title: Sequence alignments - multiple

short title: pp1_alignment2

lecture: Computational Biology 1 - Protein structure (for Informatics) - TUM summer semester
Videos: YouTube / www.rostlab.org

THANKS:

Special lectures:
- TBN

No lecture:
- 05/21 SVV (Student Representation)
- 05/30 Ascension Day
- 06/08-11 Whitsun holiday
- 06/11 TBA: likely no lecture
- 06/13 TBA: likely no lecture
- 06/20 Corpus Christi

LAST lecture: bef: Jul 18 after: Jul 9

Examen: JULY 17 (!!) 18:00-20:00 TBA
- Makeup: NONE!

EXERCISES:

CONTACT: pp1ex@rostlab.org
Announcements

CONTACT: pp1ex@rostlab.org

OFFICIAL course web site rostlab.org

Maria Schelling

Sofie Kemper
Today: Alignments 1

- LAST
  - 3D comparison / pairwise alignments

- TODAY
  - Multiple alignments and “reach of comparative modeling”

- NEXT
  - Alignment contd / Comparative modelling
Science is communication

questions are often the first step
Notation: protein structure 1D, 2D, 3D

P PP P 128 110
Q QQQY 175 97
I FFQVI 70 E 60
T SSIVR 77 E 69
L LLSTL 120 E 14
W WWQED 238 E 81
Q RKQAK 169 E 97
R RRRPF 200 62
P PPPPP 24 48
L VVTKF E 71 E 59
V VVIIE E 14 E 0
T TTTEK E 74 E 69
I AALIV E 0 E 0
K HYKKF E 90 E 73
I IIVVI 4 E 0
G EENGG 46 41
G GGGTG 62 53
Q QQKRR 68 71
L PPLMW E 118 E 59
K VVFKV E 31 E 73
E EESKK E 124 E 95
A VVGLG E 1 E 0
L LLLL E 29 E 0
L LLLL V 24 E 0
D DDDDD 49 E 58
T TTTE T 72 51
G GGGGG 62 30
A AAAAA 17 0
D DDDDD 102 79
D DDAKE 69 58
T SSTTV 1 69
V IIVIV E 14 E 0
L VVIIVL E 0 E 0

0 10 20 30 40 50 60 70 80 90 kcal/mol
CATH: 50 structures - 1 superfamily

superfamily
3.40.640.10

Type I PLP-dependent aspartate aminotransferase-like (Major domain)
Dynamic programming: optimal alignment

Pair of protein sequences
U  GGQLAKEEAL
T  EGQPVEVL

Optimal alignment (no gaps)
U  GGQLAKEEAL
T1 EVL
T2 EGQPVEVL

Optimal alignment (with gaps)
U  GGQLAKEEAL
T  EGQP.VE.VL

• Global/no gap:
  SB Needleman and CD Wunsch 1970 J Mol Biol 48, 443-53
• Local/Gap:
  TF Smith and MS Waterman 1981 J Mol Biol 147, 195-197

\[
SW = \sum_{k=1}^{L_{\text{ali}}} M_{U_k T_k} - G_o \cdot N_{\text{gap}} - G_e \cdot (L_{\text{gap}} - N_{\text{gap}})
\]
Alignments: scoring matrix

\[
\text{Score} = \sum M_{ij} - G_o^* N_{\text{gap}} - G_e^* (L_{\text{gap}} - N_{\text{gap}})
\]

synonyms:
- [scoring | substitution | exchange | log-odds]
- [matrix | metric | table]

one particular:
- Blossum65

<table>
<thead>
<tr>
<th>C</th>
<th>S</th>
<th>T</th>
<th>P</th>
<th>A</th>
<th>G</th>
<th>N</th>
<th>D</th>
<th>E</th>
<th>Q</th>
<th>R</th>
<th>K</th>
<th>M</th>
<th>I</th>
<th>L</th>
<th>V</th>
<th>F</th>
<th>Y</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>1</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>S</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>T</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>P</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>5</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>N</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Q</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>H</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>R</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>K</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>M</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>I</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>L</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>V</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>F</td>
<td>9</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Y</td>
<td>10</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>W</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
</tr>
</tbody>
</table>

BLAST: fast matching of single ‘words’

TTYKLILNGKTLKGETTTEAVDAATAEKVFQYANDNGVDGEWTVDDATKTFTVTEK
TIYKLILQGRTIKAEILITEGVDGATGKVEKQYGNQNAVDAEYTYDNATRTFTITQK

#1 seed=3

TTYKLIL
TIYKLIL

TTYKLIL

#2 extend

AATAEKVFKQYA
GATGKVEKQY

WTYDDATKTF
YTYDNATRTF

Default “word” size for “seeds” = 3
Significance of match (e.g. BLAST E-values)

Score = \sum M_{ij} - G_o N_{gap} - G_e (L_{gap} - N_{gap})

Background (database)
Twilight zone = false positives bazoom!!

Percentage sequence identity

Number of protein pairs

Distance from HSSP threshold

Sequence identity

Number of residues aligned

B Rost 1999 Prot Engin 12, 85-94
Zones

Sequence comparisons: multiple alignment/profile-based
Profiles/PSSM/motifs
Computationally: motifs

retrieved from http://weblogo.berkeley.edu/examples.html

Transcription factor binding sites

Following slides:
thanks kudos to Theresa Wirth
Sequence motifs

Representation of a sequence with more than one possible amino acid (AA) or nuclear acid (NA) at a single position

G-[LI]-L-M-S-A-{RK}-X(1,3)

Two or more possible AA
Disallowed AA
Repetition range X(n,m) of X

One-letter code
Any AA

Matching sequences e.g.
G L L M S A C V V V
G I L M S A Y P P
G L L M S A E S

Fig. 7: M Zvelebil & JO Baum (2008) Understanding Bioinformatics, Garland
PSSM - Position specific substitution matrix: concept

- Matrix of numbers with scores for each residue or nucleotide at each position

<table>
<thead>
<tr>
<th>Building a PSSM:</th>
<th>Starting point:</th>
</tr>
</thead>
<tbody>
<tr>
<td>◦ Absolute frequencies</td>
<td>123456</td>
</tr>
<tr>
<td>◦ Add pseudo-counts if necessary</td>
<td>ATGCTA</td>
</tr>
<tr>
<td>◦ Relative frequencies</td>
<td>ATTGCT</td>
</tr>
<tr>
<td>◦ Log likelihoods</td>
<td>TCTGAG</td>
</tr>
</tbody>
</table>
<pre><code>                                                             | GTTGAG          |
                                                             | CCATCC          |
</code></pre>
PSSM - Position specific substitution matrix: one solution

123456
ATGCTA
ATTGCT
TCTGAG
TGTGAG
CATCC

Absolute frequency

<table>
<thead>
<tr>
<th></th>
<th>A: 2</th>
<th>T: 1</th>
<th>G: 1</th>
<th>C: 1</th>
</tr>
</thead>
</table>

Relative frequency

<table>
<thead>
<tr>
<th></th>
<th>A: 0.4</th>
<th>T: 0.2</th>
<th>G: 0.2</th>
<th>C: 0.2</th>
</tr>
</thead>
</table>

Σ = 1

1 2 3 4 5 6

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.4</td>
<td>0</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>0.6</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.6</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0.2</td>
<td>0.4</td>
<td>0</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Log odds

Log odds = \ln(S ÷ 0.25)

slide: Theresa Wirth
Recap: substitution matrix (BLOSUMUM)

