title: 2: Predict localization (de novo)
short title: cb2_localization1

lecture: Protein Prediction 2 (for Computer Science) - Protein function
TUM winter semester
CONTACT: Lothar Richter richter@rostlab.org

Announcements

☐ Videos: YouTube / www.rostlab.org

THANKS:

☐ Special lectures:
  • 12/15 - Jana Schmidt
  • 12/17 - Andrea Schafferhans
  • 01/07 Marco De Vivo, IIT Genoa (video)

☐ No lecture:
  • 11/05 - skip
  • 11/10 - SVV Student representation
  • 12/03 - Dies Academicus TUM
  • 12/22 - no lecture

☐ LAST lecture: January 14 (followed by 2 x wrap-up sessions)

☐ Examen: January 28, 2016: 11:30 room to be confirmed
  • Makeup: TBC: Apr 12 & Apr 14, 2016 - lecture time
Today: Predict localization 1

☐ LAST
  • Homology-based inference of function: sequence/motifs

☐ THIS
  • Prediction of sub-cellular localization

☐ NEXT WEEK
  • Prediction of sub-cellular localization contd
II. De novo Sub-cellular localization
Words

☐ by homology

☐ de novo

☐ ab initio
II.1 Predict localization: the problem
Mycoplasma genitalium

Protein synthesis (labels in black)
1. DNA
2. DNA polymerase
3. single-stranded-DNA binding protein (protects single-stranded portions during replication)
4. RNA polymerase
5. messenger RNA
6. ribosome
7. transfer RNA (in pink) and elongation factor
8. elongation factor Tu and Ts
9. elongation factor G
10. aminoacyl-tRNA synthetases
11. topoisomerase
12. Rec system for DNA repair: a) RecA, b) RecBC
13. chaperonin GroEL (helps folding of new proteins)
14. proteasome ClpA (destroys old proteins)

Enzymes for energy production (labels in red)
15. glycolytic enzymes
16. pyruvate dehydrogenase complex

Membrane proteins (labels in blue)
17. ATP synthase
18. secretory proteins
19. sodium pump
20. zinc transporter
21. magnesium transporter
22. ABC transporter (different ABC transporters transport different types of molecules-ABC is short for “ATP-binding cassette”)
23. magnesium transporter
24. lypoglycan (long carbohydrate chains connected to lipid in the membrane)

Illustration by David S. Goodsell, the Scripps Research Institute, UCSD, USA
Prokaryotic cell (E. coli)

http://www.uic.edu/classes/bios/bios100/lectures/cells.htm
Cellular compartments/localization

Prokaryotic Cell (bacillus type)  Eukaryotic Cell

K Rogers (2011) Britannica

© Tatyana Goldberg  
& Burkhard Rost (TUM Munich)
Localization: simplistic perspective

EXTRACELLULAR

NUCLEAR

CYTOPLASMIC
Predict sub-cellular localization

Homology

- sequence similarity
- text similarity

Motifs

De novo

- structure
- sequence
II.2 Predict localization: Homology-based inference
Concept of testing conservation

Proteins of known localization

All proteins

unique
# SWISS-PROT: transcription factor E2F-1

## Description and origin of the Protein

### Description
Transcription factor E2F1 (E2F-1) (Retinoblastoma binding protein 3) (RBBP-3) (PRB-binding protein E2F-1) (PBR3) (Retinoblastoma-associated protein 1) (RBAP-1).

### Gene name(s)
E2F1 OR RBBP3.

### Organism source
Homo sapiens (Human).

### Taxonomy
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

### Comments
**FUNCTION**
TRANSCRIPTION ACTIVATOR THAT BINDS DNA COOPERATIVELY WITH DP PROTEINS THROUGH THE E2 RECOGNITION SITE, TTTCC/GCGC, FOUND IN THE PROMOTER REGION OF A NUMBER OF GENES WHOSE PRODUCTS ARE INVOLVED IN CELL CYCLE REGULATION OR IN DNA REPLICAITION. THE DRTF1/E2F COMPLEX FUNCTIONS IN THE CONTROL OF CELL-CYCLE PROGRESSION FROM G1 TO S PHASE. E2F-1 BINDS PREFERENTIALLY RB1 PROTEIN, IN A CELL-CYCLE DEPENDENT MANNER. IT CAN MEDIATE BOTH CELL PROLIFERATION AND P53-DEPENDENT APOPTOSIS.

