Announcements

* Videos: SciVe / www.rostlab.org
* THANKS:
  
  Tim Karl + Manfred Roos

* Special lectures:

* NO lectures (not final):

* LAST lecture: Jul 5
* Examen: Jul 12, 10:30 (likely this room)
  
  • Makeup: likely: Oct 18 - afternoon

CONTACT: Marlena Drabik assistant@rostlab.org
Lecture plan (PP1: Structure)

01: 2012/04/17: skipped
02: 2012/04/19: welcome: who we are
03: 2012/04/24: intro I - acids/structure
04: 2012/04/26: INSERT - Machine learning in biology (cross-validation asf) - TMH1
05: 2012/05/01: holiday (May 1)
06: 2012/05/03: INSERT 2 - TMH2
07: 2012/05/08: Andrea Schafferhans: intro II - 3D comparisons: Andrea Schafferhans
08: 2012/05/10: Andrea Schafferhans: Comparative Modeling 1
09: 2012/05/15: no lecture: student assembly
10: 2012/05/17: holiday (Ascension Day)
11: 2012/05/22: Andrea Schafferhans: Comparative Modeling 2
12: 2012/05/24: Alignment 1
13: 2012/05/29: holiday (Pentecost/Whitsun)
14: 2012/05/31: Alignment 2
15: 2012/06/05: intro III - 3D->1D: sec str
16: 2012/06/07: holiday (Corpus Christi)
17: 2012/06/12: sec str pred 1 (white board)
18: 2012/06/14: no lecture
19: 2012/06/19: sec str pred 2
20: 2012/06/21: sec str pred 3
21: 2012/06/26: transmembrane helix prediction
22: 2012/06/28: transmembrane strand prediction, solvent accessibility
23: 2012/07/03: 3D prediction
24: 2012/07/05: summary: what we do in our group
25: 2012/07/10: no lecture
26: 2012/07/12: no lecture
27: 2012/07/17: no lecture
28: 2012/07/24: no lecture
29: 2012/07/26: no lecture
Today: Secondary structure prediction 1

☑️ LAST WEEKs
  • 3D->1D: secondary structure assignment

☑️ THIS WEEK
  • Secondary structure prediction methods - details

☑️ NEXT WEEK
  • Secondary structure prediction contd.
1D prediction
Notation: protein structure 1D, 2D, 3D
Words

- Secondary structure prediction
- 2ndary structure prediction
- 2D prediction
Protein function classification

Protein Space:

X = Positive  \quad Y = Negative

- Red circles: Close Homology (Sequence Id. > 60%, Psi-Blast Eval < 10^{-20})
- Blue circles: Distant Homology (Domain, Motif)
- Orange circles: Machine Learning (NN, SVM)

© Kaz Wrzeszczynski: Thesis
1D: secondary structure prediction
Coverage of structure space

The art of being humble

Comparative modelling

Exp. 3D

Percentage proteins

Percentage residues
Secondary structure prediction

- DSSP secondary assignment has 8 “states”

- H = Helix
- G = 3\textsubscript{10} helix
- I = Pi helix
- E = Extended (strand)
- B = beta-bridge, single strand residue
- T = Turn, i.e. one turn of helix
- S = bent
- “ “ = loop
Goal of secondary structure prediction

helix

AAQVKDALT
LEQWGTLLAQL
RAIWEQELTD
FPEFLTMMAR
QETWLGWLTL

strand

VNTFV
GTFAA
VAHAL
YIAQ
DIRVGL

loop

LEDKSPD
HNPTGID
HPKDDS
RNFTGRN
AKGKPMD

Secondary structure prediction methods

- L Pauling, RB Corey and HR Branson (1951) Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. PNAS 37:205-211.
Secondary structure prediction methods

- L Pauling, RB Corey and HR Branson (1951) Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. PNAS 37:205-211.
  some are more equal than others ...
Sec str pred methods: single residues

- Pauling, RB Corey and HR Branson (1951) Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. PNAS 37:205-211.
Albert Szent-Györgyi von Nagyrapolt
(Sep 16, 1893 - Oct 22, 1986)