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>S -2</th>
<th>T -2</th>
<th>P -3</th>
<th>A -2</th>
<th>G -3</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-4</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>D-5</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>E -5</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Q -5</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>H -3</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>2</td>
</tr>
<tr>
<td>R -4</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>0</td>
</tr>
<tr>
<td>K -5</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>1</td>
</tr>
<tr>
<td>M -5</td>
<td>-2</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
</tr>
<tr>
<td>L -6</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-4</td>
<td>-3</td>
</tr>
<tr>
<td>V -2</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>F -4</td>
<td>-3</td>
<td>-3</td>
<td>-5</td>
<td>-4</td>
<td>-5</td>
<td>-2</td>
</tr>
<tr>
<td>Y 0</td>
<td>-3</td>
<td>-3</td>
<td>-5</td>
<td>-3</td>
<td>-7</td>
<td>-2</td>
</tr>
<tr>
<td>W -8</td>
<td>-2</td>
<td>-5</td>
<td>-6</td>
<td>-6</td>
<td>-7</td>
<td>-3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>POS</th>
<th>PROBE</th>
<th>CONSENSUS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>E</td>
<td>C</td>
</tr>
<tr>
<td>2</td>
<td>L</td>
<td>C</td>
</tr>
<tr>
<td>3</td>
<td>V</td>
<td>C</td>
</tr>
<tr>
<td>4</td>
<td>K</td>
<td>C</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>6</td>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td>7</td>
<td>S</td>
<td>C</td>
</tr>
<tr>
<td>8</td>
<td>S</td>
<td>C</td>
</tr>
<tr>
<td>9</td>
<td>V</td>
<td>C</td>
</tr>
<tr>
<td>10</td>
<td>K</td>
<td>C</td>
</tr>
<tr>
<td>11</td>
<td>M</td>
<td>C</td>
</tr>
<tr>
<td>12</td>
<td>S</td>
<td>C</td>
</tr>
<tr>
<td>13</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>14</td>
<td>K</td>
<td>C</td>
</tr>
<tr>
<td>15</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>16</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>17</td>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td>18</td>
<td>Y</td>
<td>C</td>
</tr>
<tr>
<td>19</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>20</td>
<td>F</td>
<td>C</td>
</tr>
<tr>
<td>21</td>
<td>S</td>
<td>C</td>
</tr>
<tr>
<td>22</td>
<td>S</td>
<td>C</td>
</tr>
<tr>
<td>23</td>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td>24</td>
<td>D</td>
<td>C</td>
</tr>
<tr>
<td>25</td>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td>26</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>27</td>
<td>Y</td>
<td>C</td>
</tr>
<tr>
<td>28</td>
<td>E</td>
<td>C</td>
</tr>
<tr>
<td>29</td>
<td>L</td>
<td>C</td>
</tr>
<tr>
<td>30</td>
<td>Y</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>S</td>
<td>C</td>
</tr>
<tr>
<td>49</td>
<td>S</td>
<td>C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PROFILE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A   B   C   D   E   F   G   H   I   K   L   M   N   P   Q   R   S   T   V   W   Y</td>
</tr>
<tr>
<td>3 -2  3  4 -1  3 -1  4  1  1  1 -2  1  2  6 -6 -2</td>
</tr>
<tr>
<td>2 -2 -1 -3  1  3 -1  6 -5 -1  3  0 -1  3  1  4  1 -1</td>
</tr>
<tr>
<td>2  2 -2 -2  2  2 -3  11 -2  8 -2  1 -2 -2  0  2  15 -9 -1</td>
</tr>
<tr>
<td>6 -2  5  6 -5  4  1  0  5 -2  0  3  3  3  1  3  6  0 -6 -4</td>
</tr>
<tr>
<td>6 -1  0 -1 -2  2  0  1  0  2  2  0  8  2  0  2  2  3 -5 -4</td>
</tr>
<tr>
<td>7  1 -2 -6  2  2 -2  3 -2  4  3  6  1  6  2 -1 -6 -5</td>
</tr>
<tr>
<td>4  4  2 -2  4 -1  0  2  3 -2  2  7  0  1  10  6  0 -2 -4</td>
</tr>
<tr>
<td>5  0 -1 -1  3  1 -2 -7 -2  7 -2  7 -6 -1 -1 -3  0  2  10 -5 -1</td>
</tr>
<tr>
<td>0 -1  1 -5  0  2 -2  8 -3  1  3  3  3  10 -5  1 -2 -7 -5</td>
</tr>
<tr>
<td>0  2 -3 -2  7 -3 -3 -11 -1  11 10 -2 -2 -1 -2 -2 -2 -2 1 -9 3 1</td>
</tr>
<tr>
<td>4  6  2  2 -3 -5 -1  0  2 -3 -2  3 -4 -1  1  12 6 0 0 -4</td>
</tr>
<tr>
<td>3 15 -5 -5 -1 2 1 -3 5 -8 -6 -3 1 -6 -3 7 -3 3 -13 10</td>
</tr>
<tr>
<td>0 -2  7 -3 -3 -11 -1 11 10 -2 -2 -1 -2 -2 1 -9 3 1</td>
</tr>
<tr>
<td>0  5 -1  0  2  2  2  3  1  1 3 1 7 2 1 2 -1 2</td>
</tr>
<tr>
<td>1 2 0  4  0  0 -2  2  2  2  2  3  1  1 3 1 7 2 1 2</td>
</tr>
<tr>
<td>3 2 5  4 -4 -5 0 -1 2 -3 -2 4 3 1 1 8 -2 -1 -2 -3</td>
</tr>
<tr>
<td>23 0  2  1 -2  0  0  0  1 -2 -1 2 2 0 1 8 2 0 1 2</td>
</tr>
<tr>
<td>23 0  2  1 -2  0  0  0  1 -2 -1 2 2 0 1 8 2 0 1 2</td>
</tr>
<tr>
<td>23 0  2  1 -2  0  0  0  1 -2 -1 2 2 0 1 8 2 0 1 2</td>
</tr>
<tr>
<td>23 0  2  1 -2  0  0  0  1 -2 -1 2 2 0 1 8 2 0 1 2</td>
</tr>
<tr>
<td>0 -5 0 -1 5 -1 2 1 -1 0 -1 5 2 2 -1 -3 3 1 5 -3</td>
</tr>
<tr>
<td>5 -2  9  8 -3 3 4 -1 1 -3 -2 5 -1 4 -1 1 -1 -6 0</td>
</tr>
<tr>
<td>3 -5 -3 -1 6 -1 -2 6 -1 10 10 -2 0 0 -2 -1 0 6 -1 0</td>
</tr>
<tr>
<td>0  1 3 2 0 2 3 -1 1 -1 -1 8 0 -1 -1 2 1 1 -1</td>
</tr>
</tbody>
</table>

Profiles profit from relation of “families”
Scoring matrix: generic vs. specific

You = {TUM, Student, Age}

You = part of your circle

Generic

(Here Blossum62)

Position-specific scoring matrix

PSSM

\[
\begin{array}{cccccc}
\text{C} & \text{G} & \text{A} & \text{T} & \text{P} & \text{S} \\
\text{B} & 0 & 2 & 0 & 0 & 0 \\
\text{T} & 2 & 1 & 3 & 0 & 0 \\
\text{F} & 3 & 1 & 0 & 6 & 0 \\
\text{A} & 2 & 1 & 1 & 1 & 2 \\
\text{G} & 3 & 1 & 0 & 4 & 1 & 5 \\
\text{N} & 4 & 1 & 0 & -1 & 0 & 0 \\
\text{D} & 5 & 0 & 0 & -1 & 0 & 1 \\
\text{E} & 6 & 0 & 0 & -1 & 0 & 0 \\
\text{Q} & 7 & 4 & 1 & 0 & 0 & 4 \\
\text{H} & 8 & 1 & -1 & 0 & 1 & 1 \\
\text{R} & 9 & 0 & -1 & 0 & -3 & 0 \\
\text{K} & 10 & 0 & 0 & -1 & 0 & 0 \\
\text{M} & 11 & 0 & 0 & -1 & 0 & 0 \\
\text{l} & 12 & 1 & -1 & 0 & -3 & 0 \\
\text{L} & 13 & 2 & -2 & 0 & -4 & 0 \\
\text{V} & 14 & 2 & -2 & 0 & -4 & 0 \\
\text{F} & 15 & 2 & -2 & 0 & -4 & 0 \\
\text{Y} & 16 & 2 & -2 & 0 & -4 & 0 \\
\text{W} & 17 & 2 & -2 & 0 & -4 & 0 \\
\end{array}
\]
what is the difference: profile vs motif
PSI-Blast
Idea: replace generic by specific scoring

Generic scoring matrix (here BLOSUM62)

```
S T P A G N D E Q R K M L V F Y W
C 1 2
S 0 2
T -2 1 3
P -3 1 0 6
A -2 1 1 1 2
G -3 1 0 -1 1 5
N -4 1 0 -1 0 2
D -5 0 0 -1 0 1 2 4
E -5 0 0 -1 0 0 1 3 6
Q -5 -4 0 0 -1 1 2 2 4
H -3 -1 -1 0 -1 -2 2 1 1 3 6
R -4 0 1 0 2 -3 0 1 -1 2 6
K -5 0 0 -1 -1 -2 1 0 0 1 0 3 5
M -5 -2 -1 -2 -1 -3 2 3 -2 -1 2 0 0 6
L -1 2 -1 0 2 1 -3 2 2 2 2 2 2 5
S -6 3 2 2 3 2 4 3 4 3 2 2 3 4 2 6
V -2 1 0 0 1 -1 2 -2 -2 2 2 2 2 4 2 4 4
F -4 3 3 5 4 5 2 6 -5 5 2 4 -5 6 1 2 -1 9
Y 0 -3 -3 -5 3 -7 2 4 4 4 0 4 4 -2 -1 1 2 7 10
W -3 2 -5 -5 -6 -7 -4 -7 -7 -5 3 2 -3 -4 -5 -2 1 6 0 0 17
```


© Burkhard Rost

ROSTLAB.
Sr. Investigator - NCBI @ NIH, Bethesda, MD

- Education: Mathematics/Statistics/Biology
- PhD - Harvard Univ/MIT (Supervisor: Daniel Kleitman, comp bio)
- MS - Harvard (mathematics)

~70 publications (May 2014)
- 2 over 50,000 citations
- 8 over 1,000
- 30 over 100
- >120k citations

H-index: 44 (May 2014)
David J Lipman

- Director of NCBI @ NIH, Bethesda, MD (since 1989)

Education:
- MD - State Univ NY Buffalo
- MS - Brown Univ. (mathematics)

- 71 publications (May 2012)
  - 2 over 30,000 citations
  - 6 over 1,000
  - 38 over 100

H-index: 47 (May 2012)

Awards:
- ISCB Fellow, ISCB Sr. Scientist Award
- US National Academy of Sciences

David J Lipman
NCBI Bethesda USA

photo: [http://jamia.bmj.com/content/9/4/409.full](http://jamia.bmj.com/content/9/4/409.full)
Sequence-profile methods

