title: Alignments

short title: alignments_1

lecture: Protein Prediction I - Protein Structure / Burkhard Rost, TUM, 2011 summer
Announcements

Videos: SciVee

www.rostlab.org

THANKS:
Tim Karl + Haitam Sohby

NO lectures:
Tue Jun 28
Thu Jun 28 ?

LAST lecture: Jul 7

Examen: Jul 12 (?), 10:30 (likely this room)
• Makeup: likely: October 13 - morning

CONTACT: Marlena Drabik assistant@rostlab.org

Let it go. Let it out. Let it all unravel. Let it free and it can be A path on which to travel.
Today: Secondary structure prediction 1

- LAST WEEKs
  - Secondary structure prediction

- THIS WEEK
  - Alignments and “reach of comparative modeling”

- NEXT WEEK
  - Marc Ofmann: Molecular Dynamics (MD)
  - Comparative modeling
Sequence comparisons: pairwise methods
Alignments

PAPI
OPA
Alignments

.PAPI
OPA..
Alignments

.PAPI
OPA..
.PAPA
Alignments

.PAPI  
OPA..  
.PAPA  
.PAPU  
.PIPI
Alignments: Steps to do

- goal: align two sequences

  GGQLAKEEAL
  EGQPVEVLP

- 2 do:
  - (1) Find optimal superposition of the two
  - (2) Define “optimal”
Alignment: Local vs. Global

- **compare:**
  - global: all residues aligned
  - local: best matches

- **GGQLAKEEAL**
- **EGQPVEVLP**
- **GGQLAKEEAL**
- **EGQ..PVEVLP**
- **GGQLAKEEAL**
- **EGQ..PVEVLP**
- **GGQLAKEEAL**
- **EGQ..PVEVL**
Alignment: Assumptions about meaning

What does this mean? GGQLAKEEAL EGQ.PVEVL
Alignment: Assumptions about meaning

- What does this mean?

- Historical interpretation
  (reason to start working on the problem):

  evolutionary relation

GGQLAKEEAL
EGQ.PVEVL
Meaning of “homology”?
Alignment: Assumptions about meaning

What does this mean? GGQLAKEEAL EGQ. PVEVL

Historical interpretation (reason to start working on the problem):

- evolutionary relation
- common ancestor
- homology
Dynamic programming: brute force
How to align sequences?

1. visually
How to align sequences?

- 1. visually
Dynamic programming: Global

GGQLAKEEAL
EGQPVEVL
Dynamic programming: Global

GGQLAKEEAL
EGQPVEVL

GGQLAKEEAL
EGQPVEVL
Dynamic programming: Global

GGQLAKEEAL
EGQPVEVL

GGQLAKEEAL
EGQPVEVL

GGQLAKEEAL
EGQPVEVL
## Dynamic programming concept

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>G</th>
<th>Q</th>
<th>L</th>
<th>A</th>
<th>K</th>
<th>E</th>
<th>E</th>
<th>A</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</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>E</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>V</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>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dynamic programming concept

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>G</th>
<th>Q</th>
<th>L</th>
<th>A</th>
<th>K</th>
<th>E</th>
<th>E</th>
<th>A</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Dynamic programming concept

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>G</th>
<th>Q</th>
<th>L</th>
<th>A</th>
<th>K</th>
<th>E</th>
<th>E</th>
<th>A</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>1</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>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>2</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></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></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**GGQLAKEEAL**

**GQP..VEVL**

**GGQLAKEEAL**

**G.QP.VE.VL**
2 of 2 better than 4 of 8?

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>G</th>
<th>Q</th>
<th>L</th>
<th>A</th>
<th>K</th>
<th>E</th>
<th>E</th>
<th>A</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>P</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Gap/Insertions: penalty

- linear gap penalty:

N gaps cost N

EGQ......PVEVLP  <->  E.G.Q.P.V.EVLP
Gap/Insertions: penalty

- **linear gap penalty:**
  
  \[ \text{N gaps cost N} \]
  
  
  \[ \text{EGQ.....PVEVLP} \leftrightarrow \text{E.G.Q.P.V.EVLP} \]

- **idea of BLOCKS:**

How to enforce blocks in alignments?
Gap/Insertions: penalty

- **linear gap penalty:**
  
  N gaps cost N

  \[ \text{EGQ....PVEVLP} \leftrightarrow \text{E.G.Q.P.V.EVLP} \]