1937: Nobel Prize in Physiology or Medicine
"for his discoveries in connection with the biological combustion processes, with special reference to vitamin C and the catalysis of fumaric acid"
Simple prediction: frequency

- First step (Szent-Györgyi)
  Proline breaks a helix
  Helices span several turns, i.e. >4 residues
  -> identify helices/non-helices

- from Proline to odds for all
Simple prediction: frequency

- from Proline to odds for all

QEKSPREVTM KKGDILTLLNSTNK

E...E       EEEEEEE

|   | A | A | D | D | E | E | G | G | I | I | K | K | L | L | M | M | N | N | P | P | Q | Q | R | R | S | S | T | T | V | V |
| E | 1 |   | 1 |   | 3 |   | 1 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| L | 1 | 1 | 1 |   | 4 |   |   |   | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 |   |   |   |   |   |   |   |   |   |   |   |   |

Tuesday June 19, 2012
Secondary structure prediction methods

- single residues (1. generation)
  - Chou-Fasman, GOR
Secondary structure prediction accuracy

- Q3 : three-state per-residue accuracy

\[
Q3 = \frac{\text{number of correctly predicted residues in states helix, strand, other}}{\text{number of residues in protein}}
\]

Secondary structure prediction methods

- single residues (1. generation)
  - Chou-Fasman, GOR
    - published: 63% accuracy

Secondary Structure Assignment: DSSP

- Dictionary of protein Secondary Structure for Proteins
- ASSESSING secondary structure prediction

Secondary structure prediction methods

- **single residues**
  - Chou-Fasman, GOR
    - 1957-70/80
    - 50-55% accuracy (assessed in 1994)

2nd Generation: what would you do?
Secondary Structure Assignment: DSSP

Dictionary of protein Secondary Structure for Proteins

Secondary structure prediction: 1.+2. Generation
Secondary structure prediction: 1.+2. Generation

- single residues (1. generation)
  - Chou-Fasman, GOR
  - 50-55% accuracy

- segments (2. generation)
  - GORIII

Secondary structure prediction: 1.+2. Generation
Secondary structure prediction: 1.+2. Generation

- **single residues** (1. generation)
  - Chou-Fasman, GOR
    50-55% accuracy (Q3)

- **segments** (2. generation)
  - GORIII
    55-60% Q3
Secondary structure prediction: 1.+2. Generation
Secondary structure prediction: 1.+2. Generation

- single residues (1. generation)
  - Chou-Fasman, GOR
    - 50-55% accuracy
  - GORIII
    - 55-60% accuracy

- segments (2. generation)
  - GORIII
    - 1986-92

- problems
  - < 100%
    - they said: 65% max
Helix formation is local

THYROID hormone receptor (2nll)

residues $i$ and $i+3$
β-sheet formation is NOT local

Erbutoxin β (3ebx)
Secondary structure prediction: 1.+2. Generation
Secondary structure prediction: 1.+2. Generation

- **single residues** (1. generation)
  - Chou-Fasman, GOR
    - 50-55% accuracy
  - GORIII
    - 55-60% accuracy

- **segments** (2. generation)
  - 1957-70/80
  - 1986-92

- **problems**
  - < 100% they said: 65% max
  - < 40% they said: strand non-local
Secondary structure prediction: 1.+2. Generation
Secondary structure prediction: 1.+2. Generation

- single residues  
  • Chou-Fasman, GOR  
    1957-70/80  
    50-55% accuracy

- segments  
  • GORIII  
    1986-92  
    55-60% accuracy

- problems  
  • < 100%  
    they said: 65% max  
  • < 40%  
    they said: strand non-local  
  • short segments
Problems of secondary structure predictions (before 1994)

SEQ    KELVLALYDYEKSPREVTKKGDILTLLNSTNKDWKVEVNDQGFVPAAYVKKLD
OBS    EEEE  E  E  EEEE  EEEE  EEEE  EEEEEEHHHEEEE
TYP    EHHHH  EE  EEEE  EE  HHHHEE  EEEHH

© Burkhard Rost (TU Munich)
INSERT: concept of neural networks
Simple neural network

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

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

Diagram of a neural network with input nodes labeled 0 and 1.
Training a neural network 2

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

Error

Junctions

1
0
0
1
1
1
1
Training a neural network

- Diagram showing the structure of a neural network with inputs and outputs.
- Input values: 1, 0, 0, 1, 1, 2.
- Output values: -1, -1, 1, 2.