- **PSI-BLAST**
  - fast, partial dynamic programming

~50,000 citations in Google Scholar May 2014

Stephen Altschul
David Lipman
Tom Madden
Web Miller
Alejandro Schäffer
PSI-BLAST

sketch of solution
PSI-BLAST in steps

1. fast hashing
Like BLAST match ‘words’

Default “word” size for “seeds” = 3
Like BLAST match ‘words’

TTYKLI<ILNGKTLKGETTTEAVDAATAEKFVFKQYANDNGVDGEWTVYDDATKTFTVTEK
TTYKLI<ILLLLLLLLLLLLLLLLLLLLL<AWTVEKAFKTF<FAAAAAAAAAAWTVEKAFKTF<AAA

#1 seed=3

TTYKLI<IL TTYKLI<IL

#2 extend

AATAEKFVFKQYA
AWTVEKAFKTF
WTYDDATKTF
WTVEKAFKTF

Default “word” size for “seeds” = 3

note: word matches are not identical! (as in FASTA)
Neighborhood substitutions

BLAST matches word of 3

YES to all with substitution > threshold $T$
e.g. $T=13$ -> neighborhood

YES-YES
YES-YEA
YES-YEN

$$S_{ij} = \left( \frac{1}{\lambda} \right) \log \left( \frac{p_{ij}}{q_i \ast q_j} \right)$$
Btw. Blosum 62


Btw. Blosum 62

\[ S_{ij} = \left( \frac{1}{\lambda} \right) \log \left( \frac{p_{ij}}{q_i \times q_j} \right) \]

<table>
<thead>
<tr>
<th></th>
<th>Ala</th>
<th>Arg</th>
<th>Asn</th>
<th>Asp</th>
<th>Cys</th>
<th>Gln</th>
<th>Glu</th>
<th>Gly</th>
<th>His</th>
<th>Ile</th>
<th>Leu</th>
<th>Lys</th>
<th>Met</th>
<th>Phe</th>
<th>Pro</th>
<th>Ser</th>
<th>Thr</th>
<th>Trp</th>
<th>Tyr</th>
<th>Val</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg</td>
<td>-1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asn</td>
<td>-2</td>
<td>0</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asp</td>
<td>-2</td>
<td>-2</td>
<td>1</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cys</td>
<td>0</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gln</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-3</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glu</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>-4</td>
<td>2</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gly</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>His</td>
<td>-2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>-3</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ile</td>
<td>-1</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>-3</td>
<td>-3</td>
<td>-4</td>
<td>-3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leu</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-3</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lys</td>
<td>-1</td>
<td>2</td>
<td>0</td>
<td>-1</td>
<td>-3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phe</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-3</td>
<td>0</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pro</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>-2</td>
<td>-4</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ser</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thr</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trp</td>
<td>-3</td>
<td>-3</td>
<td>-4</td>
<td>-4</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>-1</td>
<td>1</td>
<td>4</td>
<td>-3</td>
<td>-2</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tyr</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>1</td>
<td>5</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Val</td>
<td>0</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>1</td>
<td>-2</td>
<td>-2</td>
<td>1</td>
<td>-2</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>2</td>
<td>-2</td>
<td>-2</td>
<td>0</td>
<td>3</td>
<td>-1</td>
<td>4</td>
</tr>
</tbody>
</table>

Btw. Blosum 62


\[ S_{ij} = \left( \frac{1}{\lambda} \right) \log \left( \frac{p_{ij}}{q_i * q_j} \right) \]

<table>
<thead>
<tr>
<th></th>
<th>Ala</th>
<th>Arg</th>
<th>Asn</th>
<th>Asp</th>
<th>Cys</th>
<th>Gln</th>
<th>Glu</th>
<th>Gly</th>
<th>His</th>
<th>Ile</th>
<th>Leu</th>
<th>Lys</th>
<th>Met</th>
<th>Phe</th>
<th>Pro</th>
<th>Ser</th>
<th>Thr</th>
<th>Trp</th>
<th>Tyr</th>
<th>Val</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg</td>
<td>-1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asn</td>
<td>-2</td>
<td>0</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asp</td>
<td>-2</td>
<td>-2</td>
<td>1</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cys</td>
<td>0</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gln</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-3</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glu</td>
<td>-1</td>
<td>0</td>
<td>2</td>
<td>-4</td>
<td>2</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gly</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>2</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>His</td>
<td>-2</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>-3</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ile</td>
<td>-1</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>-3</td>
<td>-3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leu</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-3</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lys</td>
<td>-1</td>
<td>2</td>
<td>0</td>
<td>-1</td>
<td>-3</td>
<td>1</td>
<td>1</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>2</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>1</td>
<td>2</td>
<td>-1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phe</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-3</td>
<td>0</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pro</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>2</td>
<td>-4</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ser</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thr</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>2</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trp</td>
<td>-3</td>
<td>-3</td>
<td>-4</td>
<td>-4</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>3</td>
<td>-3</td>
<td>1</td>
<td>1</td>
<td>-4</td>
<td>-3</td>
<td>-2</td>
<td>1</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tyr</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>2</td>
<td>-1</td>
<td>1</td>
<td>-2</td>
<td>1</td>
<td>3</td>
<td>-3</td>
<td>-2</td>
<td>2</td>
<td>2</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Val</td>
<td>0</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>-2</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

PSI-BLAST in steps

1. fast hashing
2. dynamic programming extension between matches
BLAST + Smith-Waterman

TTYKLILNGKTLKGETTTEAVDAATAEKFQKQYANDNGVDGEWTYDDATKTFVTTEK
TTYKLIPLLLLLLLLLLLLLLLLLAWTVEKAFKTFAAABBBBBBBBBBBBBBBBB

#1 seed=3

#2 extend

dynamic programming to extend
BLAST + Smith-Waterman

TTYKLILNGKTLKGETTTEAVDAATAEKFQKQANDNGVDGEWTVYDDATKTFVTVEK TTYKLILLLLLLLLLLLLLLLLLLLAWTVEKAFKTFAAAAAAWTVEKAFKTFAAAAAA

#1 seed=3
#2 extend

dynamic programming to extend

Why is this fast?
PSI-BLAST in steps

1. fast hashing
2. dynamic programming extension between matches
3. compile statistics
   EVAL - Expectation values
PSI-BLAST in steps

1. fast hashing
2. dynamic programming extension between matches
3. compile statistics
4. collect all pairs and build profile
PSI-BLAST in steps

1. fast hashing
2. dynamic programming extension between matches
3. compile statistics
4. collect all pairs and build profile
5. ?
1. **fast hashing**
2. **dynamic programming extension between matches**
3. **compile statistics**
4. **collect all pairs and build profile**
5. **iterate**
Expanding in sequence space: dynamics of PSI-BLAST
Profile-based database search