**SUBUNIT**
COMPONENT OF THE DRTF1/E2F TRANSCRIPTION FACTOR COMPLEX. FORMS HETERODIMERS WITH DP FAMILY MEMBERS. THE E2F-1 COMPLEX Binds SPECIFICALLY HYPOPHOSPHORYLATED RETINOBLASTOMA PROTEIN RB1. DURING THE CELL CYCLE, RB1 BECOMES PHOSPHORYLATED IN MID-TO-LATE G1 PHASE, DETACHES FROM THE DRTF1/ E2F COMPLEX, RENDERING E2F TRANSCRIPTIONALLY ACTIVE. VIRAL ONCOPROTEINS, NOTABLY E1A, T-ANTIGEN AND HPV E7, ARE CAPABLE OF SEQUESTERING RB PROTEIN, THUS RELEASING THE ACTIVE COMPLEX.

### SUBCELLULAR LOCATION
NUCLEAR.

### Keywords
Transcription regulation; Activator; DNA-binding; Nuclear protein; Phosphorylation; Cell cycle; Apoptosis; Polymorphism;
Statistical scores better when statistics kick in

R Nair & B Rost 2002 *Protein Science* 11, 2836-47
B Rost 2002 *J Mol Biol* 318, 595-608
B Rost 1999 *Prot Engng* 12, 85-94
Predict sub-cellular localization

**Homology**
- sequence similarity
- text similarity

**Motifs**

**De novo**
- structure
- sequence
done

not
II.3 Predict localization: Text analysis
## SWISS-PROT: transcription factor E2F-1

### Description and origin of the Protein

<table>
<thead>
<tr>
<th>Description</th>
<th>Transcription factor E2F1 (E2F-1) (Retinoblastoma binding protein 3) (RBBP-3) (PRB-binding protein E2F-1) (PBR3) (Retinoblastoma-associated protein 1) (RBAP-1).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene name(s)</td>
<td>E2F1 OR RBBP3.</td>
</tr>
<tr>
<td>Organism source</td>
<td>Homo sapiens (Human).</td>
</tr>
<tr>
<td>Taxonomy</td>
<td>Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.</td>
</tr>
</tbody>
</table>

### Comments

**FUNCTION**

TRANSCRIPTION ACTIVATOR THAT BINDS DNA COOPERATIVELY WITH DP PROTEINS THROUGH THE E2 RECOGNITION SITE, TTTCC/GCGC, FOUND IN THE PROMOTER REGION OF A NUMBER OF GENES WHOSE PRODUCTS ARE INVOLVED IN CELL CYCLE REGULATION OR IN DNA REPLICATION. THE DRTF1/E2F COMPLEX FUNCTIONS IN THE CONTROL OF CELL-CYCLE PROGRESSION FROM G1 TO S PHASE. E2F-1 BINDS PREFERENTIALLY RB1 PROTEIN, IN A CELL-CYCLE DEPENDENT MANNER. IT CAN MEDIATE BOTH CELL PROLIFERATION AND P53-DEPENDENT APOPTOSIS.

**SUBUNIT**

COMPONENT OF THE DRTF1/E2F TRANSCRIPTION FACTOR COMPLEX. FORMS HETERODIMERS WITH DP FAMILY MEMBERS. THE E2F-1 COMPLEX BINDS SPECIFICALLY HYPOPHOSPHORYLATED RETINOBLASTOMA PROTEIN RB1. DURING THE CELL CYCLE, RB1 BECOMES PHOSPHORYLATED IN MID-TO-LATE G1 PHASE, DETACHES FROM THE DRTF1/E2F COMPLEX, RENDERING E2F TRANSCRIPTIONALLY ACTIVE. VIRAL ONCOPROTEINS, NOTABLY E1A, T- ANTIGEN AND HPV E7, ARE CAPABLE OF SEQUESTERING RB PROTEIN, THUS RELEASING THE ACTIVE COMPLEX.

**SUBCELLULAR LOCATION**

NUCLEAR.

