></td>
</tr>
<tr>
<td>OC</td>
<td>Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;</td>
<td></td>
</tr>
<tr>
<td>OC</td>
<td>Mammalia; Eutheria; Euarchontognathi; Primates; Haplorrhini;</td>
<td></td>
</tr>
<tr>
<td>OC</td>
<td>Catarrhini; Hominidae; Homo.</td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>-!- SUBCELLULAR LOCATION: Nucleus membrane; Multi-pass membrane protein.</td>
<td></td>
</tr>
</tbody>
</table>

- Exclusion of proteins with non-experimental annotations and of proteins with unclear or multiple localizations
- Identification of transmembrane proteins by „Single-pass“ or „Multi-pass“ keywords in the CC lines
Data set: bias and homology

Proteins of known localization

All proteins

unique
Data set distribution

SWISS-PROT annotated (2011_04 release)

HSSP-value ≤ 0

BLAST E-value > 10^{-3}

# proteins

Archaea
Bacteria
Eukaryota

© Tatyana Goldberg

© Burkhard Rost
From annotation to non-redundant set

SWISS-PROT annotated

[Diagram showing a circle representing SWISS-PROT annotated proteins and a rectangle divided into test and training sets]
From annotation to non-redundant set

SWISS-PROT annotated

HSSP-value ≤ 0
HSSP-value ≤ 60

Size increase by almost a factor of 4!

© Tatyana Goldberg
Stratified k-fold cross-validation

**K-fold cross testing**: random partitioning into $k$ equally sized subsets, use each subset for testing exactly once and the remaining $k-1$ subsets for training

**Stratified**: each subset contains about the same proportion of class labels as the original data set
LocTree - method
Novel Method

Data Set Preparation

Training the Prediction Method

5-fold cross-validation

Testing

Comparison to External Tools

Archaea: 3 classes, 59 sequences
Bacteria: 6 classes, 479 sequences
Eukaryota: 18 classes, 1682 sequences

SVMs

Kernel Selection

Multiclass Classifier Selection

SVMs Parameter Optimization

10-fold cross-validation
- Linear separation by building a hyperplane
- Hyperplane with the maximum margin is the best

What if linear SVM not separated?

Map data to a higher-dimensional *feature space* for a linear separation

A *Kernel function* performs this mapping and computes the distance between two vectors in the feature space.
String kernels

Idea: find similarity between two protein sequences by looking at the number of common $k$-mers

   - Text classification
   - 3 parameters ($k$, decay factor for non-continuous $k$, $l=k+\text{gap}$)

   - Protein classification
   - 2 parameters ($k$, number of mismatches)

   - Protein classification
   - 2 parameters and evolutionary profile ($k$, min conservation)
Multiple sequence alignment

query MKWLGLLGLVALSECLVTIPLMKVKSDSEEPQRVAIKSLRLEK
homo1 MKTFGVLGLVTLSECLVTIPLVKIKSLRENLRKDKMKEYLEK
homo2 MMAFGVLSETTLSECLVTIPLVKILLSLRENLRKDKMMLWWYLEK
homo3 MMAFGVLSETTLSTTCLVTIPLVKILLSLRENREKRDKMMLWWYLEK

20 AAs

<table>
<thead>
<tr>
<th>A</th>
<th>M</th>
<th>.</th>
<th>T</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>M 0</td>
<td>1 0</td>
<td>0 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K 0</td>
<td>0.5</td>
<td>0 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W 0.5</td>
<td>0</td>
<td>0.25</td>
<td>0.25</td>
<td></td>
</tr>
</tbody>
</table>

-\log
(frequency)

sequence
Profile kernel example

Prot1
CATLGLLGLVAL

Prot2
CARRGLLWAVAL

C Leslie, W Noble, et al. 2004 Bioinformatics
Profile kernel example

<table>
<thead>
<tr>
<th>Prot1</th>
<th>CATLGLLLGLVAL</th>
<th>Prot2</th>
<th>CARRGGLLWAVAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>e.g. $k$-mer length=3,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>conservation threshold=5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>i.e. $-\log(\text{frequency}) &lt; 5$</td>
</tr>
</tbody>
</table>

C 2 3 1 1 ...
A 1 4 3 1 ...
T 4 2 1 4 ...