Family U

© Burkhard Rost
Profile-based database search

Family U

safe for pairwise
Profile-based database search

Family U

zone reached through position-specific family profile

safe for pairwise
Profile-based database search

Family U

zone reached through position-specific family profile

lost after iteration

safe for pairwise
Profile-based database search

Family U

safe zones of close homologues

zone reached through position-specific family profile

lost after iteration

safe for pairwise
Profile-based database search
### Sequence-profile comparison

<table>
<thead>
<tr>
<th>YDFHGVGEDDISIKRG</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>fyn_human VTLFVAPVDY EARTDSLFS HKGEKQILN SSEGDAWR SLTTEGTQY</td>
</tr>
<tr>
<td>yrk_chick VTLFVAPVDY EARTDSLFS QKGEKPHIH NTEGDAWR SLRSSGATQY</td>
</tr>
<tr>
<td>fgr_human VTLFVAPVDY EAMDDRTF TOKFKPILN NTEGDAWR SSLSGKTCIQ</td>
</tr>
<tr>
<td>yes_chick VTVFVAPVDY EAMDDRTF HKGERQIIVN NTEGDAWR SIATGKTY</td>
</tr>
<tr>
<td>src_avis2 VTVFVAPVDY ESRTETDSF HKGERQIIVN NTEGDAWR SLTTEGTQY</td>
</tr>
<tr>
<td>src_avis3 VTVFVAPVDY ESRTETDSF HKGERQIIVN NTEGDAWR SLTTEGTQY</td>
</tr>
<tr>
<td>src_chick VTVFVAPVDY ESRTETDSF HKGERQIIVN NTEGDAWR SLTTEGTQY</td>
</tr>
<tr>
<td>stk_hydat VTVFVAPVDY EARISSDLFS HKGERQIIVN TADGDAWR SLITNSEQ</td>
</tr>
<tr>
<td>src_rsvpa .......... ESRTETDSF KKRERIIVN NTEGDAWR SLTTEGTQY</td>
</tr>
<tr>
<td>hck_humam ..IVLAFYV EAIIHEDLSF QKGDQMVLLE ES.GEWWAK SLATRKEQY</td>
</tr>
<tr>
<td>blk_mouse ..FVLAFYV AAALRLDQLV KGEKIQVLR .STGDAWLR SVTQIEGQ</td>
</tr>
<tr>
<td>hck_mouse ..TVLAFYV EAIIHEDLSF QKGDQMVLLE .EAGEWAKAR SLATKKEQY</td>
</tr>
<tr>
<td>lyn_humam ..IVLAFYV DIHFPDSLFS HKGEKPVLE .EHGWAKAR SLTKEQFI</td>
</tr>
<tr>
<td>lck_humam ..LVIALHSY EPSSHGDGPF EKGQIRLE QS.GEWWAK SLTQIEQF</td>
</tr>
<tr>
<td>ss81_yeast.... ALVPY DADVDPAPF EOEIILQVSD .IEGRWAKAR R.ANGEFL</td>
</tr>
<tr>
<td>abl_mouse ..LFVAFYDF VASGINTLSI TKGEKIRVLG YnNGEWQAQ ..TKNGQQW</td>
</tr>
<tr>
<td>abl1_humam ..LFVAFYDF VASGINTLSI TKGEKIRVLG YnNGEWQAQ ..TKNGQQW</td>
</tr>
<tr>
<td>src1_drome ..VVFVAFYDF KRDESDLFS MKGRVKMEID DTESDAVRV NL TRQELI</td>
</tr>
<tr>
<td>mysd_dicdi .... ALDVDF AESSMEMSF KEGDIIYVVD QSSDLVDAE L..KGRKKY</td>
</tr>
<tr>
<td>yfj4_yeast.... ALVSF AGEESGDPF RKGDVITILK KsQNDWVTGR V..NGRECIF</td>
</tr>
<tr>
<td>abl2_humam ..LFVAFYDF VASGINTLSI TKGEKIRVLG YNNGEVSF RSKG_QQW</td>
</tr>
<tr>
<td>tec_humam ..EIVVAMYDF QAAGHDRLR ERQEGYLLE KNDVHWRAR D.KYNGQY</td>
</tr>
<tr>
<td>abl1_caeel ..LFVAFYDF VASGINTLSI TKGEKIRVLG YNNGEVSF RSKG_QQW</td>
</tr>
<tr>
<td>yha2_yeastVRRVAFYDF TTNEPDLSF RKGDVITVLE QVYRDPAKG L..RGNMCEF</td>
</tr>
<tr>
<td>abpl_sacex .... AEDYD EAGENELTF AENDKIINIE FVDDCAWLGE LETTGQKGF</td>
</tr>
</tbody>
</table>

---

**PSI-BLAST** SF Altschul 1997 *Nucl Acids Res* 25 3389-3402

© Burkhard Rost

ROSTLAB. TUM

51/157
Sequence-profile methods

- **PSI-BLAST**
  - fast, partial dynamic programming
  - SF Altschul (1997) NAR 25:3389-3402

- **ClustalW/ClustalX**
  - slow, dynamic programming, for experts
  - JD Thompson, DG Higgins, TJ Gibson (1994) NAR 22:4673-80
Clustal
Clustal - naming convention

- 1988 ClustalV
- 1994 ClustalW
- 1997 ClustalX
- 2011 ClustalOmega


Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Soeding, Julie D Thompson, Desmond G Higgins
Clustal (ClustalW, ClustalX)

- all against all (pairs)
  - by dynamic programming
  - (varying substitution matrices)
- build phylogenetic tree

JD Thompson, DG Higgins, TJ Gibson (1994) NAR 22:4673-80
Clustal (ClustalW, ClustalX)

- all pairs (dynamic programming with varying substitution matrices)
- create phylogenetic tree
- cluster and dynamic programming

JD Thompson, DG Higgins, TJ Gibson (1994) NAR 22:4673-80
Fig. 1:
Clustal (ClustalW, ClustalX)

>100,000 citations in GoogleScholar


GoogleScholar May 2014

<table>
<thead>
<tr>
<th>Method</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustal</td>
<td>3,211</td>
</tr>
<tr>
<td>ClustalV</td>
<td>2,601</td>
</tr>
<tr>
<td>ClustalW</td>
<td>47,968</td>
</tr>
<tr>
<td>ClustalX</td>
<td>30,044</td>
</tr>
<tr>
<td>ClustalOmega</td>
<td>544</td>
</tr>
</tbody>
</table>
Desmond G Higgins

- Prof. University College Dublin, Ireland
- over 110,000 citations ISI (May 2014)
- 94 publications (May 2012)
  - 2 over 30,000 citations
  - 3 over 10,000
  - 10 over 1,000
  - 30 over 100
- H-index: 52 (May 2014)
Group leader EMBL Heidelberg
over 70,000 citations ISI (May 2012)
128 publications (May 2012)
- 1 over 35,000 citations
- 2 over 20,000
- 8 over 1,000
- 28 over 100
H-index: 47 (May 2012)
Julie Dawn Thompson

Sr. Scientist, Inst. of Genetics & Mol & Cellular Biology (IGBMC), Strasbourg, France

- over 70,000 citations ISI (May 2012)
- ~30(?) publications (May 2012)
  - 1 over 35,000 citations
  - 2 over 20,000
  - 6 over 1,000
  - JD Thompson: >1000 publications top 6 from her!
Clustal (ClustalW, ClustalX)

- over 60,000 citations ISI (May 2012)

Julie D Thompson, Desmond G Higgins, Toby J Gibson (1994) NAR 22:4673-80
>47,000 citations (May 2014)

Des Higgins
Toby Gibson
Julie Dawn Thompson
Sequence-profile methods

- **PSI-BLAST**
  - fast, partial dynamic programming
  - SF Altschul (1997) NAR 25:3389-3402

- **ClustalW/ClustalX**
  - slow, dynamic programming, for experts
  - JD Thompson, DG Higgins, TJ Gibson (1994) NAR 22:4673-80

- **MaxHom**
  - relatively slow, dynamic programming, good first guess
Homology-derived protein structures and the structural meaning of sequence alignment

Reinhard Schneider (1994) Sequenz- und Struktur Vergleiche und deren Anwendung für die Struktur- und Funktionsvorhersage von Proteinen (PhD Heidelberg University)
### Maxhom/HSSP

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>90</td>
<td>80</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>90</td>
<td>80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Sweep 1**

- A → B
- A → C
- B → A
- A → D

-> Profile (P0)

**conservation weight (cw0)**

**Sweep 2**

- P0
- cw0
- B
- P1
- cw0
- C
- P1
- cw0
- D
Philosophy: family vs. protein

<table>
<thead>
<tr>
<th>Protein-centric (MaxHom/PSI-BLAST)</th>
<th>Family-centric (Clustal/HMM/T-coffee)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LWYGQQAR KSQDKAKHAF</td>
<td>LWYG.Q.R K.....AKHAF</td>
</tr>
<tr>
<td>IWyAAEER KSQDKAKHAF</td>
<td>IWYA.E.R.....AKHAF</td>
</tr>
<tr>
<td>IWYAAEAK RPDQRAKHGF</td>
<td>IWYAE..K.....RAKHGF</td>
</tr>
<tr>
<td>IWYAAEER K--QKARVVF</td>
<td>IWYAE..R..KQKARVVF</td>
</tr>
<tr>
<td>HAYIVVGR ----KARVVF</td>
<td>HAYIV..R....KARVVF</td>
</tr>
<tr>
<td>HAYI---R ----KARVVF</td>
<td>HAYI---R....KARVVF</td>
</tr>
</tbody>
</table>

original sequence

| LWYGQQAR KSQDKAKHAF |
| LWYGQ..R ....KAKHAF |
another way to do math: HMM - Hidden Markov Models
HMM: Hidden Markov Models

Fig. 1. A toy HMM, modeling sequences of as and bs as two regions of potentially different residue composition. The model is drawn (top) with circles for states and arrows for state transitions. A possible state sequence generated from the model is shown, followed by a possible symbol sequence. The joint probability $P(x, \pi | \text{HMM})$ of the symbol sequence and the state sequence is a product of all the transition and emission probabilities. Notice that another state sequence (1-2-2) could have generated the same symbol sequence, though probably with a different total probability. This is the distinction between HMMs and a standard Markov model with nothing to hide: in an HMM, the state sequence (e.g. the biologically meaningful alignment) is not uniquely determined by the observed symbol sequence, but must be inferred probabilistically from it.