**Keywords**

Transcription regulation; Activator; DNA-binding; Nuclear protein; Phosphorylation; Cell cycle; Apoptosis; Polymorphism;
Trivial ‘keywords’

Localization sometimes clear from keywords

- DNA-binding -> nuclear
- Chromatin regulator -> nuclear
- Blood coagulation -> extra-cellular

Goals:
(1) find what is trivial by machine
(2) discover ‘non-obvious’ correlations

R Nair & B Rost (2002) *Bioinformatics* 18: S78-86
Localization classes in numbers

Localization annotated SWISS-PROT

Annotated (13589)

1743 Mitochondrial
2900 Extra-cellular
2642 Cytoplasmic
3478 Nuclear
1648 Chloroplast

Endoplasmic reticulum (568)
Peroxisomal (177)
Golgi (167)
Lysosomal (163)

Lipoyl, Ubiquinone
Lignin Degradation, Digestion, Collagen
Glutamate biosynthesis, Proline biosynthesis, SOS response
DNA-binding, Transcription, Homeobox
Calvin cycle, Photo-respiration

R Nair & B Rost (2002) *Bioinformatics* 18: S78-86
Extract keywords

~15K proteins

Localization annotated SWISS-PROT

Extract Keywords
**Description and origin of the Protein**

<table>
<thead>
<tr>
<th>Description</th>
<th>Transcription factor E2F1 (E2F-1) (Retinoblastoma binding protein 3) (RBBP-3) (PRB-binding protein E2F-1) (PBR3) (Retinoblastoma-associated protein 1) (RBAP-1).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene name(s)</td>
<td>E2F1 OR RBBP3.</td>
</tr>
<tr>
<td>Organism source</td>
<td>Homo sapiens (Human).</td>
</tr>
<tr>
<td>Taxonomy</td>
<td>Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria;</td>
</tr>
</tbody>
</table>

**Comments**

**FUNCTION**

TRANSCRIPTION ACTIVATOR THAT BINDS DNA COOPERATIVELY WITH DP PROTEINS THROUGH THE E2 RECOGNITION SITE, TTTCC/GCGC, FOUND IN THE PROMOTER REGION OF A NUMBER OF GENES WHOSE PRODUCTS ARE INVOLVED IN CELL CYCLE REGULATION OR IN DNA REPLICATION. THE DRTF1/E2F COMPLEX FUNCTIONS IN THE CONTROL OF CELL-CYCLE PROGRESSION FROM G1 TO S PHASE. E2F-1 BINDS PREFERENTIALLY RB1 PROTEIN, IN A CELL-CYCLE DEPENDENT MANNER. IT CAN MEDIATE BOTH CELL PROLIFERATION AND P53-DEPENDENT APOPTOSIS.

COMPONENT OF THE DRTF1/E2F TRANSCRIPTION FACTOR COMPLEX. FORMS HETERODIMERS WITH DP FAMILY MEMBERS. THE E2F-1 COMPLEX BINDS SPECIFICALLY HYPOPHOSPHORYLATED RETINOBLASTOMA PROTEIN RB1. DURING THE CELL CYCLE, RB1 BECOMES PHOSPHORYLATED IN MID-TO-LATE G1 PHASE, DETACHES FROM THE DRTF1/E2F COMPLEX, RENDERING E2F TRANSCRIPTIONALLY ACTIVE. VIRAL ONCOPROTEINS, NOTABLY E1A, T-ANTIGEN AND HPV E7, ARE CAPABLE OF SEQUESTERING RB PROTEIN, THEREFORE RELEASING THE ACTIVE COMPLEX.

**SUBUNIT**

**SUBCELLULAR LOCATION**

NUCLEAR.

**Keywords**

Transcription regulation; Activator; DNA-binding; Nuclear protein; Phosphorylation; Cell cycle; Apoptosis; Polymorphism;
# Building database of keyword vectors

<table>
<thead>
<tr>
<th></th>
<th>FAD</th>
<th>Apoptosis</th>
<th>ATP synthesis</th>
<th>Calcium-binding</th>
<th>Cell adhesion</th>
<th>Exonuclease</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>2a5b_human</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>abs_human</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>ace2_human</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>a1a1_horse</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>a1bg_equas</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>a70a_drome</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>2aba_drome</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>a412_plafa</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>aatc_rabbit</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
</tbody>
</table>

- **cytoplasm**
- **extra-cellular**
- **nuclear**
- **cellular**
Automatic reasoning system: concept

Experimental vector database

<table>
<thead>
<tr>
<th></th>
<th>FAD</th>
<th>Apoptosis</th>
<th>ATP synthesis</th>
<th>Calcium-binding</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>2a5b_human</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>abs_human</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>a1a1_horse</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>a1bg_equas</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>2aba_drome</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>a412_plafa</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
</tbody>
</table>

nuclear
extra-cellular
cytoplasmic
Automatic reasoning system: concept

**QUERY**

Keyword vector: 1 0 0 1 0 1 0 ...