Graph showing the input vs. output relationship with a sigmoid function.
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 networks: input -> output

**two steps:**

1. **linear:** sum over all input \( \times \) connection
2. **non-linear:** sigmoid trigger, i.e., project sum onto 0-1

\[
\text{input to unit} (=\text{sum}) \quad \rightarrow \quad \text{output from unit} \quad \rightarrow \quad \text{output = secondary structure state of central residue}
\]

- **step 1:** \( \sum \text{connection}_{ij} \times \text{input}_j \)
- **step 2:**

input = 3 adjacent residues in protein sequence

```
A
C
A
A
C
C
```

```
\alpha
L
```

```
\begin{align*}
\text{sum} & \quad \rightarrow \quad \text{output} < \text{decision line} \\
\text{result:} & \quad \leftarrow \quad \text{< decision line}
\end{align*}
```

© Burkhard Rost (TU Munich)
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->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
Effect of over-training: practice

- Number of correct classifications per example
- Number of cycles

- Ratio for training set
- Ratio for testing set

Effect of over-training: practice

© Burkhard Rost (TU Munich)
Secondary structure predictions of 1. and 2. generation

- single residues (1. generation)
  - Chou-Fasman, GOR 1957-70/80
    50-55% accuracy

- segments (2. generation)
  - GORIII 1986-92
    55-60% accuracy

problems

- < 100% they said: 65% max
- < 40% they said: strand non-local
- short segments
Neural Network for secondary structure
<table>
<thead>
<tr>
<th>method</th>
<th>overall accuracy</th>
<th>helix</th>
<th>strand</th>
<th>other</th>
</tr>
</thead>
<tbody>
<tr>
<td>unbalanced</td>
<td>62%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
RETURN: secondary structure prediction
Secondary structure predictions of 1. and 2. generation

single residues (1. generation)
- Chou-Fasman, GOR 1957-70/80
  50-55% accuracy

segments (2. generation)
- GORIII 1986-92
  55-60% accuracy

problems
- < 100% they said: 65% max
- < 40% they said: strand non-local
- short segments
Neural Network for secondary structure

| A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y |

- D (L)
- R (E)
- Q (E)
- G (E)
- F (E)
- V (E)
- P (E)
- A (H)
- A (H)
- Y (H)
- V (E)
- K (E)
- K (E)

H
E
L

© Burkhard Rost (TU Munich)
<table>
<thead>
<tr>
<th>method</th>
<th>overall accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>neural network</td>
<td>62%</td>
</tr>
</tbody>
</table>
NN predicts secondary structure

<table>
<thead>
<tr>
<th>method</th>
<th>overall accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>neural network</td>
<td>62%</td>
</tr>
</tbody>
</table>
### NN predicts secondary structure

<table>
<thead>
<tr>
<th>method</th>
<th>overall accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>neural network</td>
<td>62%</td>
</tr>
</tbody>
</table>

... and developer believes that application of machine learning is all the intelligence he will ever need...
NN predicts secondary structure

<table>
<thead>
<tr>
<th>method</th>
<th>overall accuracy</th>
<th>helix</th>
<th>strand</th>
<th>other</th>
</tr>
</thead>
<tbody>
<tr>
<td>neural network</td>
<td>62%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

full pie: all correctly predicted residues
Performance

NN sec str: training dynamics

\[ E^\mu = \sum_i (o_i^\mu - d_i^\mu)^2 \]

\[ \Delta J^\mu \propto - \frac{\partial E^\mu}{\partial J} \{J\} \]

time: 1 step = 20,000 training samples
**NN sec str: training dynamics**

![Graph showing training dynamics](image)