Marco Punta contributed the slides

- PhD in Trieste (MD for membrane proteins)
- Postdoc @ Columbia Univ in the City of New York (contact predictions)
- Senior scientist in NYCOMPS (Target selection for membrane proteins)
- IAS Fellow @ TUM
- Project manager @ Pfam @ Sanger Inst. Hinxton (Cambridgeshire)
- now CNRS Paris
Example traffic light
Example traffic light: NOT HMM

Deterministic

R → Y → G
HMM example 2: weather
HMM example 2: weather
HMM example 2: weather

Probabilistic

*In fact, chaotic, deterministic
HMM example 2: weather

$p_{ij}$ Transition probability from symbol $i$ to symbol $j$

*In fact, chaotic, deterministic*
HMM example 2: weather

Example 2: The weather

Probabilistic, 1\textsuperscript{st} order Markov model

$$P(X_{n+1} = x | X_1 = x_1, \ldots, X_n = x_n) = P(X_{n+1} = x | X_n = x_n)$$
# ProfileHMM: example football

## FC Augsburg 2013/14 Bundesliga Fixtures

<table>
<thead>
<tr>
<th>Date</th>
<th>Status</th>
<th>Home</th>
<th>Score</th>
<th>Away</th>
<th>Attendance</th>
<th>Competition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aug 10</td>
<td>FT</td>
<td>FC Augsburg</td>
<td>0-4</td>
<td>Borussia Dort.</td>
<td>30,660</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Aug 17</td>
<td>FT</td>
<td>Werder Bremen</td>
<td>1-0</td>
<td>FC Augsburg</td>
<td>40,112</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Aug 25</td>
<td>FT</td>
<td>FC Augsburg</td>
<td>2-1</td>
<td>VfB Stuttgart</td>
<td>30,030</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Aug 31</td>
<td>FT</td>
<td>Nurnberg</td>
<td>0-1</td>
<td>FC Augsburg</td>
<td>37,239</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Sep 14</td>
<td>FT</td>
<td>FC Augsburg</td>
<td>2-1</td>
<td>SC Freiburg</td>
<td>28,453</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Sep 21</td>
<td>FT</td>
<td>Hannover 96</td>
<td>2-1</td>
<td>FC Augsburg</td>
<td>39,200</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Sep 27</td>
<td>FT</td>
<td>FC Augsburg</td>
<td>2-2</td>
<td>Borussia Mon.</td>
<td>30,352</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Oct 5</td>
<td>FT</td>
<td>Schalke 04</td>
<td>4-1</td>
<td>FC Augsburg</td>
<td>60,731</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Oct 20</td>
<td>FT</td>
<td>FC Augsburg</td>
<td>1-2</td>
<td>VfL Wolfsburg</td>
<td>27,554</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Oct 26</td>
<td>FT</td>
<td>Bayer Leverk</td>
<td>2-1</td>
<td>FC Augsburg</td>
<td>27,811</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Nov 3</td>
<td>FT</td>
<td>FC Augsburg</td>
<td>2-1</td>
<td>Mainz</td>
<td>28,007</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Nov 9</td>
<td>FT</td>
<td>Bayern Munich</td>
<td>3-0</td>
<td>FC Augsburg</td>
<td>71,000</td>
<td>Bundesliga</td>
</tr>
</tbody>
</table>
### ProfileHMM: example football

**FC Augsburg 2013/14 Bundesliga Fixtures**

<table>
<thead>
<tr>
<th>Date</th>
<th>Status</th>
<th>Home</th>
<th>Score</th>
<th>Away</th>
<th>Attendance</th>
<th>Competition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aug 10</td>
<td>FT</td>
<td>FC Augsburg</td>
<td>L</td>
<td>L</td>
<td>0-4</td>
<td>Borussia Dort.</td>
</tr>
<tr>
<td>Aug 17</td>
<td>FT</td>
<td>Werder Bremen</td>
<td>L</td>
<td>L</td>
<td>1-0</td>
<td>FC Augsburg</td>
</tr>
<tr>
<td>Aug 25</td>
<td>FT</td>
<td>FC Augsburg</td>
<td>W</td>
<td>W</td>
<td>2-1</td>
<td>VfB Stuttgart</td>
</tr>
<tr>
<td>Aug 31</td>
<td>FT</td>
<td>Nurnberg</td>
<td>W</td>
<td>W</td>
<td>0-1</td>
<td>FC Augsburg</td>
</tr>
<tr>
<td>Sep 14</td>
<td>FT</td>
<td>FC Augsburg</td>
<td>W</td>
<td>W</td>
<td>2-1</td>
<td>SC Freiburg</td>
</tr>
<tr>
<td>Sep 21</td>
<td>FT</td>
<td>Hannover 96</td>
<td>L</td>
<td>L</td>
<td>2-1</td>
<td>FC Augsburg</td>
</tr>
<tr>
<td>Sep 27</td>
<td>FT</td>
<td>FC Augsburg</td>
<td>D</td>
<td>D</td>
<td>2-2</td>
<td>Borussia Mon.</td>
</tr>
<tr>
<td>Oct 5</td>
<td>FT</td>
<td>Schalke 04</td>
<td>L</td>
<td>L</td>
<td>4-1</td>
<td>FC Augsburg</td>
</tr>
<tr>
<td>Oct 20</td>
<td>FT</td>
<td>FC Augsburg</td>
<td>L</td>
<td>L</td>
<td>1-2</td>
<td>VfL Wolfsburg</td>
</tr>
<tr>
<td>Oct 26</td>
<td>FT</td>
<td>Bayer Leverk</td>
<td>L</td>
<td>L</td>
<td>2-1</td>
<td>FC Augsburg</td>
</tr>
<tr>
<td>Nov 3</td>
<td>FT</td>
<td>FC Augsburg</td>
<td>W</td>
<td>W</td>
<td>2-1</td>
<td>Mainz</td>
</tr>
<tr>
<td>Nov 9</td>
<td>FT</td>
<td>Bayern Munich</td>
<td>L</td>
<td>L</td>
<td>3-0</td>
<td>FC Augsburg</td>
</tr>
</tbody>
</table>
ProfileHMM: example football

W | D | L

Probabilistic

W → D → L → W
ProfileHMM: example football

Our model for Augsburg’s Bundesliga results:

3 states: W, D, L

\[ S(t) = F(S(t-1)) \]

States S connected by probabilities

\[ p_{ij} \geq 0; \quad \sum_j p_{ij} = 1 \]
## ProfileHMM: example football

### FC Augsburg 2013/14 Bundesliga Fixtures

<table>
<thead>
<tr>
<th>Date</th>
<th>Status</th>
<th>Home</th>
<th>Score</th>
<th>Away</th>
<th>Attendance</th>
<th>Competition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aug 10</td>
<td>FT</td>
<td>FC Augsburg</td>
<td>0-4</td>
<td>Borussia Dort.</td>
<td>30,660</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Aug 17</td>
<td>FT</td>
<td>Werder Bremen</td>
<td>1-0</td>
<td>FC Augsburg</td>
<td>40,112</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Aug 25</td>
<td>FT</td>
<td>FC Augsburg</td>
<td>2-1</td>
<td>VfB Stuttgart</td>
<td>30,030</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Aug 31</td>
<td>FT</td>
<td>Nurnberg</td>
<td>0-1</td>
<td>FC Augsburg</td>
<td>37,239</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Sep 14</td>
<td>FT</td>
<td>FC Augsburg</td>
<td>2-1</td>
<td>SC Freiburg</td>
<td>28,453</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Sep 21</td>
<td>FT</td>
<td>Hannover 96</td>
<td>2-1</td>
<td>FC Augsburg</td>
<td>39,200</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Sep 27</td>
<td>FT</td>
<td>FC Augsburg</td>
<td>2-2</td>
<td>Borussia Mon.</td>
<td>30,352</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Oct 5</td>
<td>FT</td>
<td>Schalke 04</td>
<td>4-1</td>
<td>FC Augsburg</td>
<td>60,731</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Oct 20</td>
<td>FT</td>
<td>FC Augsburg</td>
<td>1-2</td>
<td>VfL Wolfsburg</td>
<td>27,554</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Oct 26</td>
<td>FT</td>
<td>Bayer Leverk</td>
<td>2-1</td>
<td>FC Augsburg</td>
<td>27,811</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Nov 3</td>
<td>FT</td>
<td>FC Augsburg</td>
<td>2-1</td>
<td>Mainz</td>
<td>28,007</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Nov 9</td>
<td>FT</td>
<td>Bayern Munich</td>
<td>3-0</td>
<td>FC Augsburg</td>
<td>71,000</td>
<td>Bundesliga</td>
</tr>
</tbody>
</table>
ProfileHMM: example football
Sequence-profile methods

- **PSI-BLAST**
  - fast, partial dynamic programming
  - SF Altschul (1997) NAR 25:3389-3402

- **ClustalW/ClustalX**
  - slow, dynamic programming, for experts
  - JD Thompson, DG Higgins, TJ Gibson (1994) NAR 22:4673-80

- **MaxHom**
  - relatively slow, dynamic programming, good first guess

- **SAM/HMMer**
  - slow, need preprocess, HMM (statistics), very accurate
HMM & biology: SAM & HMMer

- R Durbin, S Eddy, A Krog & G Mitchison: Probabilistic models of proteins and nucleic acids, Cambridge University Press

Anders Krogh

Sean Eddy

Kevin Karplus

David Haussler
HMM for alignment
Generic Profile-HMM for alignment

- Captures matches, insertions and deletions
- Transition and emission probabilities
- Gap penalty handled by variation of transition probabilities
- Calculation of probability by multiplication of path variables
Hidden Markov Models (HMM) - SAM

SAM-T02 web site, UCSC, Kevin Karplus

Entropy in alignment

☐ consider residue position i
BEFORE any amino acid is aligned, we expect a particular acid according to some prior or background probability, P0, with entropy H0

☐ now consider same column AFTER alignment posterior probability Pi + priors -> Hi

☐ if conserved: Hi -> 0; if varied: Hi -> H0

☐ Hi-H0 reflects the “bits saved” by the alignment

2crd  XFTNVSCCTTSKECWSVCQRLHNTSRGKCMNKKCRCYS

2
tbits saved

1

0
Alignment entropy for small families

- few members / little divergence
  entropy dominated by priors
- the background signal dominates
Alignment entropy for large families

- many members/high divergence:
  - entropy dominated by observed profile
    - profile dominates

