title: Computational Biology 2 - Protein function:

Prediction of sub-cellular localization:

Development of a prediction tool

short title: cb2_localization2

lecture: Protein Prediction 2 - Protein function
Computational Biology 2 - TUM Winter 2014/15
Announcements

- Videos: YouTube / [www.rostlab.org](http://www.rostlab.org)

THANKS:
- Tim Karl + Jonas Reeb

Special lectures:
- 12/16&18 – Andrea Schafferhans

No lecture:
- Nov 12 Tue (Student assembly)
- Dec 12 Thu (TUM Dies Academicus)

LAST lecture: January 20

Examen: January 22
- Makeup: Apr 14, 2015 - morning/noon

CONTACT: Tanya Goldberg [goldberg@rostlab.org](mailto:goldberg@rostlab.org)
Protein function
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

© http://www.pdb.org

© Burkhard Rost (TUM Munich)
Sequence to Function Gap

- UniProtKB: more than 87 Mio protein sequences (2014_11)
- UniProtKB/Swiss-Prot: 0.5 Mio manually annotated entries (2014_11)

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

MTSHSYYKDRLGFDPNEQ QPGSNNSMKRSSRRQTTHHHQSYHHATTSSSQSPA RISVSPGGNNGTLEYQQV QRENNW…

Describing Protein Function

Gene Ontology (GO)

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

Stats: Nov 2013

GO: Cellular Component for HIST1H2AI

Cellular Compartmentalization

Prokaryotic Cell
(bacillus type)

Eukaryotic Cell

Common compartment  Common physiological function

K Rogers (2011) Britannica
Sub-cellular localization: application
Drug targets tend to be found in membranes, cytoplasm or are extra-cellular!
<table>
<thead>
<tr>
<th>Subcellular target</th>
<th>Drug or compound</th>
<th>Mode of action</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plasma membrane</td>
<td>Enfuvirtide C34</td>
<td>Lipid linkage promotes membrane targeting of inhibitors and better inhibition of the HIV fusion complex</td>
<td>HIV fusion inhibitors</td>
</tr>
<tr>
<td></td>
<td>Myr-proS1</td>
<td>Myristoylated or stearoylated ProS1 domain of HBV targets the fusion complex</td>
<td>HBV fusion inhibitor</td>
</tr>
<tr>
<td></td>
<td>Pepducin</td>
<td>Palmitoylation of the i3 loop of GPCRs efficiently modulates GPCR signalling</td>
<td>GPCR modulator</td>
</tr>
<tr>
<td>Early endosomes</td>
<td>Cholesterol-linked β-secretase inhibitor</td>
<td>Addition of cholesterol promotes membrane tethering and endocytosis into endosomes, in which the active enzyme is localized</td>
<td>Efficient inhibition of the β-secretase enzyme; a therapeutic target in Alzheimer's disease</td>
</tr>
<tr>
<td></td>
<td>Anticancer drugs,</td>
<td>Ligands of cell surface receptors (folate, LDL cholesterol and transferrin) mediate endocytosis of the heterologous conjugates</td>
<td>Effective transport of anticancer drugs to the interior of the cell</td>
</tr>
<tr>
<td></td>
<td>Denileukin dfitotox IL-4 and IL-13</td>
<td>Diptheria toxin conjugates and Pseudomonas exotoxin conjugates enable the uptake of the IL and release in the intracellular site</td>
<td>Malignant lymphomas</td>
</tr>
<tr>
<td>Late endosomes and lysosomes</td>
<td>β-glucosidase,</td>
<td>Replacement of lysosomal enzymes directly or by mannose-6-phosphate-mediated uptake</td>
<td>Enzyme replacement therapy for lysosomal storage diseases (Gaucher's and Fabry's disease)</td>
</tr>
<tr>
<td></td>
<td>β-hexosaminidase</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cyclo-2</td>
<td>Cholesterol-sequestering agent</td>
<td>Niemann–Pick's disease</td>
</tr>
<tr>
<td>Endoplasmic reticulum and Golgi complex</td>
<td>Antigenic peptides</td>
<td>Delivery of conjugated antigens for presentation on MHC class I complex by conjugation to STX-B</td>
<td>Malignant lymphomas, ovarian cancer, and intestinal adenocarcinomas</td>
</tr>
<tr>
<td></td>
<td>Fluorescent cancer imaging dyes</td>
<td>Imaging of tumours by STX-B conjugates of fluorescent dyes, as STX-B binds to GB3, which is overexpressed by tumours</td>
<td>Colon cancer, liver metastasis and ovarian tumours</td>
</tr>
<tr>
<td></td>
<td>Shiga holotoxin</td>
<td>Selective killing of GB3-overexpressing tumours</td>
<td></td>
</tr>
<tr>
<td>Cytosolic delivery</td>
<td>Anticancer drugs, siRNAs and plasmids</td>
<td>Conjugation with cell-penetrating peptides such as Tat and VP22 enables transport of heterologous conjugates into the cell</td>
<td>Delivery of anticancer drugs, siRNAs, plasmid DNA and proteins</td>
</tr>
<tr>
<td>Mitochondria</td>
<td>Gamitrinibs</td>
<td>Selective targeting of the HSP90 network in cancerous mitochondria</td>
<td>Rapid tumour cell death</td>
</tr>
<tr>
<td></td>
<td>Antioxidant (ubiquinol and α-tocopherol)</td>
<td>Lipophilic cations such as triphenylphosphonium cations conjugated with antioxidants target mitochondria and confer protection</td>
<td>Neurodegenerative diseases</td>
</tr>
<tr>
<td>Nucleus</td>
<td>Antitumour drugs, cisplatin, doxorubicin and DNA</td>
<td>Delivery of genes by viral-mediated vectors, viral-like particles or liposomes Nanoparticles encapsulate the drug and enable slow and effective release Targeted charge-reversal nanoparticles carry conjugates to the nucleus</td>
<td>Carcinomas</td>
</tr>
</tbody>
</table>