- **Performance** vs **time**: 1 step = 20,000 training samples

- **Equation for performance**: 
  \[ E^\mu = \sum_i (o_i^\mu - d_i^\mu)^2 \]

- **Equation for change in performance**: 
  \[ \Delta J^\mu \propto - \frac{\partial E^\mu}{\partial J} \]

© Burkhard Rost (TU Munich)

Tuesday June 19, 2012
### NN predicts secondary structure

<table>
<thead>
<tr>
<th>method</th>
<th>overall accuracy</th>
<th>helix</th>
<th>strand</th>
<th>other</th>
</tr>
</thead>
<tbody>
<tr>
<td>neural network</td>
<td>62%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comparison: data bank distribution

**full pie: all correctly predicted residues**
NN predicts secondary structure

<table>
<thead>
<tr>
<th>method</th>
<th>overall accuracy</th>
<th>helix</th>
<th>strand</th>
<th>other</th>
</tr>
</thead>
<tbody>
<tr>
<td>neural network</td>
<td>62%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comparison: data bank distribution

Comparison: 33:33:33

full pie: all correctly predicted residues
Balanced training

\[
E^\mu = \sum_i \left( o^\mu_i - d^\mu_i \right)^2
\]

\[
\Delta J^\mu \propto - \frac{\partial E^\mu \{J\}}{\partial J}
\]
Balanced training

normal training

\[ E^\mu = \sum_i \left( o_i^\mu - d_i^\mu \right)^2 \]

\[ \Delta J^\mu \propto - \frac{\partial E^\mu \{J\}}{\partial J} \]

balanced training

\[ E = \sum_{\mu=\alpha,\beta,L} \sum_i \left( o_i^\mu - d_i^\mu \right)^2 \]
Balanced training: dynamics

Other Strand Helix

train: unbalanced

$E^\mu = \sum_i (o_i^\mu - d_i^\mu)^2$

$\Delta J^\mu \propto - \frac{\partial E^\mu}{\partial J}$

balanced

$E^\mu = \sum_{\mu = \alpha, \beta, L} \sum_i (o_i^\mu - d_i^\mu)^2$
Balanced training: dynamics

![Graph showing dynamics for unbalanced and balanced training]

**train: unbalanced**

\[ E^\mu = \sum_i (o_i^\mu - d_i^\mu)^2 \]

\[ \Delta J^\mu \propto - \frac{\partial E^\mu}{\partial J} \]

**balanced**

\[ E^\mu = \sum_{\mu=\alpha,\beta,L} \sum_i (o_i^\mu - d_i^\mu)^2 \]
<table>
<thead>
<tr>
<th>method</th>
<th>overall accuracy</th>
<th>helix</th>
<th>strand</th>
<th>other</th>
</tr>
</thead>
<tbody>
<tr>
<td>unbalanced</td>
<td>62%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>balanced</td>
<td>60%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comparison: data bank distribution

Comparison: 33:33:33

Full pie: all correctly predicted residues
Secondary structure predictions of 1. and 2. generation

- **single residues** (1. generation)
  - Chou-Fasman, GOR
    - 1957-70/80
    - 50-55% accuracy

- **segments** (2. generation)
  - GORIII
    - 1986-92
    - 55-60% accuracy

- **problems**
  - < 100%
    - they said: 65% max
  - < 40%
    - they said: strand non-local
  - short segments
Neural networks DO improve if developer does something more than dream the machine learning dream...
<table>
<thead>
<tr>
<th>method</th>
<th>overall accuracy</th>
<th>helix</th>
<th>strand</th>
<th>other</th>
</tr>
</thead>
<tbody>
<tr>
<td>unbalanced</td>
<td>62%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>balanced</td>
<td>60%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

comparison: data bank distribution

comparison: 33:33:33

full pie: all correctly predicted residues
Secondary structure predictions of 1. and 2. generation

- Single residues (1. generation)
  - Chou-Fasman, GOR
    - 1957-70/80
    - 50-55% accuracy

