title: Secondary structure prediction 2

short title: cb1_sec2

lecture: Protein Prediction 1 - Protein structure
Computational Biology 1
TUM summer 2014
Announcements

☐ Videos:   YouTube / www.rostlab.org

THANKS :

Tim Karl + Jonas Reeb

☐ Special lectures:
   • Apr 15 - Andrea Schafferhans

☐ No lecture:
   • May 29 Thu Ascension day
   • Jun 03 Tue no room
   • Jun 10 Tue Whitsun holidays
   • Jun 19 Thu Corpus Christi

☐ LAST lecture:    July 1

☐ Examen:    July 8
   • Makeup:    Oct 21 - morning
Announcing two courses winter 2014/15

2x15 slots for 2 courses during semester break (electable as modules IN0014 and IN2107)

1. Java Script Technology
2. Patterns and anti-patterns: examples for better coding

Supervisor: Guy Yachdav <yachdav@rostlab.org>

The pre-meeting will be in June, the talks will be held in a block at the end of September. Please register via email to the supervisor. TUMonline and a seminar website will be online soon.
Recap: secondary structure prediction
Goal of structure prediction

Epstein & Anfinsen, 1961: sequence uniquely determines structure

• INPUT: sequence

• OUTPUT:

3D structure and function
Zones


Midnight Zone ➔ Twilight Zone ➔ Daylight Zone

sequence - profile

profile - profile

sequence - sequence

structure similar
Comparative modeling applicable to about 1/3 of all proteins
Notation: protein structure 1D, 2D, 3D

| P | PP | 128 | 110 |
| Q | QQQY | 175 | 97 |
| I | FPQVI | 70 | E | 60 |
| T | SSIVR | 77 | E | 69 |
| L | LLSTL | 120 | E | 14 |
| W | WWQED | 238 | E | 81 |
| Q | RKQAK | 169 | E | 97 |
| R | RRRPQ | 200 | 62 |
| P | PPPPP | 24 | 48 |
| L | VVTKF | E | 71 | E | 59 |
| V | VVLII | E | 14 | E | 0 |
| T | TTKEK | E | 74 | E | 69 |
| I | AALIV | E | 0 | E | 0 |
| K | HYKKF | E | 90 | E | 73 |
| I | IILVI | E | 4 | E | 0 |
| G | EENGG | 46 | 41 |
| G | GGGTG | 62 | 53 |
| Q | QKAR | 68 | 71 |
| L | PFLMW | E | 118 | E | 59 |
| K | VVFKV | E | 31 | E | 73 |
| E | EESKK | E | 124 | E | 95 |
| A | VVGLG | E | 1 | E | 0 |
| L | LLILL | E | 29 | E | 0 |
| L | LLLVV | E | 24 | E | 0 |
| D | DDDDD | 49 | E | 58 |
| T | TTHTT | 72 | 51 |
| G | GGGGG | 62 | 30 |
| A | AAAA | 17 | 0 |
| D | DDDDD | 102 | 79 |
| D | DDANE | 69 | 58 |
| T | SSTVV | 1 | 69 |
| V | IVIV | E | 14 | E | 0 |
| L | VVIVL | E | 0 | E | 0 |
Pauling’s H-bond pattern used in DSSP

DSSP \( E_{HB} < -0.5 \text{ kcal/mol} \)

L Pauling & RB Corey (1953) PNAS 39:247-252
L Pauling, RB Corey & HR Branson (1951) PNAS 37:205-234

Monday May 26, 2014
Sec struc pred: 1st gen

1st generation (1957-1978):
e.g. Chou-Fasman / GOR
single residue odds

\[ p(SEC|AA_i) = \]

probability for observing secondary structure state SEC for amino acid AA at position

\[ = p(SEC|AA_j) - \forall \ i \land j \]
1st generation (1957-1978): single residue odds
  e.g. Chou-Fasman/GOR
  \[ p_1(\text{SEC}_i|\text{AA}_i) = \]
  probability for observing secondary structure state SEC for amino acid AA at position i

  e.g. GORIII
  odds for windows
  \[ p(\text{SEC}|\text{AA}_i) \neq p(\text{SEC}|\text{AA}_i) \]
Secondary structure prediction: 1.+2. Generation

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

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

- **problems**
  - < 100%
    - they said: 65% max
  - < 40%
    - they said: strand non-local
  - short segments
Simple neural network with hidden layer

\[
\text{out}_i = f \left( \sum_j J^2_{ij} \cdot f \left( \sum_k J^1_{jk} \cdot \text{in}_k \right) \right)
\]
Effect of over-training: practice