C Leslie, W Noble, et al. 2004 *Bioinformatics*
Profile kernel example

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

<table>
<thead>
<tr>
<th>Prot2</th>
<th>CARRGLLWAVAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

AAT CAT CTQ CAW \[\text{ all possible } 3\text{-mers } = 20^3 \]

\[ \text{e.g. } \] \[ k\text{-mer length}=3, \] \[ \text{conservation threshold}=5, \] \[ \text{i.e. } -\log(\text{frequency})<5 \]

C Leslie, W Noble, et al. 2004 Bioinformatics

© Tatyana Goldberg
© Burkhard Rost

44/99
### Profile kernel example

**Prot1**  
CATLGLLGLVAL

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>R</th>
<th>T</th>
<th>Q</th>
<th>...</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1 ...</td>
<td>5</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1 ...</td>
<td>5</td>
</tr>
<tr>
<td>T</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>1 ...</td>
<td>3</td>
</tr>
</tbody>
</table>

**Prot2**  
CARRGLLWAVAL

E.g. $k$-mer length=3,  
conservation threshold=5  
i.e. $-\log(\text{frequency})<5$

$$\Phi(\text{Prot1}) = (0, \ldots, 0, 0, 0, 0, 0, \ldots, 0, \ldots, 0)$$

\[
\begin{align*}
\text{CAT} & \rightarrow \text{AAT} \\
& \quad (2+1+1<5) \\
\text{CAT} & \rightarrow \text{CAW} \\
& \quad (1+1+3=5) \\
\text{CAT} & \rightarrow \text{CQT} \\
& \quad (1+1+1<5) \\
\end{align*}
\]
Profile kernel example

<table>
<thead>
<tr>
<th>Prot1</th>
<th>Prot2</th>
</tr>
</thead>
<tbody>
<tr>
<td>CATLGLLGLVAL</td>
<td>CARRGLLWAVAL</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>R</th>
<th>T</th>
<th>Q</th>
<th>...</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>5</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>...</td>
<td>5</td>
</tr>
<tr>
<td>T</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>...</td>
<td>3</td>
</tr>
</tbody>
</table>

... 

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

- e.g. \( k \)-mer length=3, conservation threshold=5
  - i.e. \(-\log(\text{frequency})<5\)
  - \(\Phi(\text{Prot1})\) for \(k\)-mer length=3, conservation threshold=5:
    - CATLGLLGLVAL
      - \(\text{CAT}\): \(2+1+1<5\)
      - \(\text{AAT}\): \(1+1+3=5\)
      - \(\text{CAN}\): \((1+1+1<5)\) 
      - \(\text{CAT}\): \((1+1+1<5)\) 
      - \(\text{CAT}\): \((1+1+1<5)\) 

C Leslie, W Noble, et al. 2004 *Bioinformatics*
Profile kernel example

**Prot1**
CATLGLLGLVAL

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>R</th>
<th>T</th>
<th>Q</th>
<th>...</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>5</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>...</td>
<td>5</td>
</tr>
<tr>
<td>T</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>...</td>
<td>3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Prot2**
CARRGLLWAVAL

e.g. $k$-mer length=3,
conservation threshold=5
i.e. $-\log(\text{frequency})<5$

$$\Phi(\text{Prot1}) = (0, ..., 1, 1, 1, 0, ..., 0, ..., 0)$$

C Leslie, W Noble, et al. 2004 *Bioinformatics*
Profile kernel example

\[
\Phi(\text{Prot1}) = (0, \ldots, 1, 1, 1, 0, \ldots, 0, \ldots, 0)
\]
\[
\Phi(\text{Prot2}) = (0, \ldots, 0, 1, 1, 0, \ldots, 0, \ldots, 0)
\]
\[
K(\text{Prot1}, \text{Prot2}) = \Phi(\text{Prot1}) \cdot \Phi(\text{Prot2}) = 32
\]

