Videos: YouTube / www.rostlab.org

THANKS:

Special lectures:
- 07/xx Predrag Radivojac - Indiana Univ.
- 06/xx Yana Bromberg - Rutgers Univ.

No lecture:
- 05/09 no lecture
- 05/15 Ascension day
- 05/23 Student assembly (SVV)
- 06/06 Whitsun holiday
- 06/15 Corpus Christi

LAST lecture: bef: Jul 11 after: Jul 28

Examen: WEDNESDAY(!!) July 12: 18:00-19:30 TBA
- Makeup: TBC: Oct 17 & 19, 2017 - lecture time

CONTACT: Lothar Richter richter@rostlab.org

Exercises:
teaching@rostlab.org

Dmitrij Nechaev

Jonas Reeb
Lothar Richter
Michael Bernhofer
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

<table>
<thead>
<tr>
<th></th>
<th>1D</th>
<th>2D</th>
<th>3D</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>PP P</td>
<td>128</td>
<td>110</td>
</tr>
<tr>
<td>Q</td>
<td>QQQY</td>
<td>175</td>
<td>97</td>
</tr>
<tr>
<td>I</td>
<td>FFPQVI</td>
<td>70</td>
<td>E</td>
</tr>
<tr>
<td>T</td>
<td>SSIVR</td>
<td>77</td>
<td>E</td>
</tr>
<tr>
<td>L</td>
<td>LLSTL</td>
<td>120</td>
<td>E</td>
</tr>
<tr>
<td>W</td>
<td>WWQED</td>
<td>238</td>
<td>E</td>
</tr>
<tr>
<td>Q</td>
<td>RKQAK</td>
<td>169</td>
<td>E</td>
</tr>
<tr>
<td>R</td>
<td>RRRFPQ</td>
<td>200</td>
<td>E</td>
</tr>
<tr>
<td>P</td>
<td>PPPPP</td>
<td>24</td>
<td>E</td>
</tr>
<tr>
<td>L</td>
<td>VVTKF</td>
<td>E</td>
<td>71</td>
</tr>
<tr>
<td>V</td>
<td>VVLII</td>
<td>E</td>
<td>14</td>
</tr>
<tr>
<td>T</td>
<td>TTKEK</td>
<td>E</td>
<td>74</td>
</tr>
<tr>
<td>I</td>
<td>AALIV</td>
<td>E</td>
<td>0</td>
</tr>
<tr>
<td>K</td>
<td>HYKKF</td>
<td>E</td>
<td>90</td>
</tr>
<tr>
<td>I</td>
<td>IILVI</td>
<td>E</td>
<td>4</td>
</tr>
<tr>
<td>G</td>
<td>EENGG</td>
<td>E</td>
<td>46</td>
</tr>
<tr>
<td>G</td>
<td>GGTTG</td>
<td>E</td>
<td>62</td>
</tr>
<tr>
<td>Q</td>
<td>QQKRR</td>
<td>E</td>
<td>68</td>
</tr>
<tr>
<td>L</td>
<td>PPLNW</td>
<td>E</td>
<td>118</td>
</tr>
<tr>
<td>K</td>
<td>VVFKV</td>
<td>E</td>
<td>31</td>
</tr>
<tr>
<td>E</td>
<td>EESK</td>
<td>E</td>
<td>124</td>
</tr>
<tr>
<td>A</td>
<td>VVGLG</td>
<td>E</td>
<td>1</td>
</tr>
<tr>
<td>L</td>
<td>LILLL</td>
<td>E</td>
<td>29</td>
</tr>
<tr>
<td>L</td>
<td>LLLVV</td>
<td>E</td>
<td>24</td>
</tr>
<tr>
<td>D</td>
<td>DDDDD</td>
<td>E</td>
<td>49</td>
</tr>
<tr>
<td>T</td>
<td>TTTTT</td>
<td>72</td>
<td>E</td>
</tr>
<tr>
<td>G</td>
<td>GGGGG</td>
<td>E</td>
<td>62</td>
</tr>
<tr>
<td>A</td>
<td>AAAAA</td>
<td>E</td>
<td>17</td>
</tr>
<tr>
<td>D</td>
<td>DDDDD</td>
<td>E</td>
<td>102</td>
</tr>
<tr>
<td>D</td>
<td>DDAKE</td>
<td>E</td>
<td>69</td>
</tr>
<tr>
<td>T</td>
<td>SSSTTV</td>
<td>E</td>
<td>1</td>
</tr>
<tr>
<td>V</td>
<td>IIVIV</td>
<td>E</td>
<td>14</td>
</tr>
<tr>
<td>L</td>
<td>VVIVL</td>
<td>E</td>
<td>0</td>
</tr>
</tbody>
</table>
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

| J | GGQLAXEEAL |
| T | ECQPVEVL |

Optimal alignment (no gaps)

| J | GGQLAXEEAL |
| T1 | EVL |
| T2 | EGQPVEVL |

Optimal alignment (with gaps)

| J | GGQLAXEEAL |
| T | EGQPVEVL |

- 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

\[
SW = \sum_{k=1}^{L_\text{ali}} M_{U_k T_k} - G_0 \cdot N_{\text{gap}} - G_e \cdot (L_{\text{gap}} - N_{\text{gap}})
\]
Alignments: scoring matrix

synonyms:

- [scoring substitution exchange log-odds]
- [matrix metric table]

one particular: Blossum65

BLAST: fast matching of single ‘words’

Default “word” size for “seeds” = 3
the major challenge for word search algorithms is to get the statistics right
Significance of match (e.g. BLAST E-values)

![Graph showing the distribution of scores with a peak at score 15, indicating hits and a background distribution.](image-url)
Twilight zone = false positives bazoom!!