- problem: possible over-training

2crd  XFTNVSCTTSKECWSVCQRLHNTSRGKCMNKKCRCYS
SAM-T98: Build alignment

SAM-T98 Alignment Building

1. Start: a single sequence
2. Build a model from the sequence or alignment
3. Use the model to search for additional homologs
4. Reestimate the alignment with the new homologs (Iterations 1 - 3)
5. (Iteration 4)

End: a SAM-T98 alignment
Hidden Markov Models (HMM) - SAM

SAM-T02 web site, UCSC, Kevin Karplus


Sequence-profile methods

- **PSI-BLAST**
  - fast, partial dynamic programming
  - SF Altschul (1997) NAR 25:3389-3402

- **ClustalW/ClustalX**
  - slow, dynamic programming, for experts
  - JD Thompson, DG Higgins, TJ Gibson (1994) NAR 22:4673-80

- **MaxHom**
  - relatively slow, dynamic programming, good first guess

- **SAM/HMMer**
  - slow, need preprocess, HMM (statistics), very accurate

- **T-Coffee**
  - much slower, requires preprocessing, Genetic Algorithm
Genetic Algorithm for alignment
Independence assumption

dynamic programming (smith-waterman)

PSI-BLAST

sequence-sequence or sequence-profile

SAM HMMer
Independence assumption

Dynamic programming
(smith-waterman)

PSI-BLAST

SAM
HMMer

Sequence-sequence
or
Sequence-profile

All assume that
alignment at position i
independent of
alignment at position j
Genetic algorithm does not make the independence assumption
Genetic algorithm operates on segments
Genetic algorithm - concept

The 2006 NASA ST5 spacecraft antenna. This complicated shape was found through a genetic algorithm optimizing the radiation pattern. It is known as an Evolved antenna.
Genetic algorithm - concept

The 2006 NASA ST5 spacecraft antenna. This complicated shape was found through a genetic algorithm optimizing the radiation pattern. It is known as an Evolved antenna.

- initialize population
- compute fitness
- crossover parents
- child
- mutation
- roulette parents
- cross-over to children
- mutate children
- compute fitness
- stop
- end

© Burkhard Rost
© Wikipedia
T-Coffee

- much slower, requires preprocessing, Genetic Algorithm

Cedric Notredame

Des Higgins

Jaap Heringa
Cedric Notredame: thanks for slides!
T-Coffee Genetic algorithm (GA)

Begin with “library” of local and global pairwise alignments
T-Coffee: Mix local and global alignment

ClustalW Primary Library (Global Pairwise Alignment)

Lalign Primary Library (Local Pairwise Alignment)

Extension

Multiple Sequence Alignment
T-Coffee: Use more information

Local Alignment

Multiple Alignment

Global Alignment

Specialist

Structural

Multiple Sequence Alignment
### Sequence-profile comparison

<p>| | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>fyn_human</td>
<td>VTLFVALDY</td>
<td>EARTEDDL SF</td>
<td>HKGEKFQILN</td>
<td>SSEGDAEWR</td>
<td>SLTTGETQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>yrk_chick</td>
<td>VTLFIALDY</td>
<td>EARTEDDL SF</td>
<td>QKGEKPHIIN</td>
<td>NTEGDWAR</td>
<td>SLSSGATQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>fgr_human</td>
<td>VTLFIALDY</td>
<td>EARTEDDL TF</td>
<td>TKGEKFIHNL</td>
<td>NTEGDWEAR</td>
<td>SLSSGKTQCI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>yes_chick</td>
<td>VTVFVALDY</td>
<td>EARTDDEL SF</td>
<td>KKGFERQIIN</td>
<td>NTEGDWEAR</td>
<td>SIATGKTQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>src_avis2</td>
<td>VTTFVALDY</td>
<td>ESRTETDL SF</td>
<td>KKGFERQIIVN</td>
<td>NTEGDWLAH</td>
<td>SLTTGQTQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>src_avissa</td>
<td>VTTFVALDY</td>
<td>ESRTETDL SF</td>
<td>KKGFERQIIVN</td>
<td>NTEGDWLAH</td>
<td>SLTTGQTQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>src_chick</td>
<td>VTVFVALDY</td>
<td>ESRTETDL SF</td>
<td>KKGFERQIIVN</td>
<td>NTEGDWLAH</td>
<td>SLTTGQTQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>stk_hydat</td>
<td>VTVFVALDY</td>
<td>ESRTETDL SF</td>
<td>KKGFERQIIVN</td>
<td>NTEGDWLAH</td>
<td>SLTTGQTQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>stk_hydat</td>
<td>VTVFVALDY</td>
<td>ESRTETDL SF</td>
<td>KKGFERQIIVN</td>
<td>NTEGDWLAH</td>
<td>SLTTGQTQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>yes_chick</td>
<td>VTVFVALDY</td>
<td>EARTEDDL SF</td>
<td>HKGEKFQILN</td>
<td>SSEGDAEWR</td>
<td>SLTTGETQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>src_rsvpa</td>
<td>VTVFVALDY</td>
<td>EARTEDDL SF</td>
<td>HKGEKFQILN</td>
<td>SSEGDAEWR</td>
<td>SLTTGETQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>src_chick</td>
<td>VTVFVALDY</td>
<td>EARTEDDL SF</td>
<td>HKGEKFQILN</td>
<td>SSEGDAEWR</td>
<td>SLTTGETQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>stk_hydat</td>
<td>VTVFVALDY</td>
<td>EARTEDDL SF</td>
<td>HKGEKFQILN</td>
<td>SSEGDAEWR</td>
<td>SLTTGETQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>src_avis2</td>
<td>VTVFVALDY</td>
<td>ESRTETDL SF</td>
<td>KKGFERQIIVN</td>
<td>NTEGDWLAH</td>
<td>SLTTGQTQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>src_avissa</td>
<td>VTVFVALDY</td>
<td>ESRTETDL SF</td>
<td>KKGFERQIIVN</td>
<td>NTEGDWLAH</td>
<td>SLTTGQTQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>src_chick</td>
<td>VTVFVALDY</td>
<td>ESRTETDL SF</td>
<td>KKGFERQIIVN</td>
<td>NTEGDWLAH</td>
<td>SLTTGQTQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>stk_hydat</td>
<td>VTVFVALDY</td>
<td>ESRTETDL SF</td>
<td>KKGFERQIIVN</td>
<td>NTEGDWLAH</td>
<td>SLTTGQTQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>stk_hydat</td>
<td>VTVFVALDY</td>
<td>ESRTETDL SF</td>
<td>KKGFERQIIVN</td>
<td>NTEGDWLAH</td>
<td>SLTTGQTQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>src_rsvpa</td>
<td>VTVFVALDY</td>
<td>ESRTETDL SF</td>
<td>KKGFERQIIVN</td>
<td>NTEGDWLAH</td>
<td>SLTTGQTQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>src_chick</td>
<td>VTVFVALDY</td>
<td>ESRTETDL SF</td>
<td>KKGFERQIIVN</td>
<td>NTEGDWLAH</td>
<td>SLTTGQTQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>yes_chick</td>
<td>VTVFVALDY</td>
<td>EARTEDDL SF</td>
<td>HKGEKFQILN</td>
<td>SSEGDAEWR</td>
<td>SLTTGETQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>src_rsvpa</td>
<td>VTVFVALDY</td>
<td>EARTEDDL SF</td>
<td>HKGEKFQILN</td>
<td>SSEGDAEWR</td>
<td>SLTTGETQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>src_chick</td>
<td>VTVFVALDY</td>
<td>EARTEDDL SF</td>
<td>HKGEKFQILN</td>
<td>SSEGDAEWR</td>
<td>SLTTGETQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>stk_hydat</td>
<td>VTVFVALDY</td>
<td>EARTEDDL SF</td>
<td>HKGEKFQILN</td>
<td>SSEGDAEWR</td>
<td>SLTTGETQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>src_rsvpa</td>
<td>VTVFVALDY</td>
<td>EARTEDDL SF</td>
<td>HKGEKFQILN</td>
<td>SSEGDAEWR</td>
<td>SLTTGETQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>src_chick</td>
<td>VTVFVALDY</td>
<td>EARTEDDL SF</td>
<td>HKGEKFQILN</td>
<td>SSEGDAEWR</td>
<td>SLTTGETQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>stk_hydat</td>
<td>VTVFVALDY</td>
<td>EARTEDDL SF</td>
<td>HKGEKFQILN</td>
<td>SSEGDAEWR</td>
<td>SLTTGETQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>stk_hydat</td>
<td>VTVFVALDY</td>
<td>EARTEDDL SF</td>
<td>HKGEKFQILN</td>
<td>SSEGDAEWR</td>
<td>SLTTGETQYI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**YDFHGVDGGEDSIKRG**

---

**PSI-BLAST** SF Altschul *1997 Nucl Acids Res* 25 3389-3402

© Burkhard Rost

[ROSTLAB. TUM]
Sequence-profile methods

- **PSI-BLAST**
  - fast, partial dynamic programming
    SF Altschul (1997) NAR 25:3389-3402