C Leslie, W Noble, et al. 2004 Bioinformatics

© Tatyana Goldberg
© Burkhard Rost
ROSTLAB. TUM
## Kernel selection

<table>
<thead>
<tr>
<th></th>
<th>String subsequence kernel</th>
<th>Mismatch Kernel</th>
<th>Profile Kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Archaea (2 classes)</strong></td>
<td>97 ± 3</td>
<td>97 ± 3</td>
<td>97 ± 3</td>
</tr>
<tr>
<td><strong>Bacteria (6 classes)</strong></td>
<td>88 ± 2</td>
<td>88 ± 2</td>
<td>93 ± 2</td>
</tr>
<tr>
<td><strong>Eukaryota (10 classes)</strong></td>
<td>69 ± 1</td>
<td>NA</td>
<td>81 ± 1</td>
</tr>
</tbody>
</table>

- for each kernel several parameter combinations tested
- values average over five training sets
- one-against-all multiclass classifiers

The Profile Kernel is a clear winner.
MultiClass classifier selection

1. One – Against – All
(V Vapnik 1995 The Nature of Statistical Learning Theory)

2. Ensemble on Nested Dichotomies
(S Kramer and E Frank 2004 Proceedings of ICML)

3. Ensemble of Class Balanced Dichotomies
(E Frank, S Dong, et al. 2005 PKDD)

4. Ensemble of Data Balanced Dichotomies
(E Frank, S Dong, et al. 2005 PKDD)

5. Nested Dichotomy of a fixed structure

V Vapnik 1995 The Nature of Statistical Learning Theory
MultiClass classifier #1: one vs all

Classification of data with labels from >2 classes

- Train $n$ binary classifiers, one for each class against all other classes
- Predicted class is the class of the classifier with the highest value

V Vapnik 1995 *The Nature of Statistical Learning Theory*
MultiClass classifier #2: Ensemble of Nested Dichotomies (ENDs)

decompose classes into a system of *nested dichotomies* (NDs)

• Every ND is equally probable!
• Take an *Ensemble of Nested Dichotomies* (ENDs)
• Result of END = average of estimates over individual NDs

J Fox 1997 Applied regression analysis, linear models, and related methods.
S Kramer and E Frank (2004) *Proceedings of ICML*
MultiClass classifiers #3-#4: ECBNDs+EDBNDs

ECBNDs: *Ensemble of Class Balanced NDs*
- Each internal node of a binary tree is class balanced

EDBNDs: *Ensemble of Data Balanced NDs*
- Each internal node of a binary tree is data balanced

Limited number of possible NDs
MultiClass classifier #5: MyND

Nested Dichotomy of a fixed architecture
LocTree2 prokaryotes

Archaea (3 classes)

SVM

SVM

CYT

PM

EXT

CYT: cytosol, EXT: extra-cellular, PM: plasma membrane
LocTree2 prokaryotes

Archaea (3 classes)

- SVM
- CYT
- PM
- EXT

Plasma membrane

Cytosol

Extra-cellular

**CYT**: cytosol, **EXT**: extra-cellular, **PM**: plasma membrane
LocTree2 prokaryotes

Archaea (3 classes)

- SVM
  - CYT
    - PM
    - EXT

Plasma membrane

Cytosol

Extra-cellular

CYT: cytosol, EXT: extra-cellular, PM: plasma membrane
**LocTree2 prokaryotes**

Archaea (3 classes)

Bacteria (6 classes)

**CYT**: cytosol, **EXT**: extra-cellular, **FIM**: fimbrium, **OM**: outer membrane, **PERI**: periplasmic space, **PM**: plasma membrane

T Goldberg, T Hamp & B Rost (2012) Bioinformatics 28:i458-65
Archaea (3 classes)

- PM: plasma membrane
- EXT: extra-cellular
- CYT: cytosol

SVM

Plasma membrane

Cytosol

Extra-cellular

CYT: cytosol, EXT: extra-cellular, PM: plasma membrane
**LocTree2 eukaryotes**


T Goldberg, T Hamp & B Rost (2012) Bioinformatics 28:i458-65
What prediction is shown?
What prediction is shown?
LocTree2 eukaryotic “path”

soluble → SVM → trans-membrane

© Tatyana Goldberg
LocTree2 eukaryotic “path”

soluble \( \sim \) SVM \( \sim \) trans-membrane

SVM

SVM

T Goldberg, T Hamp & B Rost (2012) Bioinformatics 28:i458-65
xxbr: continue here
LocTree2 eukaryotic “path”

soluble

trans-membrane

not

secretory pathway

© Tatyana Goldberg

© Burkhard Rost

T Goldberg, T Hamp & B Rost (2012) Bioinformatics 28:i458-65
## MultiClass classifier selection