Sub-vector: 1 0 0 0 0 1 0 ...

Make sub-vectors:
- 1 0 0 1 0 1 0 ...
- 1 0 0 0 0 1 0 ...
- 1 0 0 1 0 0 0 ...
- 0 0 0 1 0 1 0 ...

**Experimental vector database**

<table>
<thead>
<tr>
<th></th>
<th>FAD</th>
<th>Apoptosis</th>
<th>ATP synthesis</th>
<th>Calcium-binding</th>
</tr>
</thead>
<tbody>
<tr>
<td>2a5b_human</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>abs_human</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>a1a1_horse</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>a1bg_equas</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2aba_drome</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>a412_plafa</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Automatic reasoning system: concept

QUERY

Keyword vector 1 0 0 1 0 1 0 ...

Sub-vector 1 0 0 0 0 1 0 ...

Make sub-vectors 1 0 0 1 0 1 0 ...
1 0 0 0 0 1 0 ...
1 0 0 1 0 0 0 ...
0 0 0 1 0 1 0 ...

Experimental vector database

Entropy based classification system

Infer localization for query

R Nair & B Rost (2002) Bioinformatics 18: S78-86
Entropy based reasoning system: formula

Pick sub-vector which gives minimum entropy

\[
Shannon\ Information = - \sum_{i=1}^{N} P_i \log P_i
\]

\( P_i = \text{fraction of vectors belonging to localization class } i. \)
## Entropy based reasoning system: lookup

### Keywords Database

|   | N | U | C | Y | T |   | N | U | C | Y | T |   | N | U | C | Y | T |   | N | U | C | Y | T |
|   | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| N | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| U | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| C | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Y | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| T | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

### Query Sub-vector

\[ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 1 \ 1 \ 0 \ 0 \ 0 \]

---

R Nair & B Rost (2002) *Bioinformatics* 18: S78-86

© Burkhard Rost

ROSTLAB. TUM
What about 2nd best hit?
Automatic text analysis: 2nd good, too

![Graph showing Coverage vs. Accuracy with two lines: TopHit and Top2Hits.](image-url)
Second hit better?

![Graph showing Accuracy vs Coverage with two curves: TopHit and Top2Hits.](Image)
Automatic text analysis: 2nd good, too

R Nair & B Rost (2002) *Bioinformatics* 18: S78-86
“Tricky” correlations discovered

- **Plasmid, NADP, Acetylation, NAD, Metal-binding, Oxidoreductase**
  - Cytoplasmic

- **Copper, Signal, Metal-binding, Oxidoreductase**
  - Extracellular

- **Transit peptide, Tricarboxylic acid cycle, NAD, Oxidoreductase**
  - Mitochondrial
II.5 Predict localization: Localization motifs
Annotation transfer of localization motifs

Find a motif or a pattern in a functionally characterized family

Search for the motif pattern in a new protein

Transfer function annotation

Review: Rost et al. (2003) *CMLS* 60:2637-2650
Trafficking in the cell

B Alberts et al. (1994) The Cell Garland
Signal peptides (secretory pathway)
Signal peptides

- N-terminal (typically 15-30 residues)
- Cleaved during translocation across membrane
- Existing in all 3 kingdoms of life
- Simple "architecture"

n-region__h-region__c-region__cleavage site

n-region: often + charges
h-region: >6 residues
c-region: polar/uncharged
SignalP web site: CBS, Copenhagen
SignalP: history/methods

- **using NN (neural networks):**

- **combining methods (TargetP):**

- **using HMM (HMM, SignalP 3):**
SignalP and beyond (from CBS)