- ratio for training set
- ratio for testing set

number of correct classifications per example vs. number of cycles
1D: secondary structure prediction
RETURN: secondary structure prediction
Secondary structure predictions of 1. and 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
  - < 40% they said: strand non-local
  - short segments
Neural Network for secondary structure
## 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>
<tr>
<td>method</td>
<td>overall accuracy</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------</td>
</tr>
<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 sec str: training dynamics

Other \quad Strand \quad Helix

Performance

\[ E^{\mu} = \sum_{i} \left( o_{i}^{\mu} - d_{i}^{\mu} \right)^{2} \]

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

time: 1 step = 20,000 training samples

Monday May 26, 2014
NN sec str: training dynamics

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

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

Performance

1  2  3  4  5  6  7  8  9  10

0  0.2  0.4  0.6  0.8  1

time: 1 step = 20,000 training samples

© Burkhard Rost (TUM Munich)
### 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*
### 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

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

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

**normal 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**

\[ E = \sum_{\mu=\alpha,\beta,L} \sum_i \left( o^\mu_i - d^\mu_i \right)^2 \]
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

Unbalanced training:

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

Balanced training:

\[ E^\mu = \sum_{\mu=1}^{\alpha, \beta, L} \sum_i (o_i^\mu - d_i^\mu)^2 \]

\[ \Delta J^\mu \propto - \frac{\partial E^\mu}{\partial J} \]
<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
- 33:33:33

**Full pie: all correctly predicted residues**
Neural networks DO improve if developer does something more than dream the machine learning dream...
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

<table>
<thead>
<tr>
<th>SEQ</th>
<th>KELVLALYDQEKSQREVVTMKGDILTLNNSTNKNKDVWNKVEVNDQGFVPAAAYVKKLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>OBS</td>
<td>EEEE E E E EEEEEE EEEEEE EEEEEEHHHEEEE</td>
</tr>
<tr>
<td>TYP</td>
<td>EHHHH EE EEEE EE HHHEE EEEHH</td>
</tr>
</tbody>
</table>

1st level

<table>
<thead>
<tr>
<th>OBS</th>
<th>EEE EEEEHHHHHHHHHH</th>
</tr>
</thead>
</table>

comparison: observed:
Select samples at random

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

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
Local correlations in reality

residue $i$ and $i+3$
How to get those into the prediction?
PHDsec: structure-to-structure network

Better segment prediction

<table>
<thead>
<tr>
<th>1st level</th>
<th>2nd level</th>
<th>comparison: observed:</th>
</tr>
</thead>
<tbody>
<tr>
<td>EEE</td>
<td>EEEE</td>
<td>EEEEEE</td>
</tr>
<tr>
<td>HHHHH</td>
<td>HHHHHHHHH</td>
<td>HHHHHHHHH</td>
</tr>
</tbody>
</table>

Monday May 26, 2014
Better prediction of segment lengths
Better prediction of segment lengths

A

<table>
<thead>
<tr>
<th>Segment length</th>
<th>DSSP</th>
<th>PHD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

B

<table>
<thead>
<tr>
<th>Segment length</th>
<th>helix</th>
<th>strand</th>
<th>loop</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

Difference in number of observed - predicted segments

© Burkhard Rost (TUM Munich)
Structure-to-structure network: Invented?

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

Number of residues aligned

Percentage sequence identity

C Sander & R Schneider 1991 Proteins 9:56-68
B Rost 1999 Prot Engin 12:85-94
SH3
Src-homology 3 domain
one domain of proteins such as Src tyrosine kinase (STK)
SH3
Src-homology 3 domain
one domain of proteins such as Src tyrosine kinase (STK)
Evolution improves prediction

Evolutionary profile implicitly captures history of and individual protein!
Evolution improves prediction

Evolutionary profile implicitly captures history of and individual protein!
Evolution improves prediction

Evolutionary profile implicitly captures history of and individual protein!
**PHD: Neural network & evolutionary information**