- **ClustalW/ClustalX**
  - slow, dynamic programming, for experts
    JD Thompson, DG Higgins, TJ Gibson (1994) NAR 22:4673-80

- **MaxHom**
  - relatively slow, dynamic programming, good first guess

- **SAM/HMMer**
  - slow, need preprocess, HMM (statistics), very accurate

- **T-Coffee**
  - much slower, requires preprocessing, Genetic Algorithm

- **SSEARCH/PSI-Search**
  - SSEARCH - similar to MaxHom with SW only
  - PSI-Search: iterated SSEARCH
Sequence-profile comparison

YDFHGVEQEDDISIKRG

1
fyn_human VTLFVALDY EARTEDLSF HKGEKFQILN SSEGDWAER SLTTGETYI
yrk_chick VTLFIALDY EARTEDLSF QKGEKFHIIN NTEGDWARAR SLSSGATYI
fgr_human VTLFIALDY EARTEDLSF TKGEKFHILN NTEGDWARAR SLSSGRTYI
yes_chick VTVFVALDY EARTEDLSF RKGERQIIVN NTEGDWARAR SIATGKTYI
src_avis2 VTVFVALDY ESRTETDLSF RKGERQIIVN NTEGDWARAR SLTTGQTYI
src_aviss VTVFVALDY ESRTETDLSF RKGERQIIVN NTEGDWARAR SLTTGQTYI
src_chick VTVFVALDY ESRTETDLSF RKGERQIIVN NTEGDWARAR SLTTGQTYI
stk_hydat VTVFVALDY EARIISDLFS RKGERQIIVN TADGDWWAR SLTITSEGFI
src_rsvpa ....... ESRIEDLSF KKRERQIVN NTEGDWARAR SLTTGQTYI
hck_human ...IVIVALY EAIESDMLE QKGDMQVLE ES.GEWWWAR SLATRKEQFYI
blk_mouse ...FVVALDY AAVNTRLQTV KGEKIQVLR .STGDWVAR SLVTGREGQYV
hck_mouse .TVIVALY EAIESDMLE QKGDMQVLE .EAGEWVAR SLATKKEQFYI
lyn_human ...IVIVALPY DHIPDLISF KGEKKMLE .EHGFWAK SLLTKKEQFI
lck_hum...LVIALHYS EPSHDGALF EGKTEIRILE QS.GEWWWQ SLTTGQEQFYI
ss81_yeast.... ALIPY DADDDseIF EQNEIQQVSD ..IEQWRVAR R.ANGETQFYI
abl_mouse ..LFWALYDF VAGSNLTLI TKGEKIRVGY YnNGWTEAQ ..TKNGQQQYV
abl1_human ..LFWALYDF VAGSNLTLI TKGEKIRVGY YnNGWTEAQ ..TKNGQQQV
src1_drome ..VVALYDF KSRDESDLFS MKGDRIEVID DTESDWRVY NLTRQEGLI
mysd_dicdi..... ALDF DAESMSELFS KEGDIITVLD QSSGDAEADL .KGRRKRV
yfj4_yeast.... VALSF AEGESDLPF RKGDVITIY KsQNDWVGWR V..NGRECIY
abl2_human ..LFWALYDF VAGSNLTLI TKGEKIRVGY YNNGQEGFYV RSKNG.QQW
tec_human ...EIVVAYDF QAEGHDRLR ERQEQILE KNVDHVKWR D.KYNGEQGY
abl1_caeel ..LFWALYDF VAGSNLTLI TKGEKIRVGY YNNGQEGFYV RSKNG.QQW
txk_human ..... ALDF LPRPCNLAL RAAEYLILE KYNPHKWVAR D.RLGRRGQY
yha2_yeastVVRLYALYL TTMEDLFS RKGDVITLME QYRDWNGA L...RGCMIF
abp1_sacex.... AEDDY EAGEDNELTF AENDKINIE FVDDWGLGE LETTQKQLF
Zones

Anything more fancy?
beyond seq-prof?
Profile-profile alignments
evolution of alignment methods

☐ pairwise
☐ multiple
☐ sequence-profile
☐ ?
Problems with profile-profile

- complicated: too many free parameters
Problems with profile-profile

- complicated: too many free parameters
- may be unwanted:

```
ATRLLTTAKKDGPCD
ATRLLTTAKKDGPCD
ATRLLTTAKKDGPCD
ATKLLCTAKKEGPKD
ATKLLCTAKKEGPKD
ATKLLGAKKEGGCD
ATWLLTTAKKVGPCD
ATWLLTTAKKVGPCD
```
“cheap” hack toward profile-profile
Dariusz Przybylski

Thanks for slides

Dariusz Przybylski
Broad Inst, Boston
profile-profile: CPU intensive

D Przybylski & B Rost 2008 *Bioinformatics* 24: 1987-93
| A  | R  | N  | D  | C  | Q  | E  | G  | H  | I  | L  | K  | M  | F  | P  | S  | T  | W  | Y  | V  |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1  | I  | -4 | -5 | -6 | -6 | -4 | -5 | -5 | -6 | -6 | 4   | 5   | -5 | -1 | -2 | -5 | -5 | -3 | -5 | -4 | 3   |
| 2  | S  | -2 | -4 | -4 | -4 | -5 | -3 | -2 | 1  | -5 | -6 | -6 | 0   | -5 | -6 | 7   | 1   | -1 | -6 | -6 | -5 |
| 3  | P  | -3 | -3 | -1 | -3 | -6 | -1 | -2 | -5 | 3   | -6 | -6 | 3   | -5 | -6 | 8   | 3   | -3 | -3 | -7 | -5 | -6 |
| 4  | I  | -3 | 6  | -4 | -1 | -5 | -2 | -2 | -4 | 1   | 3   | -3 | 0   | 0   | -3 | 0   | -2 | -1 | -6 | -4 | 0   |
| 5  | E  | -3 | 4  | -2 | 4  | -6 | 0   | 4   | -5 | -2 | -6 | -5 | 0   | -5 | -6 | 2   | -1 | -4 | -6 | -5 | -5 |
| 6  | T  | -3 | -4 | -3 | -2 | -5 | -3 | -2 | -1 | -5 | 4   | 1   | -5 | -1 | -3 | 3   | -2 | 1   | -6 | -2 | 3   |
| 7  | V  | -3 | -3 | -4 | -6 | 4  | -4 | -3 | -6 | 0   | 5   | -1 | -5 | -1 | 4   | -5 | -3 | -1 | -5 | -2 | 4   |
| 8  | P  | -4 | 3  | -4 | -1 | -6 | 0   | 4   | -5 | 0   | -4 | -4 | 3   | 1   | -6 | 2   | -1 | -1 | -6 | -5 | -2 |
| 9  | V  | -3 | -5 | -5 | -5 | -5 | -4 | -3 | -6 | 8   | 6   | 0   | -2 | -3 | -4 | -6 | -5 | -4 | -6 | -2 | 2   |
| 10 | K  | -3 | 1  | 1  | 3  | -4 | 0   | 2   | -3 | 3   | -5 | -5 | 3   | -4 | -4 | 3   | 0   | 0   | -6 | -2 | -4 |
| 11 | L  | -5 | -6 | -7 | -7 | -5 | -6 | -6 | -7 | -6 | 4   | 6   | -6 | 0   | -2 | -6 | -4 | -3 | 2   | -3 | 0   |
| 12 | K  | 0  | 2  | -3 | 0  | -3 | -1 | 0   | -3 | -4 | 1   | -1 | 5   | -2 | -4 | -3 | -3 | -1 | -6 | -4 | 0   |
| 13 | P  | -3 | -5 | 0  | 0  | -6 | -2 | 1   | -4 | -1 | -4 | -3 | -2 | -5 | -6 | 7   | -1 | 1   | -2 | -6 | -4 |
| 14 | G  | -3 | -3 | 0  | 4  | -6 | -3 | 1   | 5   | 0   | -6 | -4 | -3 | -5 | -5 | 0   | -2 | -1 | -6 | -6 | -5 |
| 15 | M  | 3  | -4 | 0  | -2 | 0  | -3 | -1 | -1 | -1 | -1 | -3 | -1 | 3   | 0   | -1 | 0   | 3   | -2 | -2 | 0   |

D Przybylski & B Rost 2008 *Bioinformatics* 24: 1987-93
Consensus-profile: middle ground
... a little more complicated: statistics

\[ E\text{-value} \sim K \cdot L_1 \cdot L_2 \cdot e^{-\lambda \cdot \text{score}} \]

open symbols: full consensus
full: partial

D Przybylski & B Rost 2008 Bioinformatics 24: 1987-93
Consensus-profile PSI-BLAST fast & good
Profile-profile
modern:
HHblits
HHblits - a tool from Munich/Tuebingen

HHblits - a tool from Munich/Tuebingen

• J Söding, M Remmert, AN Lupas (2006) HHsenser: exhaustive transitive profile search using HMM-HMM comparison. NAR 33:W244-8

initial development & distribution

Andrei Lupas/MPI Tuebingen

Johannes Söding
MPI/Goettingen
HHblits:

Fig. 1:
HHblits:

Fig. 2:
Sequence-profile methods

- **PSI-BLAST**
  - fast, partial dynamic programming
  - SF Altschul (1997) NAR 25:3389-3402