<table>
<thead>
<tr>
<th></th>
<th>one-vs-all</th>
<th>ENDs</th>
<th>ECBNDs</th>
<th>EDBNDs</th>
<th>MyND</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Archaea</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 classes</td>
<td>97 ± 3</td>
<td>97 ± 3</td>
<td>97 ± 3</td>
<td>97 ± 3</td>
<td>97 ± 3</td>
</tr>
<tr>
<td><strong>Bacteria</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 classes</td>
<td>93 ± 2</td>
<td>96 ± 2</td>
<td>96 ± 2</td>
<td>96 ± 2</td>
<td>96 ± 2</td>
</tr>
<tr>
<td><strong>Eukaryota</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 classes</td>
<td>81 ± 1</td>
<td>86 ± 1</td>
<td>86 ± 1</td>
<td>86 ± 1</td>
<td>86 ± 1</td>
</tr>
</tbody>
</table>

- Averaged overall accuracies over five training sets are reported
- No difference in accuracy between ND-based approaches
- Selected ENDs (ensemble of random NDs) and MyND (ND of a fixed structure) for the next step

T Goldberg, T Hamp & B Rost (2012) Bioinformatics 28:i458-65
SVM parameter optimization

- Increased performance by 1%
- No difference in accuracy between END and MyND

<table>
<thead>
<tr>
<th>Class</th>
<th>END</th>
<th>MyND</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archaea</td>
<td>$6.8 \times 10^3$</td>
<td>$40 \times 10^3$</td>
</tr>
<tr>
<td>3 classes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacteria</td>
<td>$2 \times 10^3$</td>
<td>$15 \times 10^3$</td>
</tr>
<tr>
<td>6 classes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eukaryota</td>
<td>$0.2 \times 10^3$</td>
<td>$6.1 \times 10^3$</td>
</tr>
<tr>
<td>10 classes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MyND up to 30 times faster!

Note: PSI-BLAST profiles and kernel values pre-computed

T Goldberg, T Hamp & B Rost (2012) Bioinformatics 28:i458-65
LocTree - testing
Accuracy (or precision or specificity)

<table>
<thead>
<tr>
<th>Predicted to be in $L$</th>
<th>Predicted to be in not-$L$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed to be in $L$</td>
<td>True Positives; $TP$</td>
</tr>
<tr>
<td>Observed to be in not-$L$</td>
<td>False Positives; $FP$</td>
</tr>
</tbody>
</table>

$$Acc(L) = 100 \frac{TP}{TP + FP}$$

- Number of **correctly** predicted proteins that are observed to be in localization $L$ divided by the **total** number of proteins predicted in $L$

- How often the predicted localization class is correct (also called „precision“ or „specificity“)
### Coverage (or recall or sensitivity)

<table>
<thead>
<tr>
<th>Predicted to be in $L$</th>
<th>Predicted to be in not-$L$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed to be in $L$</td>
<td>True Positives; $TP$</td>
</tr>
<tr>
<td>Observed to be in not-$L$</td>
<td>False Positives; $FP$</td>
</tr>
</tbody>
</table>

\[
Cov(L) = 100 \frac{TP}{TP + FN}
\]

- Number of **correctly** predicted proteins that are observed to be in localization $L$ divided by the **total** number of proteins observed in $L$
- How often the observed localization class is predicted correctly (also called „recall“ or „sensitivity“)
More digits imply better analysis?
Details reported in results

<table>
<thead>
<tr>
<th>Method</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>AND1</td>
<td>73.7654321</td>
</tr>
<tr>
<td>Web2</td>
<td>69.99345</td>
</tr>
</tbody>
</table>
**Details reported in results**

<table>
<thead>
<tr>
<th>Method</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web2</td>
<td>73.76543</td>
</tr>
<tr>
<td>Web1</td>
<td>69.99345</td>
</tr>
</tbody>
</table>

Better method -> more
digits=NONSENSE!
# How many digits?