<table>
<thead>
<tr>
<th>Protein</th>
<th>Alignments</th>
<th>Profile Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>G G G G</td>
<td>GSAPD NTEKQ CVHIR LMYFW</td>
</tr>
<tr>
<td>Y</td>
<td>Y Y Y Y</td>
<td>5 . . . . . . . . . . 5</td>
</tr>
<tr>
<td>I</td>
<td>I I E E</td>
<td>. . 2 . 3 . . . . . . 5</td>
</tr>
<tr>
<td>Y</td>
<td>Y Y Y Y</td>
<td>. . . . . . . . . . . 5</td>
</tr>
<tr>
<td>D</td>
<td>D D D D</td>
<td>. . 5 . . . . . . . .</td>
</tr>
<tr>
<td>P</td>
<td>P P P P</td>
<td>. . 5 . . . . . . . .</td>
</tr>
<tr>
<td>E</td>
<td>A E A A</td>
<td>. . 3 . 2 . . . . . .</td>
</tr>
<tr>
<td>D</td>
<td>V V E E</td>
<td>. . 1 . 2 . . . . . .</td>
</tr>
<tr>
<td>G</td>
<td>G G G G</td>
<td>5 . . . . . . . . . . 5</td>
</tr>
<tr>
<td>D</td>
<td>D D D D</td>
<td>5 . . . . . . . . . . 5</td>
</tr>
<tr>
<td>P</td>
<td>P P P P</td>
<td>5 . . . . . . . . . . 5</td>
</tr>
<tr>
<td>D</td>
<td>D T D D</td>
<td>4 . 1 . . . . . . . .</td>
</tr>
<tr>
<td>D</td>
<td>N Q N N</td>
<td>4 . 1 . . . . . . . .</td>
</tr>
<tr>
<td>G</td>
<td>G N G G</td>
<td>. . 1 . . . . . . . .</td>
</tr>
<tr>
<td>V</td>
<td>V V V V</td>
<td>. . 4 . 1 . . . . . .</td>
</tr>
<tr>
<td>N</td>
<td>E P K K</td>
<td>1 . 1 . 1 . . . . . .</td>
</tr>
<tr>
<td>P</td>
<td>P P P P</td>
<td>5 . . . . . . . . . . 5</td>
</tr>
<tr>
<td>G</td>
<td>G G G G</td>
<td>5 . . . . . . . . . . 5</td>
</tr>
<tr>
<td>T</td>
<td>T T T T</td>
<td>. . 5 . . . . . . . .</td>
</tr>
<tr>
<td>D</td>
<td>E K S A</td>
<td>1 . 1 . 1 . . . . . .</td>
</tr>
<tr>
<td>F</td>
<td>F F F F</td>
<td>1 . 1 . 1 . . . . . .</td>
</tr>
</tbody>
</table>

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

---

From sequence to profile

1. Sequence data bank
2. Sequence identity
   - 100%
   - 25%
   - 25%
3. Extract alignment
   - PHD

Filter MaxHom

Protein A
- Protein C
- Protein M

MaxHom

Protein A
- Protein B
- Protein N


© Burkhard Rost (TUM Munich)
PHDsec: more details

**PHDsec**

### input local in sequence

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Values</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A C L I G S V ins del cons</td>
<td>100 0 0 0 0 0 0 0 0 1.17</td>
<td></td>
</tr>
<tr>
<td>100 0 0 0 0 0 0 33 0 0.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 0 100 0 0 0 0 0 33 0.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 0 33 66 0 0 0 0 0 0.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>66 0 0 0 33 0 0 0 0 1.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 66 0 0 0 33 0 0 0 0.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 0 0 33 0 66 0 0 0 0.48</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### input global in sequence

- Percentage of each amino acid in protein
- Length of protein
- Distance: centre, N-term
- Distance: centre, C-term

**PHDsec**


Monday May 26, 2014
<table>
<thead>
<tr>
<th>single network vs. jury decision</th>
<th>architecture 1 architecture 2 centre of mass = jury over 1-4 architecture 3 architecture 4</th>
</tr>
</thead>
</table>

Monday May 26, 2014
PROFsec: Evolutionary information + more

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

<table>
<thead>
<tr>
<th>AA</th>
<th>KELVLALYDYQEKSREPVTMKGDILTLLNSTNKDWKVEVNRQGFPVAAAYVKKLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>OBS</td>
<td>EEEE E--E EEEEEEE EEEEE EE------HHH------EEE</td>
</tr>
<tr>
<td>C+F</td>
<td>HHHHHHHH HHHHHH EEEEEEE HHHHHH EEEEEEHHHNNNNHHH</td>
</tr>
<tr>
<td>GOR</td>
<td>HHHHHHHHH HHHHH EEEEEEE EEEEHH HHH HHHHHHHH</td>
</tr>
<tr>
<td>PHD</td>
<td>EEEEEEE  EEE EEEEEEEE HHHHHH EEEE HHHHHH</td>
</tr>
<tr>
<td>Rel</td>
<td>948999972587775211443884899847697314344045955111321221558</td>
</tr>
</tbody>
</table>

*:**6, **6, **7%
Prediction accuracy varies!