- **affine gap penalty:**
  
  gap open
  gap extension (elongation)

  typical go=10 x ge
Dynamic programming: optimal alignment

- 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
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

\[
SW = \sum_{k=l}^{L_{ali}} M_{U_kT_k} - Go \cdot N_{gap} - Ge \cdot (L_{gap} - N_{gap})
\]
Identity the best criterion?
Substitution matrices
Alignments in brief

>1bl8_A  mol:protein length:97  Potassium Channel Protein
ALHWRAAGAATVLIVLLAGSYLAVLAERGAPGAQLIT
YPRALWWSVETATTGYPDLYPVTLWGRCAVVVMVAGITSFGLVTAALATWFVGREQ

>1orq_C  mol:protein length:223  Potassium Channel
IGDVMEHPLVELGVSYAALLSVIVVVVECTMQLSGEYLVRLYVLVLVILYAWAYRAYKSGDPAGYVKKTLEY
PALVPAGLALIEGHLAGLGLFRLVRLRRFLRILHLHSRGSKFLSAIAADAADKIRFYHLFGAVMLTVLYGAFAIYIVEYPDPNSSIKSVFDALWWAVVTATTGYPDVPATIGKVIGIAVMLTGISAALTLLIGTVSNMFQKILV

Query= 1bl8_A  mol:protein length:97  Potassium Channel Protein
(97 letters)

>1orq_C  mol:protein length:223  Potassium Channel
Length = 223

Score = 58.5 bits (140), Expect = 4e-14
Identities = 26/72 (36%), Positives = 43/72 (59%)

Query: 21  GSYLAERGAPGACLITYPRALWWSVETATTGYPDLYPVTLWGRCAVVVMVAGITS 80
G++ E  P ++ +  ALWW+V TATTGYPD+ P T G+ + + VM+ GI++
Sbjct: 147  GAFFKIVVEYDPDNSSIKSVFDALWWAVVTATTGYPDVPATIGKVIGIAVMLTGISA 206

Query: 81  FGLVTAALATWF 92
L+ ++ F
Sbjct: 207  LTLLIGTVSNMF 218
Alignments in brief

>1bl8_A  mol:protein length:97  Potassium Channel Protein
ALHWRAAGAATVLLVLLAGSYLAVLAERGAPGAQLITYPRALWWSVETATTTVGYGDLYPVTLWGRCVAVVVMA
GITSFGLVTAALATWFVGVREQ

>1orq_C  mol:protein length:223  Potassium Channel
IGDVMELHPLVELGVYAAALLSVIVVVECTMQLSGEYLVRLYLDLILVLWADYAYRAYKSGDPAGYVKKTLEYIPALVPAGLLLIEGHLAGLFRILVRLLRFLRILLIHSRG
SKFLSAIAADAKIRFYHLFGAVMTVLYGAFAIYIVEYPDPNSSIKSVFDALWWAATTGVYGDPATPIGVKVI
GIAVMGLTGISALTLLIGTVSNMFQKHVL

Query= 1bl8_A  mol:protein length:97  Potassium Channel Protein
(97 letters)

>1orq_C  mol:protein length:223  Potassium Channel
Length = 223

Score = 58.5 bits (140), Expect = 4e-14
Identities = 26/72 (36%), Positives = 43/72 (59%)

Query: 21  GSYI AVLAERGAPGACQLITYPRALWWSVETATTTVGYGDLYPVTLWGRCVAVVVMA
G++ + E   P + + +  ALWW+V TATTVGYGD+ P T  G+ + + VM+ GI++
Sbjct: 147  GAFAIYIVEYPDPNSSIKSVFDALWWAATTGVYGDPATPIGVKVI
GIAVMGLTGISALTLLIGTVSNMFQKHVL

Query: 81  FGLVTAALATWF 92
L+  ++  F
Sbjct: 207  LTLLIGTVSNMF 218

Scoring matrix

Algorithm to optimize score
Alignment: Substitution Matrix: PAM

PAM: Point Accepted Mutations (substitution matrix)