GB3, globotriaosylceramide (also known as CD77); GPCR, G protein–coupled receptor; HBV, hepatitis B virus; HSP90, heat shock protein 90; IL, interleukin; LDL, low-density lipoprotein; MHC, major histocompatibility complex; siRNA, small interfering RNA; STX-B, Shiga toxin B subunit; RNAi, transcription protein; VP22, viral protein 22.
### Mislocalizations Associated with Human Diseases

<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 dyschondrodystosis</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 modifications</td>
<td>Cytoplasmic mislocalization</td>
</tr>
</tbody>
</table>

MC Hung and W Link (2012) JCS

© Burkhard Rost (TUM Munich) 12/00
Mislocalization of PKC => Cytokinesis Failure

Control

Control

© http://www.jbc.org/content/279/6.cover-expansion
D Chen, AC Newton, et al. (2004) JBC
Why 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 sub-cellular localization
Localization Predictions Indispensable

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

© http://www.microscopyu.com
Intracellular mitochondrial network, microtubules, nuclei

SWISS-PROT: *Saccharomyces cerevisiae*
- 6,621 manually annotated entries
- 4,935 have localization information (Nov 2014)
Localization is an easily identifiable functional feature

Process of protein trafficking is quite well understood

Localization data available in public protein databases

Most proteins active in only one compartment
Trafficcking in the Cell

Key:
- gated transport
- transmembrane transport
- vesicular transport

A majority of signal peptides are still unknown

For signal patches the situation is even worse

B Alberts et al. (1994) *The Cell*
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“

Nucleus
Intracellular Space
Eukaryotic Cell
PKKKRKV

http://www.time.com
The Nobel Prize in Physiology or Medicine 2013

James E. Rothman, Randy W. Schekman, Thomas C. Südhof

The Nobel Prize in Physiology or Medicine 2013 was awarded jointly to James E. Rothman, Randy W. Schekman and Thomas C. Südhof “for their discoveries of machinery regulating vesicle traffic, a major transport system in our cells”.

Photo: © Yale University
James E. Rothman

Photo: H. Goren. © HHMI
Randy W. Schekman

Photo: © S. Fisch
Thomas C. Südhof
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
   (CELLO v.2.5)

5) Hybrid approaches
   (LOCtree, MultiLoc2, WoLF PSORT)
Problems to Consider

- Biological data is „noisy“
  - at the biological, experimental and curation levels

- Prediction performance must be established properly

- Sequence similarity between training and test data sets

- Sequence redundancy in public databases

- Prediction results must be interpretable by users

- Benchmarking prediction methods can be a difficult task
Novel predictor: LocTree2
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
New Method: LocTree2

- One framework for all domains of life
- Robustness to sequencing errors
- Predictions for the largest number of classes so far:
  - 3 for Archaea
  - 6 for Bacteria
  - 18 for Eukaryota
- Membrane predictions
- Decision tree resembling cellular protein sorting
- Reliability score measuring the strength of a prediction
- High performance
Novel Method

Data Set Preparation

Training the Prediction Method

Testing

Comparison to External Tools

5-fold cross-validation

SVMs

Kernel Selection

Multiclass Classifier Selection

SVMs Parameter Optimization

10-fold cross-validation

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

Data Set Preparation

5-fold cross-validation

Training the Prediction Method

Testing

Comparison to External Tools

10-fold cross-validation

SVMs

Kernel Selection

Multiclass Classifier Selection

SVMs Parameter Optimization

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

© Burkhard Rost (TUM Munich) 28/00
Data Set for Development

- 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
Homology Reduction

Bias in protein databases towards certain protein families
⇒ Construction of sequence unique sets in two steps

1. HSSP-value ≤ 0

Below this threshold no reliable annotation of subcellular localization based on sequence homology

2. BLAST E-value > 10^{-3}

B Rost (2002) J Mol Biol
SF Altschul, DJ Lipman, et al. (1990) J Mol Biol
Data Statistics

- **SWISS-PROT annotated (2011_04 release)**: The highest number of proteins.
- **HSSP-value ≤ 0** and **BLAST E-value > 10^{-3}**: Significantly lower numbers, indicating reduced homology compared to the first category.

Legend:
- Blue: Archaea
- Red: Bacteria
- Green: Eukaryota
Stratified $k$-fold Cross-Validation

*K-fold Cross Testing*: a 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.
Size Increase of the Training Sets

Size increase by almost a factor of 4!
Performance Estimates
Performance Estimate: Accuracy

<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“)
Performance Estimate: Coverage

<table>
<thead>
<tr>
<th>Observed to be in $L$</th>
<th>Predicted to be in $L$</th>
<th>Predicted to be in not-$L$</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positives; $TP$</td>
<td>False Negatives; $FN$</td>
<td></td>
</tr>
<tr>
<td>False Positives; $FP$</td>
<td>True Negatives; $TN$</td>
<td></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“)
Measure for the Precision of Estimates

- 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 by Bradley Efron in 1979
- Popularized in 1980s due to the introduction of computers in statistical practice

---

B Efron (1979) *The Annals of Statistics*
Bootstrapping: the Algorithm

- 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

\[ \Rightarrow \text{bootstrap estimates } x_1, \ldots, x_m \]

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

\[ \text{Standard Error} = \frac{\text{Standard Deviation}}{\sqrt{n - 1}} \]
Novel Method

- Data Set Preparation
- 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
Linear SVM

- Linear separation by building a hyperplane
- Hyperplane with the maximum margin is the best
What if data is linearly non-separable?

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.
To be continued...