- **ClustalW/ClustalX**
  - slow, dynamic programming, for experts: all in family treated equal
  - JD Thompson, DG Higgins, TJ Gibson (1994) NAR 22:4673-80

- **MaxHom/SSEARCH/PSI-Search**
  - relatively slow, dynamic programming, good first guess: query is special

- **SAM/HMMer**
  - slow, need preprocess, HMM (statistics), very accurate

- **T-Coffee**
  - much slower, requires preprocessing, Genetic Algorithm

- **HHBlits**
  - SSEARCH - similar to MaxHom with SW only
  - PSI-Search: iterated SSEARCH
Zones

From 3D twilight to 3D midnight zone
PDB all-against-all

proteins of known 3D structure (PDB)
Databases biased: MUST remove bias!

proteins of known 3D structure (PDB)

sequence-unique subset

proteins of known 3D structure (PDB)
Hypothetical distribution of similar structures
FAKE DATA
Evolution into the Midnight zone

![Graph showing Number of structure pairs versus Percentage pairwise sequence identity. The graph has a peak at around 10% sequence identity and shows a decrease as the sequence identity increases, with a shaded area indicating a particular zone of interest.]
Midnight zone: real - random
Midnight zone: real - random

B Rost 1997 Folding & Design 2, S19-S24

AS Yang and B Honig 2000 J Mol Biol 301, 679-689
Protein structures evolved at random - almost

- average < 10%
  - -> most pairs have ‘random’ identity levels
- 3 - 4% anchor residues
- 4 billion years of evolution reached equilibrium
  - rate of creating new structures slower than drift towards mean
- averages for convergent and divergent evolution similar
  - convergent evolution may have been a major event
Zones

- Sequence - sequence
- Sequence - profile
- Profile - profile

Midnight Zone
Twilight Zone
Safe Zone

Sequences similar
Structures similar
use structural information to intrude into midnight zone?
Secondary structure
Proteins have local “regularities”

sequence
KAVIDEEMQQDSVECATQALEKYNIEKDIAYIKKEFDKYNPTWHCIVGRNFGSYVTHETKHFIFYLYGQVAILLFKS

secondary structure:
LEEEEEELLHHHHHHHHHHHHHHHHHHHLLLHHHHHHHHHHHHHHHHHLLLLEEEEEELLLLLLEEEEEELL

2-state relative solvent accessibility:
OOOOOOOOOB000B000B00BB00B0000000000BB00BB00000000BB00BB00000000BB00BB000000

© Dariusz Przybylski, Columbia New York
Two paths to fold recognition

<table>
<thead>
<tr>
<th>3D PDB</th>
<th>Fosfos Profile</th>
<th>1D Projection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Str 1</td>
<td>EEH</td>
<td>sec 1aap</td>
</tr>
<tr>
<td>Str 2</td>
<td>HEEH</td>
<td>acc 1tcp</td>
</tr>
<tr>
<td>Str 3</td>
<td>HEHH</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>EHHÉHE</td>
<td>1btr</td>
</tr>
<tr>
<td>Str n</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1D PHD

<table>
<thead>
<tr>
<th>PHD 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHD 2</td>
</tr>
<tr>
<td>PHD 3</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>PHD n</td>
</tr>
</tbody>
</table>
**TOPITS**

**Project known 3D structure onto 1D**

**Predict 1D structure from sequence**

- **Input:**
  - sequence
  - generate sequence alignment

- **Predict 1D structure**
  - input: sequence

- **Align predicted and known structure(s)**

- **Good match to one of the known structures?**
  - Yes
    - predict fold of matching structure
    - model 3D coordinates by homology
Generalized sequences and profiles

Amino acids: A,D,G,H,V,…
Secondary Structure: E,H,L
Solvent Accessibility: B,O
1D Structure States: EB,HB,LB,EO,HO,LO

sequence______ generalized sequence
AVDG AHB VHO DLO GLO

sequence profile (PSSM) → generalized profile (GPSSM)

<table>
<thead>
<tr>
<th>\</th>
<th>A</th>
<th>D</th>
<th>G</th>
<th>H</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>7</td>
<td>-1</td>
<td>-2</td>
<td>-5</td>
<td>...</td>
</tr>
<tr>
<td>V</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>-6</td>
<td>...</td>
</tr>
<tr>
<td>D</td>
<td>-2</td>
<td>8</td>
<td>3</td>
<td>-1</td>
<td>...</td>
</tr>
<tr>
<td>G</td>
<td>-1</td>
<td>3</td>
<td>6</td>
<td>-1</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>\</th>
<th>AHB</th>
<th>DLO</th>
<th>GLO</th>
<th>HEB</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>AHB</td>
<td>9</td>
<td>-1</td>
<td>-3</td>
<td>-7</td>
<td>...</td>
</tr>
<tr>
<td>VHO</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>-6</td>
<td>...</td>
</tr>
<tr>
<td>DLO</td>
<td>-1</td>
<td>9</td>
<td>5</td>
<td>-2</td>
<td>...</td>
</tr>
<tr>
<td>GLO</td>
<td>-2</td>
<td>4</td>
<td>8</td>
<td>-2</td>
<td>...</td>
</tr>
</tbody>
</table>

D Przybylski & B Rost 2004 J Mol Biol 341, 255-269
we have to establish first that similar structures have similar 1D strings (secondary structure & solvent accessibility)
Secondary structure conserved

D Przybylski & B Rost 2004 J Mol Biol 341, 255-269
Secondary structure conserved

D Przybylski & B Rost 2004 J Mol Biol 341, 255-269
Secondary structure conserved

D Przybylski & B Rost 2004 J Mol Biol 341, 255-269

© Dariusz Przybylski, Columbia New York
Secondary structure conserved

D Przybylski & B Rost 2004 J Mol Biol 341, 255-269
Thanks for slides
Fold recognition without folds: AGAPE

1D prediction errors correlate!

Fold recognition without better than with folds

D Przybylski & B Rost 2004 *J Mol Biol* 341, 255-269
Aligning Generalized Profiles
# Lecture plan (PP1 structure/comp biol)

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>01: 04/23</td>
<td>Tue: No lecture</td>
</tr>
<tr>
<td>02: 04/25</td>
<td>Thu: No lecture</td>
</tr>
<tr>
<td>03: 04/30</td>
<td>Tue: No lecture</td>
</tr>
<tr>
<td>04: 05/02</td>
<td>Thu: Intro 1: organization of lecture: intro into cells &amp; biology</td>
</tr>
<tr>
<td>05: 05/07</td>
<td>Tue: Intro 2: amino acids, protein structure (comparison), domains</td>
</tr>
<tr>
<td>06: 05/09</td>
<td>Thu: Alignment 1</td>
</tr>
<tr>
<td>07: 05/14</td>
<td>Tue: Alignment 2</td>
</tr>
<tr>
<td>08: 05/16</td>
<td>Thu: Comparative modeling &amp; exp structure determination &amp; secondary structure assignment</td>
</tr>
<tr>
<td>09: 05/21</td>
<td>Tue: SKIP: Student Representation (SVV)</td>
</tr>
<tr>
<td>10: 05/23</td>
<td>Thu: 1D: Secondary structure prediction 1</td>
</tr>
<tr>
<td>11: 05/28</td>
<td>Tue: 1D: Secondary structure prediction 2</td>
</tr>
<tr>
<td>12: 05/30</td>
<td>Thu: SKIP: Ascension Day</td>
</tr>
<tr>
<td>13: 06/04</td>
<td>Tue: 1D: Secondary structure prediction 3 - Deep Learning</td>
</tr>
<tr>
<td>14: 06/06</td>
<td>Thu: 1D: Secondary structure prediction 4</td>
</tr>
<tr>
<td>15: 06/11</td>
<td>Tue: SKIP: Whitsun</td>
</tr>
<tr>
<td>16: 06/13</td>
<td>Thu: No lecture (but exercises)</td>
</tr>
<tr>
<td>17: 06/18</td>
<td>Tue: 1D: Transmembrane structure prediction 1</td>
</tr>
<tr>
<td>18: 06/20</td>
<td>Thu: SKIP: Corpus Christi</td>
</tr>
<tr>
<td>19: 06/25</td>
<td>Tue: 1D: Transmembrane structure prediction 2 / Solvent accessibility prediction</td>
</tr>
<tr>
<td>20: 06/27</td>
<td>Thu: 1D: Disorder prediction; 2D prediction / 3D prediction</td>
</tr>
<tr>
<td>21: 07/02</td>
<td>Tue: 2D: contact prediction - Deep Learning</td>
</tr>
<tr>
<td>22: 07/04</td>
<td>Thu: Recap 1</td>
</tr>
<tr>
<td>23: 07/09</td>
<td>Tue: Recap 2</td>
</tr>
<tr>
<td>24: 07/11</td>
<td>Thu: TBA</td>
</tr>
<tr>
<td>25: 07/16</td>
<td>Tue: TBA</td>
</tr>
<tr>
<td>26: 07/18</td>
<td>Thu: TBA</td>
</tr>
<tr>
<td>27: 07/23</td>
<td>Tue: TBA</td>
</tr>
<tr>
<td>28: 07/25</td>
<td>Thu: TBA</td>
</tr>
</tbody>
</table>

**EXAM: 07/17 18:00-20:00**