![Image of twilight zone]

**Graph:**
- **Y-axis:** Number of protein pairs
- **X-axis:** Distance from HSSP threshold
- **Legend:**
  - Red line: Percentage sequence identity
  - Blue crosses: Number of residues aligned

**Equations:**
- Sequence identity:
  - 0 to 100
- Residues aligned:
  - 0 to 400

**References:**
B Rost 1999 *Prot Engin* 12, 85-94
Sequence comparisons: multiple alignment/profile-based
Multiple alignments problems & hacks
Multiple alignments

- Dynamic programming?
  for 3 sequences: $O(N_1 \times N_2 \times N_3)$
  NP-complete (L Wang & T Jiang (1994) JCB 1: 337-48)
- claim: computer: up to 6
  ~60 TB main memory
  no quote -> unsure
Multiple alignments

- Dynamic programming? for 3 sequences: $O(N_1 \times N_2 \times N_3)$
  NP-complete (L Wang & T Jiang (1994) JCB 1: 337-48)

- hack 1:
  dynamic programming: pairwise, only space in vicinity of intersection searched n-wise
Multiple alignments

- Dynamic programming?
  for 3 sequences: $O(N_1 \times N_2 \times N_3)$
  NP-complete (L Wang & T Jiang (1994) JCB 1: 337-48)

- hack 1:
  dynamic programming: pairwise, only space in vicinity of intersection searched n-wise

- hack 2:
  map to tree / pairwise
  Russell Doolittle, UCSD

Shapers and Shakers

© Burkhard Rost
Multiple alignment: progressive 1

<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>90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

GGQLAKEEAL
GGQLAKDEAL
GGQIAKDEAL
GGQIAKDEAI
Multiple alignment: progressive

<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>90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Step 1**

GGQLAKEEAL
GGQLAKDEAL
GGQIAKDEAI
ggqlakeeal
Multiple alignment: progressive

Step 1
GGQLAKEEAL
GGQLAKDEAL
gqqlakeeal

Step 2
GGQIAKDEAL
GGQIAKDEAI
gqqiakdeal
Multiple alignment: progressive 1

<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>90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Step 1

GGQLAKEEAL
GGQLAKDEAL
ggqlakeeal

Step 2

GGQIAKDEAL
GGQIAKDEAI
gqiqakdeal

Step 3

GGQIAKDEAL
GGQIAKDEAI
ggqiakdeal

Step 1

Step 2

Step 3
# Multiple alignment: progressive 2

<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>90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Step 1**

GGQLAKEEAL
GGQLAKDEAL
GGQIAKDEAI
ggqlakeeal
## Multiple alignment: progressive 2

<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>90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Step 1**
- GGQLAKEEAL
- GGQLAKDEAL
- ggqlakeeal

**Step 2**
- GGQLAKEEAL
- GGQLAKDEAL
- GGQIAKDEAL
- GGQIAKDEAI

**Step 1**
- GGQLAKEEAL
- GGQLAKDEAL
- ggqlakeeal

**Step 2**
- ggqlakeeal
- GGQIAKDEAL
- ggqlakeeal
Zones

## Multiple alignment: progressive 2

<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>

**Step 1**
- GGQLAKEEAL
- GGQLAKDEAL
- ggqlakeeal

**Step 2**
- ggqlakeeal
- GGQIAKDEAL
- ggqlakeeal

**Step 3**
- ggqlakeeal
- GGQIAKDEAI

The image illustrates a multiple alignment process using progressive alignment methods. Each step involves comparing the sequences and aligning them progressively. The example shows how different sequences are aligned in stages, with each step refining the alignment until the final sequence is obtained.

© Burkhard Rost

ROSTLAB. M
Profiles/PSSM/motifs
Computationally: motifs

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

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.
- GLLMSA CVVV
- GI LMSA YPP
- GLLMSAES

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

Building a PSSM:
- Absolute frequencies
- Add pseudo-counts if necessary
- Relative frequencies
- Log likelihoods

Starting point:
- 123456
- ATGCTA
- ATTGCT
- TCTGAG
- GTTGAG
- CCATCC
PSSM - Position specific substitution matrix: one solution

\[ \begin{array}{ccccccc}
1 & 2 & 3 & 4 & 5 & 6 \\
A & 0.4 & 0 & 0.2 & 0 & 0.4 & 0.2 \\
T & 0.2 & 0.6 & 0.6 & 0.2 & 0.2 & 0.2 \\
G & 0.2 & 0.2 & 0.6 & 0 & 0.4 \\
C & 0.2 & 0.4 & 0 & 0.2 & 0.4 & 0.2 \\
\end{array} \]

\[ \begin{array}{ccccccc}
1 & 2 & 3 & 4 & 5 & 6 \\
A & 0.47 & \infty & -0.22 & \infty & 0.47 & -0.22 \\
T & -0.22 & 0.86 & 0.96 & -0.22 & -0.22 & -0.22 \\
G & -0.22 & \infty & -0.22 & 0.86 & \infty & 0.47 \\
C & -0.22 & 0.47 & \infty & -0.22 & 0.47 & -0.22 \\
\end{array} \]

\[ \text{Log odds} = \ln(S \div 0.25) \]
Recap: substitution matrix (BLOSUMUM)

|   | C | S | T | P | A | G | N | D | E | Q | H | R | K | M | L | V | F | Y | W |
| C | 12|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 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 | 0 | 2 |   |   |   |   |   |   |   |   |   |   |   |   |   |
| D | -5| 0 | 0 | -1| 0 | 1 | 2 | 4 |   |   |   |   |   |   |   |   |   |   |   |   |
| E | -5| 0 | 0 | -1| 0 | 0 | 1 | 3 | 4 |   |   |   |   |   |   |   |   |   |   |   |
| Q | -5| -1| -1| 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| 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 |   |   |   |   |   |   |
| I | -2| -1| 0 | -2| -1| -3| -2| -2| -2| -2| -2| -2| -2| 2 | 5 |   |   |   |   |   |
| L | -6| -3| -2| -3| -2| -4| -3| -4| -3| -2| -2| -3| -3| 4 | 2 | 6 |   |   |   |   |
| V | -2| -1| 0 | -1| 0 | -1| -2| -2| -2| -2| -2| -2| -2| 2 | 4 | 2 | 4 |   |   |
| F | -4| -3| -3| -5| -4| -5| -2| -6| -5| -5| -2| -4| -5| 0 | 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 | -8| -2| -5| -6| -6| -7| -4| -7| -7| -5| -3| 2 | -3| -4| -5| -2| -6| 0 | 0 | 17 |