Margaret Oakley Dayhoff (1925-1983)

introduced 1978 based on 1572 observed mutations in 71 families of closely related (85% PIDE) proteins
Protein Atlas (Atlas of Protein Sequence and Structure)
(1st sequence: insulin: 1955 Fred Sanger, 1st structures: 1958: Perutz et al. (hemoglobin) / Kendrew et al. (myoglobin) )


PAM1: 1 point mutation per hundred comparisons
PAMn=(PAM1)^n
BLOSUM

- BLOcks of amino acid SUbstitution Matrices
  Align only conserved regions
- compile log-odd ratios

\[ S_{i,j} = \log \frac{p_i \cdot M_{i,j}}{p_i \cdot p_j} = \log \frac{M_{i,j}}{p_j} = \log \frac{\text{observed frequency}}{\text{expected frequency}} \]

- BLOSUM\(n\)=threshold at \(n\)% pairwise sequence identity

S Henikoff & Jorja Henikoff (1992)
PNAS 89:10915-9
## Alignments: scoring matrix

| C   | S | T | P | A | G | N | D | E | Q | H | R | K | M | L | V | F | Y | W |
| 12  | 2 | 1 | 0 | 6 | 2 | 0 | 1 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 2 |
| S   | 0 | 2 | 1 | 3 | 0 | 1 | 0 | 6 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 2 |
| T   | -2| 1 | 3 | 0 | 1 | 0 | 6 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 2 | 0 |
| P   | -3| 1 | 0 | 6 | 2 | 1 | 0 | 1 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 2 |
| A   | -2| 1 | 1 | 2 | 0 | 1 | 0 | 6 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 2 |
| G   | -3| 1 | 0 | 1 | 5 | 2 | 1 | 0 | 6 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 0 |
| N   | -4| 1 | 0 | -1| 0 | 0 | 2 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 2 |
| D   | -5| 0 | 0 | -1| 0 | 1 | 2 | 4 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 0 |
| E   | -5| 0 | 0 | -1| 0 | 0 | 1 | 3 | 4 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 0 |
| Q   | -5| -1| 0 | -1| 0 | -1| 1 | 2 | 4 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 0 |
| H   | -3| -1| 0 | -1| 0 | -2| 2 | 1 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| R   | -4| 0 | -1| 0 | -2| 3 | 0 | 1 | -1| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| K   | -5| 0 | 0 | -1| 0 | -2| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| M   | -5| -2| -1| -2| -1| 3 | -2| 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| I   | -2| -1| 0 | -2| -1| 3 | -2| 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| L   | -6| -3| -2| -3| -2| 4 | 0 | -4| 3 | 4 | 2 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| V   | -2| -1| 0 | -1| 0 | -1| 2 | -2| 4 | 2 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F   | -4| -3| -3| -5| -4| -5| -2| -4| 5 | 0 | 1 | 2 | -1| 9 | 0 | 0 | 0 | 0 | 0 | 0 |
| Y   | 0 | -3| -3| -5| -3| -7| -2| -4| 4 | 0 | -4| 4 | 2 | -1| 1 | -2| 7 | 10 | 0 | 0 |
| W   | 8 | -2| -5| -6| -6| -7| -4| -7| 7 | 5 | 3 | 2 | 3 | 4 | 5 | 2 | 6 | 0 | 0 | 17 |

**Scoring Matrix**
Alignments: scoring matrix

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

Scoring Matrix

© Burkhard Rost (TU Munich)
Other substitution matrices

- BLOSUM-62 took the day
- many more (sequence based):
  - GONNET
  - McLachlan
  - Claverie
- structure based
  - PHAT, SLIM
  - STROMA, SDM, HSDM, SM_SAUSAGE, AGAPE
- function based
- for threading/fold recognition
- asf. ....
Interactive software tool

Ignacio Ibarra & Francisco Melo:

Interactive software tool to comprehend the calculation of optimal sequence alignments with dynamic programming

Bioinformatics, 2010, in press

http://melolab.org/sat
Dynamic programming

- time used?
- how to choose parameters?
Alignment: hashing (fast & dirty)
BLAST: fast matching of single ‘words’

TTYKLILNGKTLKGETTTEAVDAATAAEKVFKQYANDNGVDGEWTYDDATKTFTVTEK
TTYKLILLLLLLLLLLLLLLLLLLLAWTVEKAFKTFAAAAAAAAAAWTVEKAFKTFAAAAA

Default “word” size for “seeds” = 3
BLAST: fast matching of single ‘words’