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

Per-residue accuracy (\( Q_3 \))

Number of protein chains
Stronger predictions more accurate!

H = 0.5  H = 0.8
E = 0.4  E = 0.1
L = 0.1  L = 0.1

Per-residue accuracy ($Q_3$)

$\langle Q_3 \rangle = 72.3\% \; ; \; \sigma = 10.5\%$

Number of protein chains

Reliability index averaged over protein

$Q_3$ per protein

fit: $Q_3$ fit = 21 + 8.7 * $Q_3$
Correct prediction of correctly predicted residues

- PHDse
- PHDac
- PHDhtm

Overall per-residue accuracy vs. percentage of residues predicted.

RI=9
RI=4
RI=7
RI=0

Monday May 26, 2014
BAD errors are frequent!

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

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

<table>
<thead>
<tr>
<th>AA</th>
<th>TTYKLILNGKTLKGETTTEAVDAATAEKVFQYANDNGVDGEWTYDDATKTFVTVEK</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSSP</td>
<td>EEEEEEEE EEEEEEEE HHHHHHHHHHHHHHHHHH EEEEEEEE EEEEEEEE</td>
</tr>
</tbody>
</table>

| PHD 30 | EEEEEEE E EHHHHHHHHHHHHHHHHH EEEE EEEEE EEEEE |
| PHD no  | EEEEEEE EEEEEHHHHHHHHHHHHHHHHHHHHHHH EEEE EEEEE EEEEE |

AATAEKVFQY
AWTVEKAFKTF

| PHD 30 | EEEEEEE EEEEEHHHHHHHHHHHHHHHHH EEEE EEEEE EEEEE |
| PHD no  | EEEEEEE EEEEEHHHHHHHHHHHHHHHHHHHHHHHH EEEE EEEEE EEEEE |

EHYDDATKTF
AWTVEKAFKTF

| PHD 30 | EEEEEEE EEE EHHHHHHHHHHHHHHHHH EEEE EEEEE EEEEE |
| PHD no  | EEEEEEE E E EHHHHHHHHHHHHHHHHHHHHHHHHH HHHHHHH EEEE |
Announcing two courses winter 2014/15

2x15 slots for 2 courses during semester break (electable as modules IN0014 and IN2107)

1. Java Script Technology
2. Patterns and anti-patterns: examples for better coding

Supervisor: Guy Yachdav <yachdav@rostlab.org>

The pre-meeting will be in June, the talks will be held in a block at the end of September. Please register via email to the supervisor. TUMonline and a seminar website will be online soon.
Proper comparison of methods
Method A = 60\% 
Method B = 63\%
B better?
Method A=60% B=63%, B better?

☐ same measure?

e.g. both Q3?
Method A=60% B=63%, B better?

- use same (meaningful) measure e.g. both Q3
- same data set
Method A=60% B=63%, B better?

- Use same (meaningful) measure e.g. both Q3
- Same data set:
  - Note both used 100 proteins, and both used random splits to take one half for testing, ok?
Method A=60% B=63%, B better?

- Use same (meaningful) measure e.g. both Q3
- Same data set: must contain ALL available proteins!
Method A=60% B=63%, B better?

- Use same (meaningful) measure e.g. both Q3
- Same data set: must contain ALL available proteins!
- Split training/testing: random ok?
Method A=60% B=63%, B better?

- Use the same (meaningful) measure, e.g., both Q3.
- Use the same data set: must contain ALL available proteins!
- Split training/testing: must ascertain that there was NO overlap between sets.
  Overlap defined as, e.g., comparative modeling cannot be applied.

C Sander & R Schneider 1991 *Proteins* 9:56-69
B Rost 1999 *Prot Engin* 12, 85-94
Method A=60% B=63%, B better?

- use same (meaningful) measure e.g. both Q3
- same data set: must contain ALL available proteins!
- split training/testing: must ascertain that there was NO overlap between sets.
- $63-60=3$
  - significant?
Method A=60% B=63%, B better?