### PSSM - Position specific substitution matrix: example

<table>
<thead>
<tr>
<th>POS</th>
<th>PROBE</th>
<th>CONSENSUS</th>
<th>PROFILE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>1</td>
<td>E</td>
<td>-2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>L</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>3</td>
<td>V</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>K</td>
<td>6</td>
<td>-2</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>6</td>
<td>-1</td>
</tr>
<tr>
<td>6</td>
<td>G</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>S</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>S</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>V</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>R</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>11</td>
<td>M</td>
<td>0</td>
<td>-2</td>
</tr>
<tr>
<td>12</td>
<td>S</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>13</td>
<td>C</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>14</td>
<td>K</td>
<td>1</td>
<td>-2</td>
</tr>
<tr>
<td>15</td>
<td>A</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>16</td>
<td>T</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>17</td>
<td>G</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>Y</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>19</td>
<td>I</td>
<td>1</td>
<td>-2</td>
</tr>
<tr>
<td>20</td>
<td>F</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>21</td>
<td>S</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>22</td>
<td>S</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>23</td>
<td>G</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>D</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>25</td>
<td>G</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>26</td>
<td>A</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>27</td>
<td>Y</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>28</td>
<td>E</td>
<td>2</td>
<td>-2</td>
</tr>
<tr>
<td>29</td>
<td>L</td>
<td>3</td>
<td>-5</td>
</tr>
<tr>
<td>30</td>
<td>N</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

---

Profiles profit from relation of “families”
what is the difference: profile vs motif
PSI-Blast
basic idea
**Profile-based comparison**

<table>
<thead>
<tr>
<th>1</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>fyn_human</td>
<td>VTLFVLYDY EARTEDDLSF HKGEKFQILN SSEGDWEAR SLTTGETQYI</td>
</tr>
<tr>
<td>yrk_chick</td>
<td>VTLFIALYD Y EARTEDDLSF QKGEKFHIIN NTEGDWEAR SLSSGATQYI</td>
</tr>
<tr>
<td>fgr_human</td>
<td>VTLFIALYD EARTEDDLTF TKGEKFHILN NTEGDWEAR SLSSGKTQYI</td>
</tr>
<tr>
<td>yes_chick</td>
<td>VTVFVALYD EARTDDLSF KKGERFQIIN NTEGDWEAR SIATGKTQYI</td>
</tr>
<tr>
<td>src_avis2</td>
<td>VTTFVALYD ESRTETDLK KKKERIQLIVN NTEGDWLAH SLTTGQTQYI</td>
</tr>
<tr>
<td>src_avisr</td>
<td>VTTFVALYD ESRTETDLK KKKERIQLIVN NTEGDWLAH SLTTGQTQYI</td>
</tr>
<tr>
<td>src_chick</td>
<td>VTTFVALYD ESRTETDLK KKKERIQLIVN NTEGDWLAH SLTTGQTQYI</td>
</tr>
<tr>
<td>stk_hydat</td>
<td>VTIIVALYD EARISEDLSF KKKERIQLIN TADGDWWAR SLTNSEGQYI</td>
</tr>
<tr>
<td>src_rsvpa</td>
<td>EAIHEDLSF QKGDQMVLK EAGWVRAR SLATKKEQYI</td>
</tr>
<tr>
<td>hck_hydat</td>
<td>EAIHEDLSF QKGDQMVLK EAGWVRAR SLATKKEQYI</td>
</tr>
<tr>
<td>blk_mouse</td>
<td>FVVALFDY AAVNDRDLQV LKEKIQVLR STGDWLR SLVTGREGYV</td>
</tr>
<tr>
<td>hck_mouse</td>
<td>EAIHEDLSF QKGDQMVLK EAGWVRAR SLATKKEQYI</td>
</tr>
<tr>
<td>lck_human</td>
<td>EAIHEDLSF QKGDQMVLK EAGWVRAR SLATKKEQYI</td>
</tr>
<tr>
<td>ss81_yeast</td>
<td>ALPYDADDGDIFS EUNHISQVSD IEGRWKRAR R.ANGTQI</td>
</tr>
<tr>
<td>abl1_mouse</td>
<td>LFVALYDF VASGDNTLSI TKGEKIRVLG YnnGEWCAQ ..TKNGQGWV</td>
</tr>
<tr>
<td>abl1_human</td>
<td>LFVALYDF VASGDNTLSI TKGEKIRVLG YnnGEWCAQ ..TKNGQGWV</td>
</tr>
<tr>
<td>src1_drome</td>
<td>VVVALYD KSRDESDLF MKGDRMEVID DTESDWVRV VNLTRQEGL</td>
</tr>
<tr>
<td>mysd_dicdi</td>
<td>ALYDF DAESSTNSLF KEGDITVLD QSSGDNWDAE L..KGRKRV</td>
</tr>
<tr>
<td>yfj4_yeast</td>
<td>ALYSF AGESGDLDF RKGDVITILK ksQNDWCR V..NGREGY</td>
</tr>
<tr>
<td>abl2_human</td>
<td>LFVALYDF VASGDNTLSI TKGEKIRVLG YnnGEWCAQ ..TKNGQGWV</td>
</tr>
<tr>
<td>tec_human</td>
<td>EIVVAMEDF QAAEHDRLR ERGQEYIL KNDVHWAR D.KYQNEYI</td>
</tr>
<tr>
<td>abl1_caeel</td>
<td>LFVALYDF HGVEEQSL RKKGDQRVLG YNNKEWCEA RlrlGEIQW</td>
</tr>
<tr>
<td>txk_human</td>
<td>LFVALYDF VASGDNTLSI TKGEKIRVLG YnnGEWCAQ ..TKNGQGWV</td>
</tr>
<tr>
<td>yha2_yeast</td>
<td>VTVALYD VTVIAMTS NFLNTPEWLSF RKKGDQVRVLG YNNKEWCEA RlrlGEIQW</td>
</tr>
<tr>
<td>yfj4_yeast</td>
<td>ALYSF AGESGDLDF RKGDVITILK ksQNDWCR V..NGREGY</td>
</tr>
<tr>
<td>abp1_sacex</td>
<td>AEYDY EAGEDNELTF AENDKIINIE FVDDDEWAG LETTGQKGLF</td>
</tr>
</tbody>
</table>
Idea: replace generic scoring matrix ...