TTYKLILNGKTLKGETTTEAVDAATAEKFQYANDNGVDGEWTYDDATKTFTVTEK
TTYKLILLLLLLLLLLLLAWTWVEKAFKTFAAAATWVEKAFKTFAAA

Default “word” size for “seeds” = 3
BLAST: fast matching of single ‘words’

TTYKLIILNGKTLKGETTTTEAVDAATAEKFQYANDNGVDGEWTYDDATKTFTVTEK
TTYKLIILLLLLLLLLLLLLLLAWTVEKAFKTFAAAAAAAAAAWTVEKAFKTFAAAAA

Default “word” size for “seeds” = 3
BLAST: fast matching of single ‘words’

Default “word” size for “seeds” = 3
BLAST: fast matching of single ‘words’

TTYKLI LNLGLKTLKGETTTEAVDAATAE}KVFQYANDNGVDGEWTYDDATKTFVTEK
TTYKLILLLLLLLLLLLLLLLLAWTVEKAFKTFFAAAAAAAAWTVEKAFKTFFAAAA

Default “word” size for “seeds” = 3
Examples for hashing

- FASTA (extension of FASTN/FASTP)
  DJ Lipman & WR Pearson (1985)
  Science 227: 1435-41

- BLAST
  Basic Local Alignment Search Tool
  SF Altschul, W Gish, W Miller, EW Myers, DJ Lipman (1990) JMB
  215:403-10
BLAST: fast matching of single ‘words’

TTYKLILNGKTLKGETTTEAVDAATAEKFQYANDNGVDGEWTYDDATKTFTVTEK
TTYKLLLLLLLLLLLLLLLLAWTVEKAFKTFAAAAAAAAAAWTVEKAFKTFAAAAA

Default “word” size for “seeds” = 3
BLAST: fast matching of single ‘words’

TTYKLIILNGKTLKGETTTEAVDAATAEKEVFQYANDNGVDGEWTYDDATKTFVTVEKT
TTYKLIILLLLLLLLLLLLLLAWTVEKAFKTFAAAAAATWVEKAFKTFaaaaa

Default “word” size for “seeds” = 3
BLAST: fast matching of single ‘words’

Default “word” size for “seeds” = 3
BLAST: fast matching of single ‘words’

Default “word” size for “seeds” = 3
BLAST: fast matching of single ‘words’

```
TTYKLILNGKTLKGETTTEAVAATAEKVFKQYANDNGVGEWTYDDATKTFVTTEK
TTYKLILLLLLLLALLLLAWTVEKAFKTFAAAAAAAAWTVEKAFKTFAAAAA
```

```
TTYKLIL  ?
TTYKLIL  ?
```

```
AATAEKVFKQYA
AWTVEKAFKTF
```

```
WTYDDATKTF
WTLVEKAFKTF
```

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)
Different approaches for background

- SEARCHSEQ
  - permutations of SEARCHSEQ

  FASTA

  - pre-compile distribution for entire database

  BLAST
How accurate are pairwise alignments?
Annotation transfer

similar sequence $\Rightarrow$ similar function

similar sequence $\Rightarrow$ similar structure
Zones

Midnight Zone

Twilight Zone

Save Zone

sequences similar

structures similar
How to assess alignment accuracy?
How to assess alignment accuracy?
All-vs-all: PDB

1D = sequence alignment
All-vs-all: PDB

3D = structural alignment

1D = sequence alignment
All-vs-all: PDB

3D = structural alignment

1D = sequence alignment

<0.2nm rmsd — SAME 3D
in between — ignore
>0.5nm rmsd — DIFFER in 3D
PDB all-against-all ok?