- **using NN (neural networks):**

- **combining methods (TargetP):**

- **using HMM (HMM, SignalP 3):**

- **twin-arginine signal peptides:**

- **non-classical secretion:**
Secretory pathway: other methods

plenty
Other “signal peptides”
Motif-based predictions

- **Secreted (signal peptides):**
  - SignalP (latest version 3.0)

- **Transit peptides - chloroplast:**
  - ChloroP
    - O Emanuelsson, H Nielsen and G von Heijne (1999) Protein Science 8: 978-84

- **Targeting peptides - mitochondria:**
  - MitoP

- **TargetP: all 3**
Nuclear localization signals
Types of zip-codes

sequence

structure

signal peptide

target motif


© Burkhard Rost

ROSTLAB. TUM
Shuttle into the nucleus

- **NLS**
- **M9**
- **Importin**
- **Transportin**

**Cytoplasm**

**Nucleus**
### How many NLS motifs in databases?

**ONE in PROSITE**

**bi-partite motif**

<table>
<thead>
<tr>
<th>Set</th>
<th>N NLS</th>
<th>Nprot nuc</th>
<th>Nfam nuc</th>
<th>Accuracy</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROSITE</td>
<td>1</td>
<td>96</td>
<td>31</td>
<td>90 %</td>
<td>3 %</td>
</tr>
</tbody>
</table>

# 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>KKKKRKRKEK</td>
<td>LEF-1</td>
<td>Prieve et al., 1998</td>
</tr>
<tr>
<td>KKKRRSREK</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>RRMKWKK</td>
<td></td>
<td>Moede et al., 1999</td>
</tr>
<tr>
<td>PKKKRKRV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRRRK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GKKRSKA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KAKRQR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RGRRRRQR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PPVKRERTS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PYNKRKGKP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KRx{(7,9)}PQPKKKP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KVTKRKHDNEGSGKRPK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RLKLCX{19}KTKR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RRKRIREDRX{18}RKRKR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RRRERX{4}RPKRP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KKKKEEDEGKK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRPRKIPR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PPIYPQLPAPSAPT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KDCVINKHHRNRCQYCRLQR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>APKRKGVSKC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RKKRRQRRR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPKTRRRPQRSQRKRPT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KRPMNAFIVWSDQRKK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KRPMNAFVMWAQAR RK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PPRKRTVV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YKRPCRKFIRFI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LKDVRRKRKLGPGH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KRP RP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RSSMKRK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAKRRARGY K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RKCLQAGMNLEARKT K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RRERNKMAAACKRRRR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KRMRNIAASKRKRKL</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Experimental NLS: more complicated

<table>
<thead>
<tr>
<th>NLS</th>
<th>Protein</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYGSKNTGAKKRKIDDA</td>
<td>DNAhelicaseQ1</td>
<td>Miyamoto et al., 1997</td>
</tr>
<tr>
<td>GGGx{3}KNRRx{6}RGRN</td>
<td>Nab2</td>
<td>Truant et al., 1998</td>
</tr>
<tr>
<td>KRxxxxxxxxxxKTKK</td>
<td>THOV NP</td>
<td>Weber et al., 1998</td>
</tr>
<tr>
<td>EYLSRKGKLEL</td>
<td>VirD2-Nterm</td>
<td>Tinland et al., 1992</td>
</tr>
<tr>
<td>KRPACTLKPECVQQLVCQSQEAKK</td>
<td>HCDA</td>
<td>Somasekaram et al., 1999</td>
</tr>
<tr>
<td>RVHPYQR</td>
<td>QKI-5</td>
<td>Wu et al., 1999</td>
</tr>
<tr>
<td>HARN</td>
<td>Eguchi et al., 1997</td>
<td>Bonifaci et al., 1997</td>
</tr>
<tr>
<td>YNNQSSNFGPMKGN</td>
<td>M9</td>
<td></td>
</tr>
<tr>
<td>SxGTKRSYxxM</td>
<td>InfluenzaNP</td>
<td>Wang et al., 1997</td>
</tr>
<tr>
<td>TKRSxxxxM</td>
<td>InfluenzaNP</td>
<td>Wang et al., 1997</td>
</tr>
<tr>
<td>VNEAFETLKRC</td>
<td>MyoD</td>
<td>Vandromme et al., 1995</td>
</tr>
<tr>
<td>MNKIPIKDLLNPG</td>
<td>Mat-alpha</td>
<td>Hall et al., 1984</td>
</tr>
</tbody>
</table>