Generic scoring matrix  
(here BLOSUM62)

<table>
<thead>
<tr>
<th>C</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>G</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Idea: replace generic by specific scoring

**Generic scoring matrix**

(here BLOSUM62)

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

Stephen F Altschul

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

NCBI: National Center for Biotechnology Information
NIH: National Institutes of Health, USA
ISCB: International Society for Computational Biology

David J Lipman
NCBI Bethesda USA

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

- PSI-BLAST
  - fast, partial dynamic programming
  - Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. NAR 25:3389-3402

- ~50,000 citations in Google Scholar May 2014

Stephen Altschul
David Lipman
Web Miller
Tom Madden
Alejandro Schäffer
PSI-BLAST
sketch of solution
PSI-BLAST in steps

1. fast hashing
Like BLAST match ‘words’

TTYKLILNGKTLKGETTTEAVDAATAEKFVKQYANDNGVDGEWTYDDATKTFVTVEK
TTYKLILLLLLLLLLLLLLLLLAWTVEKAFKTFFAAAAAAAAWTVEKAFKTFAA

---

#1 seed=3

---

TTYKLIL
TTYKLIL

#2 extend

??

AATAEKFVKQYA
AWTVEKAFKTF

TTYDDATKTF
WTVEKAFKTF

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

TTYKLILNGKTLKGETTTEAVDAATAEKFQKQANDNGVDGEWTYDDATKTFTVTEK
TTYKLILLLLLLLLLLLLLLLLLLLLAWTVEKAFKTFAAAAAAAAAWTVEKAFKTF

#1 seed=3

#2 extend

TTYKLIL
TTYKLIL

TTYKLIL

AATAEKFQKQYA
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 $> \text{threshold } T$

e.g. $T=13$ \rightarrow neighborhood

YES-YES

YES-YEA

YES-YEN

\begin{equation}
S_{ij} = \left( \frac{1}{\lambda} \right) \log \left( \frac{p_{ij}}{q_i \times q_j} \right)
\end{equation}
Btw. Blosum 62


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


**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>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>
<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>-1</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>
<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>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>
</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>6</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>5</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>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>
</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>-2</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-2</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>-1</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Tyr</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>1</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>-3</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>1</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>2</td>
<td>-3</td>
<td>7</td>
</tr>
</tbody>
</table>

PSI-BLAST in steps

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

TTYKLILNGKTLKGETTTTEAVDAATAEKFQKQYANDNGVDGEWTYDDATKTFTVTEKTYYKLILLLLLLLLLLLLLLLLLLLAWTVEKAFKTFAAAAAAAAWTVEKAFKTFAAAAAA

#1 seed=3

#2 extend

dynamic programming to extend
BLAST + Smith-Waterman

TTYKLIILNGKTLKGETTTEAVDAATAEKFVKQYANDNGVDGEWTYDDATKTFTVTEK
TTYKLIILLLLLLLLLLLLLLLLLLLLLLAWTVEKAFKTFAAAATAAWTVEKAFKTFAAA

---

#1 seed=3

#2 extend

TTYKLIIL
TTYKLIIL

AATAEKFVKQYA
AWTVEKAFKTF

WTYDDATKTTF
WTVEKAFKTF

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
Sequence-profile comparison

YDFHGVGEDDISIKRGG

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

[Diagram of sequence-profile comparison]
PSI-BLAST in steps

1. fast hashing
2. dynamic programming extension between matches
3. compile statistics
4. collect all pairs and build profile
5. ?

YDFHGVGEDDISIKRG
PSI-BLAST in steps

1. fast hashing
2. dynamic programming extension between matches
3. compile statistics
4. collect all pairs and build profile
5. iterate
Sequence-profile comparison

YDFHGTVGEDDIESIKRG

**PSI-position specific iteration**

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

© Burkhard Rost

ROSTLAB.UMI
Expanding in sequence space: dynamics of PSI-BLAST
Profile-based database search
Profile-based database search
Profile-based database search

Family U

safe for pairwise

zone reached through position-specific family profile

safe zone
Profile-based database search

Family U

zone reached through position-specific family profile

safe for pairwise

lost after iteration
Profile-based database search

- Family U
- Safe zones of close homologues reached through position-specific family profile
- Safe zones of close homologues
- Lost after iteration
- Safe for pairwise
Profile-based database search
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
INSERT: reproduce phylogeny?
Reproduce phylogeny
Reproduce phylogeny
Reproduce phylogeny
Reproduce phylogeny
Reproduce phylogeny
Reproduce phylogeny

story 2
Reproduce phylogeny

story 2
Reproduce phylogeny

story 2
Reproduce phylogeny: reality

story 1

story 3

story 2
Reproduce phylogeny: our view
Reproduce phylogeny: our view
Reproduce phylogeny: our view

Truth

Fiction
Clustal (ClustalW, ClustalX)

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

<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>90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

Fig. 1: F Sievers et al & DG Higgins (2011) Mol Syst Biol:7:539
# Clustal (ClustalW, ClustalX)

<table>
<thead>
<tr>
<th>Date</th>
<th>GoogleScholar</th>
<th>May 2014</th>
</tr>
</thead>
</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)
Toby Gibson

- 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)

Desmond G 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
MaxHom
Maxhom/HSSP

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>90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sweep 1

-> Profile (P0)
conservation weight (cw0)

Sweep 2

© Burkhard Rost
ROSTLAB. TUM
**Philosophy: family vs. protein**

<table>
<thead>
<tr>
<th>Protein-centric</th>
<th>Family-centric</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(MaxHom/PSI-BLAST)</strong></td>
<td><strong>(Clustal/HMM/T-coffee)</strong></td>
</tr>
<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 RPDQRAKHRGF</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.

Shapers and Shakers

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
**ProfileHMM: example football**

<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>Borussia Dort.</td>
<td>30,660</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Aug 17</td>
<td>FT</td>
<td>Werder Bremen</td>
<td>L</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>W</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>L</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>W</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>W</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>W</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>D</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>L</td>
<td>VfL Wolfsburg</td>
<td>27,554</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Oct 26</td>
<td>FT</td>
<td>Bayer Leverkusen</td>
<td>L</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>L</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>W</td>
<td>FC Augsburg</td>
<td>71,000</td>
<td>Bundesliga</td>
</tr>
</tbody>
</table>
ProfileHMM: example football

Probabilistic
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>L</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>L</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>W</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>L</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>W</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>W</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>L</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>D</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>L</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>L</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>L</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>W</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>Borussia Dort.</td>
<td>30,660</td>
<td>Bundesliga</td>
</tr>
<tr>
<td>Aug 17</td>
<td>FT</td>
<td>Werder Bremen</td>
<td>W L W</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>W</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>W</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>W</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>W</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>L</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>D</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>D</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>D D</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>L</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>W W</td>
<td>FC Augsburg</td>
<td>71,000</td>
<td>Bundesliga</td>
</tr>
</tbody>
</table>
ProfileHMM: example football