- Use same (meaningful) measure e.g. both Q3
- Same data set: must contain ALL available proteins!
- Split training/testing: must ascertain that there was NO overlap between sets.
- 63-60=3, whether significant or not depends on distribution and number:
DeltaQ3=3%, 100 proteins->significant?

<Q3>=72.3% ; sigma=10.5%

Per-residue accuracy ($Q_3$)

Number of protein chains

Monday May 26, 2014
DeltaQ3=3% for 100 proteins is significant!

rule-of-thumb:

\[
\text{Stderror} = \frac{\sigma}{\sqrt{\text{proteins}}}
\]

here:

\[
\text{StdErr} = \frac{10.5}{\sqrt{100}} = \pm 1.05
\]

> DeltaQ3=3

-> statistically significant
Method B 20 years older than A, still better?
Difference
statistically significant
→ age no difference!
Any other test to do? (mind you B is 20 years old)
pre-release test: ideally use data added after both methods had been developed
Cross-validation: how to

Table 1

<table>
<thead>
<tr>
<th>Nhidden</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>62</td>
</tr>
<tr>
<td>30</td>
<td>64</td>
</tr>
<tr>
<td>45</td>
<td>63</td>
</tr>
</tbody>
</table>

Conclusion:
Q3=64%
best method has 30 hidden units
Cross-validation: how to

Table 1

<table>
<thead>
<tr>
<th>Nhiddenn</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>62</td>
</tr>
<tr>
<td>30</td>
<td>64</td>
</tr>
<tr>
<td>45</td>
<td>63</td>
</tr>
</tbody>
</table>

Conclusion: Q3=64%
best method has 30 hidden units

OK?
Cross-validation: need 3 sets!

<table>
<thead>
<tr>
<th>Nhid</th>
<th>cross-train</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>62</td>
<td>60</td>
</tr>
<tr>
<td>30</td>
<td>64</td>
<td>61</td>
</tr>
<tr>
<td>45</td>
<td>63</td>
<td>62</td>
</tr>
</tbody>
</table>

**Conclusion:**
Q3=61%
best method has 30 hidden units
Announcing two courses winter 2014/15

2x15 slots for 2 courses during semester break (electable as modules IN0014 and IN2107)

1. Java Script Technology
2. Patterns and anti-patterns: examples for better coding

Supervisor: Guy Yachdav <yachdav@rostlab.org>

The pre-meeting will be in June, the talks will be held in a block at the end of September. Please register via email to the supervisor. TUMonline and a seminar website will be online soon.
Lecture plan (CB1: Structure)-generic

01: 2014/04/08 Tue: sorry
02: 2014/04/10 Thu: welcome: who we are
03: 2014/04/15 Tue: Intro I - acids/structure (Andrea Schafferhans)
04: 2014/04/17 Thu: SKIP: Easter vacation
05: 2014/04/22 Tue: SKIP: Easter vacation
06: 2014/04/24 Thu: Intro 2 - domains
07: 2014/04/29 Tue: Intro 3 - 3D comparisons
08: 2014/05/01 Thu: SKIP: “May day” - (NOT to be confused with “m’aidez”)
09: 2014/05/06 Tue: SKIP: student assembly (SVV)
10: 2014/05/08 Thu: Alignment 1
11: 2014/05/13 Tue: Alignment 2
12: 2014/05/15 Thu: Comparative modeling 1
13: 2014/05/20 Tue: Secondary structure prediction 1
14: 2014/05/22 Thu: Secondary structure prediction 2
15: 2014/05/27 Tue: 1D: Secondary structure prediction 1
16: 2014/05/29 Thu: SKIP: holiday (Ascension Day)
17: 2014/06/03 Tue: SKIP: no room
18: 2014/06/05 Thu: 1D: Secondary structure prediction 2
19: 2014/06/10 Tue: SKIP: Whitsun holidays
20: 2014/06/12 Thu: 1D: Transmembrane helix prediction
21: 2014/06/17 Tue: Nobel prize symposium
22: 2014/06/19 Thu: SKIP: Corpus Christi (Fronleichnam)
23: 2014/06/24 Tue: 1D: Transmembrane strand prediction, solvent accessibility
24: 2014/06/26 Thu: 2D prediction
25: 2014/07/01 Tue: 3D prediction/wrap up
26: 2014/07/03 Thu: wrap up again
27: 2014/07/08 Tue: examen, no lecture
28: 2014/07/10 Thu: no lecture