- Segments (2. generation)
  - GORIII
    - 1986-92
    - 55-60% accuracy

- Problems
  - < 100%  
    - they said: 65% max
  - < 40%  
    - they said: strand non-local
  - Short segments
\( \beta \)-sheet formation is NOT local

Erabutoxin \( \beta \) (3ebx)
Conclusion: not all sound explanations are right!
Secondary structure predictions of 1. and 2. generation

- single residues (1. generation)
  - Chou-Fasman, GOR
    1957-70/80
    50-55% accuracy

- segments (2. generation)
  - GORIII
    1986-92
    55-60% accuracy

- problems
  - < 100%
    they said: 65% max
  - < 40%
    they said: strand non-local
  - short segments
Bad segment prediction

1st level

EEE
HHHH

comparison:
observed:

EEEEEEE
HHHHHHHHHHHH
PHDsec: structure-to-structure network


Tuesday June 19, 2012
Better segment prediction

<table>
<thead>
<tr>
<th>1st level</th>
<th>2nd level</th>
</tr>
</thead>
<tbody>
<tr>
<td>EEE</td>
<td>EEEEE</td>
</tr>
<tr>
<td>HHHH</td>
<td>HHHHHH</td>
</tr>
<tr>
<td>comparison:</td>
<td>observed:</td>
</tr>
<tr>
<td>EEEEEE</td>
<td>HHHHHHHH</td>
</tr>
</tbody>
</table>
Better prediction of segment lengths

![Graph showing comparison of DSSP and PHD predictions for segment lengths](image)

- Number of segments vs. Segment length
- Better prediction of segment lengths

© Burkhard Rost (TU Munich)
Better prediction of segment lengths

A

B

© Burkhard Rost (TU Munich)
Other ideas

- More output units, e.g. instead of central residue: take central 3
  1. 9 output units
  2. average output -> 3 units

- Output back into neural networks:
Other ideas

- **output back into neural networks:**

Gianluca Pollastri, Dariusz Przybylski, B Rost and Pierre Baldi (2002) Proteins 47:228-235: Fig. 1
STILL ONLY 60+\% accuracy.

How to improve beyond that?
How to get more data into it?
How to get more data into it?
Evolution has it!

Sequence identity implies structural similarity!