![Diagram](attachment:image.png)
ProfileHMM: formalism

HMMs are probabilistic models defined by:

- A finite set $S$ of states
- A discrete alphabet $A$ of symbols (observed objects)
- A probability transition matrix $T=(t_{ij})$, $i,j$ states
- A probability emission matrix $E=(e_{ix})$, $i$ state, $x$ symbol
ProfileHMM - intro

Profile-HMMs are probabilistic models…

Model  -> simulates a system
Probabilistic  -> produces outcomes based on probabilities
ProfileHMM: example 2: traffic light
ProfileHMM: example 2: traffic light
ProfileHMM: example 2: traffic light

Deterministic

R → Y → G
ProfileHMM: example 3: weather
ProfileHMM: example 3: weather
ProfileHMM: example 3: weather

*In fact, chaotic, deterministic*
ProfileHMM: example 3: weather

\[ p_{ij} \]  Transition probability from symbol i to symbol j

*In fact, chaotic, deterministic*
Example 2: The weather

Probabilistic, 1\textsuperscript{st} order Markov model

\[ P(X_{n+1} = x | X_1 = x_1, \ldots, X_n = x_n) = P(X_{n+1} = x | X_n = x_n) \]
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
Thanks for slides

Following slides cut out from an ISMB Tutorial given by Kevin Karplus 1999 in Heidelberg

http://www.sccrtc.org/photos/awards/karplus00-2001.jpg

http://users.soe.ucsc.edu/~karplus/bike/karplus_recumbent.gif
SAM-T98: Build alignment

SAM-T98 Alignment Building

Start: a single sequence

Build a model from the sequence or alignment
Background: entropy/information

Entropy:

\[ P(x) : \text{probability of observing } A \]
\[ \log (\text{# microstates in a system}) \]

\[ H(x) = - \sum x(i) \{ P(x(i)) \log P(x(i)) \} \]

- minimal for PEAK distribution
- maximal for uniform distribution
Entropy in alignment

- consider residue position i
  BEFORE any amino acid is aligned,

2crd    XFTNVSCCTTSKECWSVCQRLHNTSRGKCMNKKCRCYS
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
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, $P_0$, with entropy $H_0$
- now consider same column AFTER alignment
  \textit{posterior} probability $P_i + \text{priors} \rightarrow H_i$

<table>
<thead>
<tr>
<th>2crd</th>
<th>XFTNVSCCTTSKECWSVCQRLHNTSRGKCMNKKCRCYS</th>
</tr>
</thead>
<tbody>
<tr>
<td>bits saved</td>
<td>C</td>
</tr>
</tbody>
</table>

© Bur \hspace{1cm} © Kevin Karplus UCSC
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
\textit{posterior} probability Pi + priors -> Hi

☐ if position i conserved: Hi =? 

\begin{verbatim}
2crd  XFTNVSTTSKECWSVCQRLHNTSRGKCMNKKCRCYS
\end{verbatim}
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 position i conserved: Hi -> 0
  if position i completely varied: Hi=?

2crd  XFTNVSCCTTSKECWSVCQRLHNTSRGKCMNKKCRCYS

bits saved
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, \( P_0 \), with entropy \( H_0 \)
- now consider same column after alignment posterior probability \( P_i \) + priors \( \rightarrow H_i \)
- if position \( i \) conserved: \( H_i \rightarrow 0 \)
- if position \( i \) completely varied: \( H_i \rightarrow H_0 \)

2crd

XFTNVSCTTSKECWSVCQRLHNTSRGKCMNKKCRCYS

bits saved
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

bits saved
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  XFTNVSCCTTSKECWCSVQRLHNTSRGKCMNKKCRCYS
SAM-T98: Build alignment

SAM-T98 Alignment Building

- Start: a single sequence
- Build a model from the sequence or alignment
- Use the model to search for additional homologs
- Reestimate the alignment with the new homologs

(Iterations 1 - 3)