In silico mutagenesis

M Cokol, R Nair & B Rost 2000 EMBO Rep 1:411-15
# Increasing accuracy and coverage

<table>
<thead>
<tr>
<th>Set</th>
<th>N NLS</th>
<th>Nprot nuc</th>
<th>Nfam nuc</th>
<th>Accuracy</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROSITE</td>
<td>1</td>
<td>96</td>
<td>31</td>
<td>90 %</td>
<td>3 %</td>
</tr>
<tr>
<td>SWISS-PROT</td>
<td>322</td>
<td>290</td>
<td>n.a.</td>
<td></td>
<td>9 %</td>
</tr>
<tr>
<td>NLS-lit cleaned</td>
<td>91</td>
<td>309</td>
<td>35</td>
<td>100 %</td>
<td>10 %</td>
</tr>
<tr>
<td>NLS-lit consensus</td>
<td>91</td>
<td>537</td>
<td>35</td>
<td>100 %</td>
<td>17 %</td>
</tr>
<tr>
<td>PredictNLS_DB</td>
<td>214</td>
<td>1354</td>
<td>186</td>
<td>100 %</td>
<td>43 %</td>
</tr>
</tbody>
</table>

M Cokol, R Nair & B Rost *EMBO Rep* 2000 1:411-15

© Burkhard Rost

ROSTLAB. TUM

54/140
Using NLS to bind DNA

M Cokol, R Nair & B Rost 2000 EMBO Rep 1:411-15
NLS driven import-bind-export cycle

**STEP 1:** import by binding NLS

**STEP 2:** NLS off DNA on

**STEP 3:** export by binding NLS
II.6 Predict localization: de novo prediction: “global signal”? 
1999 Nobel Prize in Medicine

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

© nobelprize.org

Guenter Blobel
Rockefeller Univ
NYC, USA
Signals NOT known for most proteins of known localization
Most proteins no known signal

☐ either:
  loads of signals remain to be discovered
☐ or:
  other types of motifs
Types of zip-codes

sequence

structure

signal peptide

target motif

Most proteins no known signal

☐ either:
   loads of signals remain to be discovered
☐ or:
   other types of motifs
☐ or:
   there are other mechanisms
Other import route WITH zip-codes
Find all “secondary cargo proteins”
Find all binders: Realistic endeavor?
Can we estimate how many secondary cargo there could be?
... while we consider, we still want to somehow predict for other proteins, how?
de novo prediction: simple
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 to surface composition!

Residue composition projected onto first two principle components

All residues

Core residues

Localization correlates 2 surface composition!

Residue composition projected onto first two principle components

Surface residues

All residues

Core residues

Surface “masked” by sugar

Surface composition projected onto first two principle components

## Surface composition

### Average surface composition

|       | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | V | Y |
| nuclear |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| cytoplasmic |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| extracellular |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

<table>
<thead>
<tr>
<th></th>
<th>15</th>
<th>10</th>
<th>5</th>
<th>15</th>
<th>10</th>
<th>5</th>
<th>15</th>
<th>10</th>
<th>5</th>
<th>15</th>
<th>10</th>
<th>5</th>
<th>15</th>
<th>10</th>
<th>5</th>
</tr>
</thead>
</table>

II.7 Predict localization: de novo prediction

Rostlab
Localization correlates 2 surface composition!
Features that discriminate localization

- amino acid
- surface
- sec str
- combined

R Nair & B Rost unpublished
INSERT: concept of neural networks & SVMs
Simple neural network

\[ \text{out0} = J_{11} \text{in1} + J_{12} \text{in2} \]

\[ \text{out} = \tanh(\text{out0}) \]
Training a neural network 1
Training a neural network 2

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

![Diagram of neural network training](image)

- **Round 1**:
  - Error vs. Junctions graph
  - Neuron connections:
    - Input 1: 1
    - Input 0: 1
    - Input 0: 1
    - Input 1: 1

- **Round 2**:
  - Continue training
  - Adjust weights based on error feedback
Training a neural network 3