Don't know region

Percentage sequence identity

Number of residues aligned

C Sander & R Schneider 1991 Proteins 9:56-68
B Rost 1999 Prot Engin 12:85-94

© Burkhard Rost (TU Munich)
<table>
<thead>
<tr>
<th>1</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>fyn_human VTLFVALYDY EARTEDLSF HKGEKFQILN SSEGDKWEAR SLTTGGETYI</td>
<td></td>
</tr>
<tr>
<td>yrk_chick VTLFIALYDY EARTEDLSF QKGEKFHIIN NTEGDKWEAR SLSSGATFYI</td>
<td></td>
</tr>
<tr>
<td>fgr_human VTLFIALYDY EARTEDDLTF TKGEKFHILN NTEGDKWEAR SLSSGKTGCI</td>
<td></td>
</tr>
<tr>
<td>yes_chick VTFVFALYDY EARTADDLSF KKGERFQIIIN NTEGDKWEAR SIATGKTGYI</td>
<td></td>
</tr>
<tr>
<td>src_avis2 VTTFVALYDY ESRTEQDLSF KKGELIQIVN NTEGDKWALAH SLTTGQTYI</td>
<td></td>
</tr>
<tr>
<td>src_aviss VTTFVALYDY ESRTEQDLSF KKGELIQIVN NTEGDKWALAH SLTTGQTYI</td>
<td></td>
</tr>
<tr>
<td>src_avisr VTTFVALYDY ESRTEQDLSF KKGELIQIVN NTEGDKWALAH SLTTGQTYI</td>
<td></td>
</tr>
<tr>
<td>stk_hydat VTIFVALYDY EAISEDLSF KKGELIQIIN TADGDWYAR SLITNSEQIY</td>
<td></td>
</tr>
<tr>
<td>src_rsvpa ........ ESIREDLFS KKRERIQIVN NTEGDKWALAH SLTTGQTYI</td>
<td></td>
</tr>
<tr>
<td>hck_human .......... IVVALYDY EAIHHEILSF QKGDQMVELE ES.GEWKKAR SALTKEQIY</td>
<td></td>
</tr>
<tr>
<td>blk_mouse .......... FVVALFY AAVNDDLQV LKGEKIQLVR .STGDWYAR SLVTREGGYV</td>
<td></td>
</tr>
<tr>
<td>hck_mouse ........... TIVVALYDY EAIHREILSF QKGDQMVELE .EAGEWKKAR SALTKEQFYI</td>
<td></td>
</tr>
<tr>
<td>lyn_human ........... IVVALYDP DGIHDLDLSF KKGEMKIVLE .EHGKWKKAR SALTKEQFI</td>
<td></td>
</tr>
<tr>
<td>lck_human ........... LVIALHSY EPSHGDGLGF EKGEQLRIQE QS.GEWWAK QSLTQEGFI</td>
<td></td>
</tr>
<tr>
<td>ss81_yeast ........ ALYPY DADDDQEIS EQNEIQYS E.IGRWKKAR R.ANGERTGI</td>
<td></td>
</tr>
<tr>
<td>abl1_mouse ........... LFVALYDF VASGNTLSI TKGEKIRVVLG YnnGEWGEA Q.TKNGQGWV</td>
<td></td>
</tr>
<tr>
<td>abll_human ........... LFVALYDF VASGNTLSI TKGEKIRVVLG YnnGEWGEA Q.TKNGQGWV</td>
<td></td>
</tr>
<tr>
<td>src1_drome ........... VVSALYDY KSRDESLSF MKGDRMEVID DTESDWARVVL NLTTRQEGLI</td>
<td></td>
</tr>
<tr>
<td>mysd_dicdi ........... ALYDF DAESEMELSF KEGD1TVLQ QSSGDWADAE L.KGRRKVI</td>
<td></td>
</tr>
<tr>
<td>yfj4_yeast ........... WLYSF AGEESGDLPP RKGDVITILK ksQNDWTRGR V..NGREGIF</td>
<td></td>
</tr>
<tr>
<td>abl2_human ........... LFVALYDF VASGNTLSI TKGEKIRVVLG YNQNGESEV RSKNG.QGWV</td>
<td></td>
</tr>
<tr>
<td>tec_human ............ EIVYMAYQ DQAEGHDLRL ERGQYELILE KNDVHWWAR D.KYNGNEQYI</td>
<td></td>
</tr>
<tr>
<td>abll_caeel ........... LFVALYDF HGVGEEQLSL RKGQIVRILG YNKNEWCEA RlrLGEIQVW</td>
<td></td>
</tr>
<tr>
<td>txk_human ............ ALYDF LPREPCNLAL RRAEXELILE KYNPHWWKAR D.RLGNEQLI</td>
<td></td>
</tr>
<tr>
<td>yha2_yeast VRRVLYDL TTNEPDELFS RKGDVITVLE QVYRDWKGAL .RGNMGIY</td>
<td></td>
</tr>
<tr>
<td>abpl_sacex ........... AEYDY EAGEDNELTF AENDKINIE FVDDDM1GE LETTGQKLF</td>
<td></td>
</tr>
</tbody>
</table>
### Protein Alignments