<table>
<thead>
<tr>
<th>Method</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>AND1</td>
<td>73.76543</td>
</tr>
<tr>
<td>Web2</td>
<td>69.99345</td>
</tr>
</tbody>
</table>
**How many digits?**

<table>
<thead>
<tr>
<th>Method</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A'</td>
<td>73.76543</td>
</tr>
<tr>
<td>Web2</td>
<td>69.99345</td>
</tr>
</tbody>
</table>
Prediction error varies

<Q3>=72.3% ; sigma=10.5%

Example: secondary structure prediction (data outdated)
Prediction error varies

\[
\langle Q_3 \rangle = 72.3\% \ ; \ \text{sigma}=10.5\%
\]

Example: secondary structure prediction (data outdated)

\[
\text{sigma} \rightarrow \text{standard error: how?}
\]
Prediction error varies

Example: secondary structure prediction (data outdated)

simplified rule of thumb:
stderr = sigma/SQRT(number of samples)
here: about ±0.4
## How many digits?

<table>
<thead>
<tr>
<th>Method</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>AND1</td>
<td>73.76543</td>
</tr>
<tr>
<td>Web2</td>
<td>69.99345</td>
</tr>
</tbody>
</table>

`stderr ±0.8: how many digits?`
### How many digits?

<table>
<thead>
<tr>
<th>Method</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>AND1</td>
<td>73.76543</td>
</tr>
<tr>
<td>Web2</td>
<td>69.99345</td>
</tr>
</tbody>
</table>

**stderr ±0.8: how many digits?**
What else to consider?

<table>
<thead>
<tr>
<th>Method</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>AND1</td>
<td>73</td>
</tr>
<tr>
<td>Web2</td>
<td>69</td>
</tr>
</tbody>
</table>
## compare to worst (random)

<table>
<thead>
<tr>
<th>Method</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>AND1</td>
<td>73</td>
</tr>
<tr>
<td>Web2</td>
<td>69</td>
</tr>
<tr>
<td>random</td>
<td>33</td>
</tr>
</tbody>
</table>
What else to compare to?

<table>
<thead>
<tr>
<th>Method</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp. error</td>
<td>95</td>
</tr>
<tr>
<td>AND1</td>
<td>73</td>
</tr>
<tr>
<td>Web2</td>
<td>69</td>
</tr>
<tr>
<td>random</td>
<td>33</td>
</tr>
</tbody>
</table>
Which method is best?

<table>
<thead>
<tr>
<th>Method</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp. error</td>
<td>95</td>
</tr>
<tr>
<td>AND1</td>
<td>73</td>
</tr>
<tr>
<td>new3</td>
<td>73.5</td>
</tr>
<tr>
<td>Web2</td>
<td>69</td>
</tr>
<tr>
<td>random</td>
<td>33</td>
</tr>
<tr>
<td>Method</td>
<td>Q3</td>
</tr>
<tr>
<td>--------</td>
<td>-------</td>
</tr>
<tr>
<td>Exp. error</td>
<td>95</td>
</tr>
<tr>
<td>AND1</td>
<td>73</td>
</tr>
<tr>
<td>new3</td>
<td>73.5</td>
</tr>
<tr>
<td>Web2</td>
<td>69</td>
</tr>
</tbody>
</table>

stderr ±0.8: (new3, AND1) do NOT differ significantly
## How to rank methods?

<table>
<thead>
<tr>
<th>Method</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp. error</td>
<td>95</td>
</tr>
<tr>
<td>A</td>
<td>73.1</td>
</tr>
<tr>
<td>B</td>
<td>73.5</td>
</tr>
<tr>
<td>C</td>
<td>72.2</td>
</tr>
<tr>
<td>D</td>
<td>71.5</td>
</tr>
<tr>
<td>E</td>
<td>69.0</td>
</tr>
<tr>
<td>random</td>
<td>33</td>
</tr>
</tbody>
</table>
Prediction error: Case 1 & 2

Example: secondary structure prediction (data outdated)

Per-residue accuracy ($Q_3$)

$<Q_3>=72.3\%$ ; $\sigma=10.5\%$

Number of protein chains

Localization (imagined data)
How to get error estimates from binary values?
Measure prediction error

• How reliable are the estimates?
• What is the **standard error**?