(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

SAM

HMMer

sequence-sequence or sequence-profile

<table>
<thead>
<tr>
<th>src_hum</th>
<th>src_aven</th>
<th>src_avis</th>
<th>src_avis2</th>
<th>src_avisr</th>
<th>src_chick</th>
<th>src_pon</th>
<th>stk_hydat</th>
<th>src_rsvpa</th>
</tr>
</thead>
<tbody>
<tr>
<td>VTLFV</td>
<td>VTLFI</td>
<td>VTVFV</td>
<td>VTVFV</td>
<td>VTVFV</td>
<td>VTVFV</td>
<td>VTVFV</td>
<td>VTIFV</td>
<td>VTTFV</td>
</tr>
<tr>
<td>ALYDY</td>
<td>ALYDY</td>
<td>ALYDY</td>
<td>ALYDY</td>
<td>ALYDY</td>
<td>ALYDY</td>
<td>ALYDY</td>
<td>ALYDY</td>
<td>ALYDY</td>
</tr>
<tr>
<td>EARTE</td>
<td>EARTE</td>
<td>EARTT</td>
<td>EARTT</td>
<td>EARTT</td>
<td>EARTT</td>
<td>EARTT</td>
<td>EARTT</td>
<td>EARTT</td>
</tr>
<tr>
<td>DDLSF</td>
<td>DDLSF</td>
<td>DDLSF</td>
<td>DDLSF</td>
<td>DDLSF</td>
<td>DDLSF</td>
<td>DDLSF</td>
<td>DDLSF</td>
<td>DDLSF</td>
</tr>
<tr>
<td>HKGEKF</td>
<td>HKGEKF</td>
<td>HKGEKF</td>
<td>HKGEKF</td>
<td>HKGEKF</td>
<td>HKGEKF</td>
<td>HKGEKF</td>
<td>HKGEKF</td>
<td>HKGEKF</td>
</tr>
<tr>
<td>QILN</td>
<td>HIIN</td>
<td>HIIN</td>
<td>HIIN</td>
<td>HIIN</td>
<td>HIIN</td>
<td>HIIN</td>
<td>HIIN</td>
<td>HIIN</td>
</tr>
<tr>
<td>SSEGDWW</td>
<td>NTEGDWW</td>
<td>NTEGDWW</td>
<td>NTEGDWW</td>
<td>NTEGDWW</td>
<td>NTEGDWW</td>
<td>NTEGDWW</td>
<td>NTEGDWW</td>
<td>NTEGDWW</td>
</tr>
<tr>
<td>EAR</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
</tr>
<tr>
<td>SLTTGETG</td>
<td>SLSSGATG</td>
<td>SLSSGATG</td>
<td>SLSSGATG</td>
<td>SLSSGATG</td>
<td>SLSSGATG</td>
<td>SLSSGATG</td>
<td>SLSSGATG</td>
<td>SLSSGATG</td>
</tr>
<tr>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>src_hum</th>
<th>src_aven</th>
<th>src_avis</th>
<th>src_avis2</th>
<th>src_avisr</th>
<th>src_chick</th>
<th>src_pon</th>
<th>stk_hydat</th>
<th>src_rsvpa</th>
</tr>
</thead>
<tbody>
<tr>
<td>VTLFV</td>
<td>VTLFI</td>
<td>VTVFV</td>
<td>VTVFV</td>
<td>VTVFV</td>
<td>VTVFV</td>
<td>VTVFV</td>
<td>VTIFV</td>
<td>VTTFV</td>
</tr>
<tr>
<td>ALYDY</td>
<td>ALYDY</td>
<td>ALYDY</td>
<td>ALYDY</td>
<td>ALYDY</td>
<td>ALYDY</td>
<td>ALYDY</td>
<td>ALYDY</td>
<td>ALYDY</td>
</tr>
<tr>
<td>EARTE</td>
<td>EARTE</td>
<td>EARTT</td>
<td>EARTT</td>
<td>EARTT</td>
<td>EARTT</td>
<td>EARTT</td>
<td>EARTT</td>
<td>EARTT</td>
</tr>
<tr>
<td>DDLSF</td>
<td>DDLSF</td>
<td>DDLSF</td>
<td>DDLSF</td>
<td>DDLSF</td>
<td>DDLSF</td>
<td>DDLSF</td>
<td>DDLSF</td>
<td>DDLSF</td>
</tr>
<tr>
<td>HKGEKF</td>
<td>HKGEKF</td>
<td>HKGEKF</td>
<td>HKGEKF</td>
<td>HKGEKF</td>
<td>HKGEKF</td>
<td>HKGEKF</td>
<td>HKGEKF</td>
<td>HKGEKF</td>
</tr>
<tr>
<td>QILN</td>
<td>HIIN</td>
<td>HIIN</td>
<td>HIIN</td>
<td>HIIN</td>
<td>HIIN</td>
<td>HIIN</td>
<td>HIIN</td>
<td>HIIN</td>
</tr>
<tr>
<td>SSEGDWW</td>
<td>NTEGDWW</td>
<td>NTEGDWW</td>
<td>NTEGDWW</td>
<td>NTEGDWW</td>
<td>NTEGDWW</td>
<td>NTEGDWW</td>
<td>NTEGDWW</td>
<td>NTEGDWW</td>
</tr>
<tr>
<td>EAR</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
</tr>
<tr>
<td>SLTTGETG</td>
<td>SLSSGATG</td>
<td>SLSSGATG</td>
<td>SLSSGATG</td>
<td>SLSSGATG</td>
<td>SLSSGATG</td>
<td>SLSSGATG</td>
<td>SLSSGATG</td>
<td>SLSSGATG</td>
</tr>
<tr>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>src_hum</th>
<th>src_aven</th>
<th>src_avis</th>
<th>src_avis2</th>
<th>src_avisr</th>
<th>src_chick</th>
<th>src_pon</th>
<th>stk_hydat</th>
<th>src_rsvpa</th>
</tr>
</thead>
<tbody>
<tr>
<td>VTLFV</td>
<td>VTLFI</td>
<td>VTVFV</td>
<td>VTVFV</td>
<td>VTVFV</td>
<td>VTVFV</td>
<td>VTVFV</td>
<td>VTIFV</td>
<td>VTTFV</td>
</tr>
<tr>
<td>ALYDY</td>
<td>ALYDY</td>
<td>ALYDY</td>
<td>ALYDY</td>
<td>ALYDY</td>
<td>ALYDY</td>
<td>ALYDY</td>
<td>ALYDY</td>
<td>ALYDY</td>
</tr>
<tr>
<td>EARTE</td>
<td>EARTE</td>
<td>EARTT</td>
<td>EARTT</td>
<td>EARTT</td>
<td>EARTT</td>
<td>EARTT</td>
<td>EARTT</td>
<td>EARTT</td>
</tr>
<tr>
<td>DDLSF</td>
<td>DDLSF</td>
<td>DDLSF</td>
<td>DDLSF</td>
<td>DDLSF</td>
<td>DDLSF</td>
<td>DDLSF</td>
<td>DDLSF</td>
<td>DDLSF</td>
</tr>
<tr>
<td>HKGEKF</td>
<td>HKGEKF</td>
<td>HKGEKF</td>
<td>HKGEKF</td>
<td>HKGEKF</td>
<td>HKGEKF</td>
<td>HKGEKF</td>
<td>HKGEKF</td>
<td>HKGEKF</td>
</tr>
<tr>
<td>QILN</td>
<td>HIIN</td>
<td>HIIN</td>
<td>HIIN</td>
<td>HIIN</td>
<td>HIIN</td>
<td>HIIN</td>
<td>HIIN</td>
<td>HIIN</td>
</tr>
<tr>
<td>SSEGDWW</td>
<td>NTEGDWW</td>
<td>NTEGDWW</td>
<td>NTEGDWW</td>
<td>NTEGDWW</td>
<td>NTEGDWW</td>
<td>NTEGDWW</td>
<td>NTEGDWW</td>
<td>NTEGDWW</td>
</tr>
<tr>
<td>EAR</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
</tr>
<tr>
<td>SLTTGETG</td>
<td>SLSSGATG</td>
<td>SLSSGATG</td>
<td>SLSSGATG</td>
<td>SLSSGATG</td>
<td>SLSSGATG</td>
<td>SLSSGATG</td>
<td>SLSSGATG</td>
<td>SLSSGATG</td>
</tr>
<tr>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
<td>YI</td>
</tr>
</tbody>
</table>

© Burkhard Rost
ROSTLAB.TUM
Independence assumption

dynamic programming
(smith-waterman)

PSI-BLAST

sequence-sequence

sequence-profile

HMMer

SAM

ALL assume that alignment at position i
independent of alignment at position j

© Burkhard Rost
ROSTLAB.