Round 1

Round 2

Error

Junctions

© Burkhard Rost
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

Round 3
1 -1
0 1
0 1
1 2

Error

Junctions

in
-2 -1 1 2
out
-1 1

© Burkhard Rost
Neural networks classify points
Simple neural network with hidden layer

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

• **output:**

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

\( \text{in}_j \) value of input unit \( j \); \( \text{out}_i \) value of output unit \( i \);
\( J_{ij} \) connection between input unit \( j \) and output unit \( i \)

• **error:**

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

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

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

• **goal:**

○ representation of set of examples (training set) for which the mapping input-\( \rightarrow \)output is known, i.e., the secondary structure state of the central residue has been observed by the network
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
Neural Network for secondary structure
Effect of over-training: practice

- Red triangles: ratio for training set
- Pink circles: ratio for testing set

Number of correct classifications per example vs. number of cycles.
SVM: Support Vector Machines

VN Vapnik 1998 Statistical learning theory, Wiley
SVM: non-linear map to feature space

Original Problem  

\[ \bar{x} \]

Transformed Problem  

\[ \Phi(\bar{x}) \]
SVM: major difference to neural networks

© Burkhard Rost

© Rajesh Nair

VN Vapnik 1998 Statistical learning theory, Wiley
Idea of machine learning explained through everyday problem
by Theresa Wirth

Theresa Wirth
LMU & TUM, Munich
Machine-learning of best wake-up
Machine-learning of best wake-up

hours awake the last day
Machine-learning of best wake-up hours awake the last day
Additional evidence

hours of sleep until getting up

hours awake the last day

© Theresa Wirth
Binary classification?
Binary classification?
Binary classification?

Simplistic perspective of sub-cellular location
How to machine-learn 3 localization classes?
Three localization classes through machine learning

EXTRACELLULAR

NUCLEAR

CYTOPLASMIC

EXT

NUC

CYT
How to binary classify 3 states?
Three localization classes through machine learning

- Extracellular
- Nuclear
- Cytoplasmic

Simplistic perspective of sub-cellular location.
Advanced NN to predict localization

R Nair and B Rost (2003) *Proteins* 53: 917-30

© Burkhard Rost
Advanced NN to predict 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
How to connect binary classifiers?

EXT
other

NUC
other

CYT
other
How to connect binary classifiers?

1. EXT 0.3 → predict=other
2. NUC 0.6 → predict=NUC
3. CYT 0.7 → predict=CYT
Prediction scheme

Protein

Discrimination algorithm like ‘neural network’ or ‘support vector machine’
Problems with simple approach

Parallel architecture

Protein trafficking is hierarchical

CYTOSOL

<table>
<thead>
<tr>
<th>NUCLEUS</th>
<th>MITOCHONDRIA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENDOPOLASMIC RETICULUM</td>
<td></td>
</tr>
<tr>
<td>GOLGI</td>
<td></td>
</tr>
<tr>
<td>EXTRA-CELLULAR SPACE</td>
<td></td>
</tr>
</tbody>
</table>

Problems with simple approach

- Parallel architecture
  - Protein trafficking is hierarchical

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

Solution: hierarchical SVM

Protein

Secretory Pathway

Intra-cellular

Cytoplasm

EXTRA-CELLULAR ORGANELLE NUCLEUS CYTOSOL MITOCHONDRIA
Hierarchical prediction system

B Alberts et al. 1994 The Cell Garland

R Nair & B Rost 2005 JMB 348:85-100
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
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.
# PSORT II - k-nearest neighbor

<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: 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
A Hoglund, P Donnes, T Blum, HW Adolph and O Kohlbacher (2006) Bioinformatics 22: 1158-65: Fig. 2

Fig. 2. The architecture of the MultiLoc prediction system. A query sequence enters a first layer of prediction methods; SVMTarget, SVMSA, SVMaac, and MotifSearch. The characteristics of each protein is collected in the PPV, which is used to obtain probability estimates by a set of one-versus-one SVMs at the second layer.
CELLO: trick in the coding

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
Sorting into sub-nuclear compartments
Shuttle into the nucleus

Cytoplasm

Nucleus

Importin

Transportin

NLS

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>

AM Mehdi, MSB Sehgal, B Kobe, TL Bailey, M Boden (2011)

*Bioinformatics* 27: 1239-46: Table 2
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