**Bootstrapping**

• “To pull oneself up by one’s bootstraps”  
  (*The Adventures of Baron Munchausen*)
• Introduced: Bradley Efron in 1979
• Popularized in 1980s with advent of computers in statistics
Bootstrapping: the concept

- Start with your set of predictions
- Randomly draw a subset of $n$ predictions from the original set
- Compute an estimate $x$ for this subset of predictions
- Repeat previous two steps $m$ times

$\Rightarrow$ bootstrap estimates $x_1, \ldots, x_m$

$$\text{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{m} (x_i - \bar{x})^2}{n}}$$

$$\text{Standard Error} = \frac{\text{Standard Deviation}}{\sqrt{n-1}}$$
LocTree2 eukaryotic "path"
Performance: **LocTree2** $Q_{18}=65\pm2\%$

**Diagram Description**

- SVM (Support Vector Machine)
- ER: endoplasmic reticulum
- GOL: Golgi apparatus
- VAC: vacuole
- CYT: cytosol
- NUC: nucleus
- PER: peroxisome
- PM: plasma membrane
- PLAS: plastid
- CHL: chloroplast
- MIT: mitochondria
- EXT: extra-cellular

**Cellular Location**

- CHL: chloroplast
- CYT: cytosol
- ER: endoplasmic reticulum
- EXT: extra-cellular
- GOL: Golgi apparatus
- MIT: mitochondria
- NUC: nucleus
- PER: peroxisome
- PLAS: plastid
- PM: plasma membrane
- SVM: support vector machine
- VAC: vacuole

**Reference**

T Goldberg, T Hamp & B Rost (2012) Bioinformatics 28:i458-65
Performance: *LocTree2 1st level*: $Q_2 = 94 \pm 2\%$

soluble $\xrightarrow{\text{SVM}}$ trans-membrane

T Goldberg, T Hamp & B Rost (2012) Bioinformatics 28:i458-65
Performance: $\text{LocTree2 1st level: } Q2=94\pm2\%$

PolyPhobius=95±2%
Performance: *LocTree2 cross-validation*

### SVM
- **Accuracy**
  - SVM: $80\pm4$
  - SVM: $91\pm3$
- **Coverage**
  - SVM: $45\pm8$
  - SVM: $67\pm6$

### ER
- **Accuracy**
  - SVM: $45\pm8$
  - SVM: $44\pm8$
- **Coverage**
  - SVM: $44\pm8$
  - SVM: $77\pm6$

### GOL
- **Accuracy**
  - SVM: $45\pm10$
  - SVM: $46\pm10$
- **Coverage**
  - SVM: $38\pm48$
  - SVM: $60\pm15$

### VAC
- **Accuracy**
  - SVM: $45\pm15$
  - SVM: $42\pm14$
- **Coverage**
  - SVM: $44\pm13$
  - SVM: $46\pm10$

### CYT
- **Accuracy**
  - SVM: $44\pm9$
  - SVM: $29\pm9$
- **Coverage**
  - SVM: $21\pm28$
  - SVM: $29\pm9$

### NUC
- **Accuracy**
  - SVM: $50\pm50$
  - SVM: $44\pm13$
- **Coverage**
  - SVM: $27\pm30$
  - SVM: $44\pm13$

### PM
- **Accuracy**
  - SVM: $44\pm15$
  - SVM: $42\pm14$
- **Coverage**
  - SVM: $44\pm13$
  - SVM: $60\pm15$

### EXT
- **Accuracy**
  - SVM: $80\pm4$
  - SVM: $91\pm3$
- **Coverage**
  - SVM: $45\pm8$
  - SVM: $67\pm6$

### MIT
- **Accuracy**
  - SVM: $45\pm10$
  - SVM: $46\pm10$
- **Coverage**
  - SVM: $38\pm48$
  - SVM: $60\pm15$

**Accuracy** = observed / predicted

Accuracy values range from $80\pm4$ to $91\pm3$ for different categories. Coverage values range from $44\pm9$ to $67\pm6$.