137/194
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.

- roulette parents
- cross-over to children
- mutate children
- compute fitness

START
Initialize population

STOP

Compute fitness

END

Crossover
parents

child

mutation

© Burkhard Rost
ROSTLAB.TUM
T-Coffee

- much slower, requires preprocessing, Genetic Algorithm

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

Extension

Multiple Sequence Alignment

ClustalW Primary Library (Global Pairwise Alignment)

Lalign Primary Library (Local Pairwise Alignment)

© Cedric Notredame, CRG Barcelona
T-Coffee: Use more information

Local Alignment

Multiple Alignment

Global Alignment

Specialist

Structural

Multiple Sequence Alignment

© Cedric Notredame, CRG Barcelona
Sequence-profile comparison

YDFHGVGEDDISIKRG

| PSI-BLAST | SF Altschul 1997 Nucl Acids Res 25 3389-3402 | © Burkhard Rost ROSTLAB. TUM | 147/194 |
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

YDFHGVGEDDISIKRGR
Zones

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

☐ pairwise
☐ multiple
☐ sequence-profile
☐ ?
Profile-profile comparison

fyn_human VTLVALDYD EARPEDLSF HKGEKQQLIN SSEQDNAER SLTDGQTOI
york_chick VTLVALDYD EARPEDLSF HKGEKQQLIN SSEQDNAER SLTDGQTOI
fgr_human VTLVALDYD EARPEDLSF HKGEKQQLIN SSEQDNAER SLTDGQTOI
yes_chick VTLVALDYD EARPEDLSF HKGEKQQLIN SSEQDNAER SLTDGQTOI
src_avis2 VTLVALDYD EARPEDLSF HKGEKQQLIN SSEQDNAER SLTDGQTOI
src_avis VTLVALDYD EARPEDLSF HKGEKQQLIN SSEQDNAER SLTDGQTOI
src_avis VTLVALDYD EARPEDLSF HKGEKQQLIN SSEQDNAER SLTDGQTOI
src_avis VTLVALDYD EARPEDLSF HKGEKQQLIN SSEQDNAER SLTDGQTOI
src_avis VTLVALDYD EARPEDLSF HKGEKQQLIN SSEQDNAER SLTDGQTOI
src_avis VTLVALDYD EARPEDLSF HKGEKQQLIN SSEQDNAER SLTDGQTOI
src_avis VTLVALDYD EARPEDLSF HKGEKQQLIN SSEQDNAER SLTDGQTOI
src_chick VTLVALDYD EARPEDLSF HKGEKQQLIN SSEQDNAER SLTDGQTOI
src_chick VTLVALDYD EARPEDLSF HKGEKQQLIN SSEQDNAER SLTDGQTOI
src_chick VTLVALDYD EARPEDLSF HKGEKQQLIN SSEQDNAER SLTDGQTOI
stk_hydat VTLVALDYD EARPEDLSF HKGEKQQLIN SSEQDNAER SLTDGQTOI
hck_human ..IVALQSLKXE ELHEDLSF KEGQDVMLE ES.GEWAH SLATQKFGI
blick mouse ..FVVALDYD AAHVHRDLQV LKGEKLQVLR .STGDQVAR SLVTQGKV
blick mouse ..TIVALQSLKXE ELHEDLSF KEGQDVMLE ES.GEWAH SLATQKFGI
lyn_human ..LIVALQSLKXE DGIMHDLQV LKGEKLQVLE .EGHWAH SLATQKFGI
lck_human ..LIVALQSLKXE EPSXHDLQV EKEQKHILE QS.GEWAH SLATQKFGI
ss81 Yeast .... ALYPY DADDDeDS NFQELQFSD .IEGRHRK R.ANGETQGI
abl1 mouse ..LFVALQSLKXE TSGQVSLG YnGQKAEQ ..TKQKQVW
abl1 human ..LFVALQSLKXE TSGQVSLG YnGQKAEQ ..TKQKQVW
src1 drome ..VVLLLLF YKRDNEQVLV DTEDQMHWV NLTRQGDL
myod dici ... ALYPY DADDDeDS NFQELQFSD .IEGRHRK R.ANGETQGI
yfj4 yeast .... ALYPY AEGKEDLPLK KQGDIYTLK KSQNDGWGR V .NQEGQF
abl2 human ..LFVALQSLKXE TSGQVSLG YnGQKAEQ ..TKQKQVW
tec_human ..EIVMMQFQ QAERQHUDLQ EKEQHILE YNQVHMAH D.KQGQF
txk human ..ALYPY DADDPNLA KREAELYL KYPHWAH D.RLGQGL
ypa2 yeast VRVALQSLK TTNDPELSL KQGDIYTLVQ QVYRDASSQG L ..KGNMZF
abpl_sacex .... ALYPY EAGEINELTF AHNKHIINIE FVDDEQLE LTTAQKGLF

© Burkhard Rost
ROSTLAB.TUM
Problems with profile-profile

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

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

```
ATRLLTTAKKDGPCD
ATRLTLTAKKDGPCD
ATRLTLTAKKDGPCD
ATKLCLLTAKKEGPKD
ATKLLTTAKKEGPKD
ATKLLTLGAKKEGGCD
ATWLLTTAKKVGPCD
ATWLLTTAKKVGPCD
```
“cheap” hack toward profile-profile
Thanks for slides
profile-profile: CPU intensive

D Przybylski & B Rost 2008 Bioinformatics 24: 1987-93
consensus sequence

profile

consensus sequence

| A | R | N | D | C | O | 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 | -7 | -5 | -6 |
| 4 | I | -3 | 6 | -4 | -1 | -5 | -2 | -2 | -4 | 1 | 3 | -3 | 0 | 0 | -3 | 0 | 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 | -3 | -1 | -3 | -1 | 3 | 0 | -1 | 0 | 3 | -2 | -2 | 0 |

D Przybylski & B Rost 2008 Bioinformatics 24: 1987-93

© Dariusz Przybylski, Columbia New York
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
Consensus-profile PSI-BLAST fast & good
Profile-profile modern: HHblits
HHblits - a tool from Munich/Tuebingen