<table>
<thead>
<tr>
<th>Protein</th>
<th>Alignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>GGGGG</td>
</tr>
<tr>
<td>Y</td>
<td>YYYYY</td>
</tr>
<tr>
<td>I</td>
<td>IEEIE</td>
</tr>
<tr>
<td>Y</td>
<td>YYYYY</td>
</tr>
<tr>
<td>D</td>
<td>DDDDD</td>
</tr>
<tr>
<td>P</td>
<td>PPPPP</td>
</tr>
<tr>
<td>E</td>
<td>AEAEE</td>
</tr>
<tr>
<td>D</td>
<td>VVEEG</td>
</tr>
<tr>
<td>G</td>
<td>GGGGG</td>
</tr>
<tr>
<td>D</td>
<td>DDDDD</td>
</tr>
<tr>
<td>P</td>
<td>PPPPP</td>
</tr>
<tr>
<td>D</td>
<td>DTTDD</td>
</tr>
<tr>
<td>D</td>
<td>NQNNN</td>
</tr>
<tr>
<td>G</td>
<td>GNGGG</td>
</tr>
<tr>
<td>V</td>
<td>VVVV</td>
</tr>
<tr>
<td>N</td>
<td>EPKKK</td>
</tr>
<tr>
<td>P</td>
<td>PPPPP</td>
</tr>
<tr>
<td>G</td>
<td>GGGGG</td>
</tr>
<tr>
<td>T</td>
<td>TTTTT</td>
</tr>
<tr>
<td>D</td>
<td>EKSSA</td>
</tr>
<tr>
<td>F</td>
<td>FFFFA</td>
</tr>
<tr>
<td>:</td>
<td>:::::</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>profile table</th>
<th>GSAPD NTEKQ CVHIR LMYFW</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>5</td>
</tr>
<tr>
<td>Y</td>
<td>.</td>
</tr>
<tr>
<td>I</td>
<td>.</td>
</tr>
<tr>
<td>Y</td>
<td>.</td>
</tr>
<tr>
<td>D</td>
<td>5</td>
</tr>
<tr>
<td>P</td>
<td>5</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>5</td>
</tr>
<tr>
<td>D</td>
<td>5</td>
</tr>
<tr>
<td>P</td>
<td>5</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
</tr>
<tr>
<td>N</td>
<td>4</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
</tr>
<tr>
<td>V</td>
<td>4</td>
</tr>
<tr>
<td>N</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
</tr>
<tr>
<td>P</td>
<td>5</td>
</tr>
<tr>
<td>G</td>
<td>5</td>
</tr>
<tr>
<td>T</td>
<td>5</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>5</td>
</tr>
</tbody>
</table>

 corresponds to the the 21*3 bits coding for the profile of one residue

---

**B Rost & C Sander (1993) PNAS** 90:7558-62

**B Rost (1996) Methods Enzymol** 266:525-39

© Burkhard Rost (TU Munich)
From sequence to profile

1. From sequence data bank.
2. Filter MaxHom.
3. Extract alignment.

Sequence identity:
- 100%
- 25%
- 80%
- Number of residues aligned.

Proteins:
- Protein A
- Protein B
- Protein C
- Protein M
- Protein N
PHDsec: more details

PHDsec

Local alignment

<table>
<thead>
<tr>
<th>13 adjacent residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
</tr>
<tr>
<td>AA.</td>
</tr>
<tr>
<td>LLL</td>
</tr>
<tr>
<td>LIT</td>
</tr>
<tr>
<td>AAG</td>
</tr>
<tr>
<td>CCS</td>
</tr>
<tr>
<td>GWV</td>
</tr>
<tr>
<td>: : :</td>
</tr>
</tbody>
</table>

Global statistics

<table>
<thead>
<tr>
<th>whole protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>ΔN-terminus</td>
</tr>
<tr>
<td>ΔC-terminus</td>
</tr>
</tbody>
</table>

Input local in sequence

A C L I G S V ins del cons
100 0 0 0 0 0 0 0 0 0 0 1.17
100 0 0 0 0 0 0 0 33 0 0.42
0 0 100 0 0 0 0 0 33 0 0.42
0 0 33 66 0 0 0 0 0 0 0.74
66 0 0 0 33 0 0 0 0 0 0.74
0 0 66 0 0 0 33 0 0 0 0.74
0 0 0 33 0 0 66 0 0 0 0.48

Input global in sequence

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

First level sequence-to-structure

Second level structure-to-structure


© Burkhard Rost (TU Munich)
Jury

single network vs. jury decision

centre of mass = jury over 1-4

decision

architecture 1

architecture 2

architecture 3

architecture 4
PROFsec: Evolutionary information + more

B Rost (2001) J Struct Biol 134, 204-18
PROFsec: Evolutionary information + more

B Rost (2001) J Struct Biol 134, 204-18
# Spectrin homology domain (SH3)