© Tatyana Goldberg
© Burkhard Rost

T Goldberg, T Hamp & B Rost (2012) Bioinformatics 28:i458-65
**LocTree2 competitive for new proteins**

<table>
<thead>
<tr>
<th>New Swiss-Prot : Bacteria</th>
<th>New Swiss-Prot : Eukaryota</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LocTree2</strong></td>
<td><strong>LocTree2</strong></td>
</tr>
<tr>
<td>PSORTb 3.0*</td>
<td>WoLF PSORT**</td>
</tr>
<tr>
<td><strong>Q5</strong>: 86 ± 16</td>
<td><strong>Q9</strong>: 65 ± 14</td>
</tr>
<tr>
<td>71 ± 21</td>
<td>62 ± 14</td>
</tr>
<tr>
<td><strong>LocTree2</strong></td>
<td><strong>LocTree2</strong></td>
</tr>
<tr>
<td>LocTree°</td>
<td>LocTree</td>
</tr>
<tr>
<td><strong>Q3</strong>: 86 ± 18</td>
<td><strong>Q5</strong>: 66 ± 15</td>
</tr>
<tr>
<td>72 ± 21</td>
<td>62 ± 17</td>
</tr>
</tbody>
</table>

* NY Yu et al. (2010) *Bioinformatics*
° R Nair and Rost (2005) *JoMB*
** P Horton et al. (2007) *NAR*
## LocTree2 handles sequencing mistakes

Sequence-unique eukaryotic proteins from Swiss-Prot and LocDB databases

<table>
<thead>
<tr>
<th></th>
<th>Full length</th>
<th>30N removed</th>
<th>30C removed</th>
<th>1/3 randomly removed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LocTree 2</strong></td>
<td>62 ± 2</td>
<td>54 ± 2</td>
<td>60 ± 2</td>
<td>53 ± 2</td>
</tr>
<tr>
<td><strong>CELLO v. 2.5</strong></td>
<td>56 ± 2</td>
<td>35 ± 2</td>
<td>47 ± 2</td>
<td>48 ± 2</td>
</tr>
<tr>
<td><strong>WoLF PSORT</strong></td>
<td>56 ± 2</td>
<td>40 ± 2</td>
<td>52 ± 2</td>
<td>49 ± 2</td>
</tr>
</tbody>
</table>
preliminary Lecture plan (CB2 function)

01: 2019/10/15: No lecture (makeup examen: PP last year)
02: 2019/10/17: No lecture (makeup)
03: 2019/10/22: No lecture
04: 2019/10/24: Welcome: who we are
05: 2019/10/29: Intro function 1: concept of protein function
06: 2019/10/31: No lecture (holiday, All Saints)
07: 2019/11/05: Localization 1 (chalk talk)
08: 2019/11/07: Localization 2 (homology, motifs)
09: 2019/11/12: No lecture (SVV)
10: 2019/11/14: Localization 3 (machine learning)
11: 2019/11/19: Localization 4 (machine learning)
13: 2019/11/26: PPI 1 - sites (chalk)
14: 2019/11/28: No lecture (Thanksgiving)
15: 2019/12/03: PPI 2 - sites
16: 2019/12/05: No lecture (Dies Academicus)
17: 2019/12/10: PPI 3 - sites / PPI pairing
18: 2019/12/12: PPI 4 - sites / DNA / RNA (Jia Jun Qiu) / small molecules (Maria Littmann)
19: 2019/12/17: PPI 3 - sites / DNA / RNA (Jia Jun Qiu)
20: 2019/12/19: No lecture
21-24: no lectures - winter break (2019/12/24 - 2020/01/06)
25: 2020/01/07: No lecture
26: 2020/01/09: PPI 4 - sites: DNA / RNA (Jia Jun Qiu) + PPI pairing 1
27: 2020/01/14: SAV effect 1 (chalk talk)
28: 2020/01/16: SAV effect 2
29: 2020/01/21: SAV effect 3
30: 2020/01/23: WRAP up 1
31: 2020/01/28: WRAP up 2
32: 2020/01/30: Examen (10:00-13:00, lecture room TBA - LMU physics?)
33: 2020/02/04: TBA
34: 2020/02/06: TBA