M Remmert, A Biegert, A Hauser & J Söding

Johannes Söding
MPI Goettingen
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
Zones

- Midnight Zone
- Twilight Zone
- Daylight Zone

Sequences similar
Structures similar
PDB all-against-all

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

proteins of known 3D structure (PDB)

sequence-unique subset
Hypothetical distribution of similar structures
FAKE DATA
Evolution into the Midnight zone

![Graph showing the number of structure pairs against percentage pairwise sequence identity. The graph indicates a peak at around 10% identity, with a significant drop in the number of pairs as the identity increases.]
Midnight zone: real - random

B Rost 1997 *Folding & Design* 2, S19-S24
AS Yang and B Honig 2000 *J Mol Biol* 301, 679-689
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

- Midnight Zone
- Twilight Zone
- Safe Zone

sequence - sequence
sequence - profile
profile - profile

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

sequence
KAVIDNADSEEMQQQDSVECATQALEKYNIEKDIAMIKKEFDKYNYNPTWHCVGRNFGSYVTHETKHFIFYYLQVAILLFSEG

secondary structure:
LLLEEEELLLLHLLLLLLLLLLLLLLLLLLLLLLLHHHHHHHHHHHHHHHHHHHHHHHLLLHLLEEEEEEELLLLLLLLLLLLLEEEEEELL

2-state relative solvent accessibility:

OOOOOOOOOOSOOOOOOOSOOOSOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO
Two paths to fold recognition

<table>
<thead>
<tr>
<th>Seq (U)</th>
<th>3D PDB</th>
<th>Fosfos Profile</th>
<th>1D Projection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Str 1</td>
<td></td>
<td>EEH</td>
</tr>
<tr>
<td></td>
<td>Str 2</td>
<td></td>
<td>HEEH</td>
</tr>
<tr>
<td></td>
<td>Str 3</td>
<td></td>
<td>HEHH</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>Str n</td>
<td></td>
<td>EHHÉHE</td>
</tr>
</tbody>
</table>

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

Two paths to fold recognition
**TOPITS**

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

- LWQRLVTIKGGQLKEALLDTGAD
  - LWQRLVTIKGGQLKEALLDTGAD
  - LWQRLVTIKGGQLKEALLDTGAD
  - LWQRLVTIKGGQLKEALLDTGAD
  - LWQRLVTIKGGQLKEALLDTGAD

**Predict 1D structure from sequence**

- LWQRLVTIKGGQLKEALLDTGAD
  - LWQRLVTIKGGQLKEALLDTGAD
  - LWQRLVTIKGGQLKEALLDTGAD
  - LWQRLVTIKGGQLKEALLDTGAD
  - LWQRLVTIKGGQLKEALLDTGAD

- LWRRPVVTAEIGLEVVLLDTGAD
  - LWRRPVVTAEIGLEVVLLDTGAD
  - LWRRPVVTAEIGLEVVLLDTGAD
  - LWRRPVVTAEIGLEVVLLDTGAD
  - LWRRPVVTAEIGLEVVLLDTGAD

**Input:**
- sequence

**Generate sequence alignment**

**Predict 1D structure**

- LWQRLVTIKGGQLKEALLDTGAD
  - LWQRLVTIKGGQLKEALLDTGAD
  - LWQRLVTIKGGQLKEALLDTGAD
  - LWQRLVTIKGGQLKEALLDTGAD
  - LWQRLVTIKGGQLKEALLDTGAD

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

- EEEEE-----EHHHH-----
  - EEEEE-----EHHHH-----
  - EEEEE-----EHHHH-----
  - EEEEE-----EHHHH-----
  - EEEEE-----EHHHH-----

**Good match to one of the known structures?**

- • 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

<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>

sequence _______ generalized sequence
AVDG _______ AHB VHO DLO GLO

sequence profile (PSSM) _______ generalized profile (GPSSM)

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

© Dariusz Przybylski, Columbia New York
we have to establish first that similar structures have similar 1D strings (secondary structure & solvent accessibility)
Secondary structure conserved

Percentage six-state 1D structure identity vs. Percentage sequence identity

- observed-observed

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

Percentage six-state 1D structure identity vs. Percentage sequence identity

- Red squares: observed-observed
- Black diamonds: predicted-observed
- Orange squares: predicted-predicted

Fold  
Superfamily  
Family

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
Dariusz Przybylski

Thanks for slides

Dariusz Przybylski
Broad Inst, Boston
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
<table>
<thead>
<tr>
<th>Date</th>
<th>Monday</th>
<th>Tuesday</th>
<th>Wednesday</th>
<th>Thursday</th>
</tr>
</thead>
<tbody>
<tr>
<td>04/25</td>
<td>no lecture</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>04/27</td>
<td>no lecture</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>05/02</td>
<td>Organization of lecture: intro into cells &amp; biology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>05/04</td>
<td>Intro 2: amino acids, protein structure/domains</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>05/09</td>
<td>No lecture</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>05/11</td>
<td>Alignment 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>05/16</td>
<td>Alignment 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>05/18</td>
<td>Comparative modeling &amp; experimental structure determination &amp; secondary structure assignment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>05/23</td>
<td>SKIP: student assembly (SVV)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>05/25</td>
<td>SKIP: Ascension Day</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>05/30</td>
<td>SKIP: Whitsun holiday (05/15-17)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/01</td>
<td>1D: Secondary structure prediction 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/06</td>
<td>SKIP: Whitsun holiday (06/03-06)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/08</td>
<td>1D: Secondary structure prediction 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/13</td>
<td>1D: Secondary structure prediction 3 / Transmembrane structure prediction 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/15</td>
<td>SKIP: Corpus Christi</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/20</td>
<td>1D: Transmembrane structure prediction 2 / Solvent accessibility prediction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/22</td>
<td>1D: Disorder prediction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/27</td>
<td>2D prediction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/29</td>
<td>3D prediction / Nobel prize symposium</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>07/04</td>
<td>TBA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>07/06</td>
<td>recap 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>07/11</td>
<td>recap 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>07/12</td>
<td>examen</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>07/13</td>
<td>TBA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>07/18</td>
<td>TBA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>07/20</td>
<td>TBA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>07/22</td>
<td>TBA</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>