<table>
<thead>
<tr>
<th>HEADER</th>
<th>CYTOSKELETON</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPND</td>
<td>ALPHA SPECTRIN (SH3 DOMAIN)</td>
</tr>
<tr>
<td>SOURCE</td>
<td>CHICKEN (GALLUS GALLUS) BRAIN</td>
</tr>
<tr>
<td>AUTHOR</td>
<td>M.NOBLE,R.FAUPITI,A.MUSACCHIO,M.SARASTE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>......1......</th>
<th>......2......</th>
<th>......3......</th>
<th>......4......</th>
<th>......5......</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AA</strong></td>
<td>KELVLALDYEQKSPREVTMKKGDI LTLLNSTNKDWWKVEVNDRQGFVPAAYVKKLD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>OBS</strong></td>
<td>EEEE</td>
<td>E--E</td>
<td>EEEEEEE</td>
<td>EEEEEEE</td>
<td>EEEEEEEHHHHEEEE</td>
</tr>
<tr>
<td><strong>C+F</strong></td>
<td>HHHHHHHH</td>
<td>HHHHHH</td>
<td>EEEEEEE</td>
<td>HHHHHH</td>
<td>EEEEEEEHHHHHHH</td>
</tr>
<tr>
<td><strong>GOR</strong></td>
<td>HHHHHHHHH</td>
<td>HHHH</td>
<td>EEEEEEE</td>
<td>EEEHHH</td>
<td>HHH HHHHHHH</td>
</tr>
<tr>
<td><strong>PHD</strong></td>
<td>EEEEEEE</td>
<td>EEE</td>
<td>EEEEEEEEEEE</td>
<td>HHHHHH</td>
<td>EEEE HHEEEE</td>
</tr>
<tr>
<td>Rel</td>
<td>948999972587775211443884899847697314344045955111321221558</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>* **********</td>
<td>* **********</td>
<td>** **** *****</td>
<td>****</td>
<td>***</td>
</tr>
</tbody>
</table>
Prediction accuracy varies!

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

Number of protein chains

Per-residue accuracy (\(Q_3\))
Stronger predictions more accurate!

Q₃ per protein

H = 0.5
E = 0.4
L = 0.1

fit: Q₃ fit = 21 + 8.7 * Q₃

Reliability index averaged over protein

Per-residue accuracy (Q₃)

Number of protein chains

© Burkhard Rost (TU Munich)
Correct prediction of correctly predicted residues

![Graph showing the correct prediction of residues](image)

- **PHDse**
- **PHDac**
- **PHDhtm**

Overall per-residue accuracy vs. percentage of residues predicted for different models and RI values.
BAD errors are frequent!

\[<\text{BAD}> = 4.0\%; \text{sigma} = 5.9\%\]
False prediction for engineered proteins!

GB1: IgG-binding domain of protein G (CHAMELEON)

TTYKLILNGKTLKGETTTEAVDAATAEKVFQYANDNGVDGEWTYDDATKFTVTKE

| AA | TTYKLILNGKTLKGETTTEAVDAATAEKVFQYANDNGVDGEWTYDDATKFTVTKE |
| DSSP | EEEEEEEE EEEEEEEE HHHHHHHHHHHHHHHHH EEEEEEEE EEEEEEEE |
| PHD 30 | EEEEEEE EEE EHHHHHHHHH HHHHHHHH EEEE EEEE |
| PHD no | EEEEEEE EEE EHHHHHHHHH HHHHHHHH EEEE EEEE |

| PHD 30 | EEEEEEE EEE EHHHHHHHHH HHHHHHHH EEEE EEEE |
| PHD no | EEEEEEE EEE EHHHHHHHHH HHHHHHHH EEEE EEEE |

| PHD 30 | EEEEEEE EEE EHHHHHHHHH HHHHHHHH EEEE EEEE |
| PHD no | EEEEEEE EEE EHHHHHHHHH HHHHHHHH EEEE EEEE |

© Burkhard Rost (TU Munich)

Tuesday June 19, 2012
Rostlab & friends @ ISMB/ECCB Vienna

www.rostlab.org