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) should not overlap with sequence-unique subset.
Sequence conservation of protein structure

B Rost 1999 Prot Engin 12, 85-94
Sequence conservation of protein structure

B Rost *Prot Engin* **12**, 85-94
Structure prediction using homology

>P1
MEDLVSVGITKHAEVEELARFESDEAVRIDGEVESFRGLSGL
VLLQTSnrveviasgardraeeelglihhdmvkgseavrhlfrvasgle8mvgqeii1rqvkaydraarlgtldealkiv
frrainlgkrrereetrisegvasi

>P2
METLILITQEEVESLISMDEAMNVEAPFLYALGKAQMPFKV
YLEFEKGDLORAPLHGAYGALKWNHPGDPKGLPVTMAL
HILNSPETGFPPLAVMDATYTTSLRTGAGGIAAKYL


Tuesday June 7, 2011
Structure prediction using homology


Tuesday June 7, 2011
Structure prediction using homology


**Figure A**

- **Percentage of identical residues**
- **Number of residues aligned**

**Figure B**

- **Percentage of identical residues**
- **Number of residues aligned**

**Query**

- P1: MEDLVSVG10HKEAEVEELEKAFESDEAVR7D1VESFG1SGS
  - VLLQTNSRVEVTASDGR5G1REDLH5DAMV5RG5AEV5H
  - LFRVASGLESMVGEQ5ILR1QV5KAD5RAARLGT1LDEALKIV
  - FRRAINLKRAREETR1SEGA5VI

- P2: METLILTQEEVESL1SMDA5MNAVEEAFRLY1GKAQ5PPK5V
  - YLEF5E5G5D5R5PA5H4LM5Y5G5L5K5WN5S5PH5NP5D5K5L5PT55MAL
  - NIL5S5PET5G5F5P5L5AV5MD5AT5T5S5L5RT5G5A5GG5IA5K

**Alignments**

- Score = 83.2 bits (205), Expect = 9e-17
- Identities = 18/101 (X%), Positives = 36/101 (35%), Gaps = 2/101 (1%)
- Query: 111 AAGGIAAKLARKNSSVFG1GCTQAYFQOLEALR5VFD1GEV5KAYD5VREKAAKF 170
  - AA +A +L + +G +G +L + V + +A +
- Sbjct: 153 AAVELAER5L5H5D5K5T5V5L5V5G5A5GM5K5T5V5A5K5L5V5D5R55VA5LV5N5R5V5E5RA5VEL 211
  - A5VELA5ERE5L5H5D5K5T5V5L5V5G5A5GM5K5T5V5A5K5L5V5D5R55VA5LV5N5R5V5E5RA5VEL

- Query: 171 EDRG15ASV5P5AE5EASSCDV5L5VTTTP55P5K5VV5K5AE5W5VE5GT 211
  - + +R DV5V5+ T + PV5 + V E
- Sbjct: 212 GGEAVRF5DE5-L5DHL5AR5SDV5V5S5ATA5AP5HP5V5I5H5DD5V5REAL 251

**Remarks**

- Structure prediction using homology
Structure prediction using homology


Tuesday June 7, 2011
Structure prediction using homology

Score = 33.9 bits (77), Expect = 0.068
Identities = 14/58 (y%), Positives = 28/58 (48%), Gaps = 2/58 (3%)
Query: 178 SVQPAAEASRCVLVTTPSRKPVKAEWVEEGTHINAIGADGPKQELD-VEILKKA 234
+ EE ++ D+LV T + +VK EW++ G + G + ++ E ++A
Sbjct: 198 TAHLDDEEVNKGDLVATGQPE-MVKGEWIKPGAIVIDCGINYKVVDAYDEAKERA 254

Score = 33.9 bits (77), Expect = 0.068
Identities = 14/58 (y%), Positives = 28/58 (48%), Gaps = 2/58 (3%)

Query: 178 SVQPAEEASRCDVLTTPSRFVKAEBVEEGTHINAIGADGPGQELD-VEILKKA 234
+ EE ++ D+LV T + +VK EW++ G + G + ++ E ++A
Sbjct: 198 TMLDEEVNKDILVATGQPE-MVKGEIKPGAIIDCGINYKVGDAVYDEAKERA 254

Structure prediction using homology


Tuesday June 7, 2011
Sequence conservation of protein structure

B Rost 1999 Prot Engin 12, 85-94

© Burkhard Rost (TU Munich)
Sequence conservation of protein structure

B Rost 1999 Prot Engin 12, 85-94
How to estimate performance from the curves?
How to estimate performance from the curves?
Distance from new HSSP-curve

Sequence identity implies structural similarity!

Don't know region

Distance from curve = +10

Distance from curve = -10

B Rost 1999 Prot Engin 12, 85-94
Distance from new HSSP-curve

B Rost 1999 Prot Engin 12, 85-94
Twilight zone = false positives explode!!

Percentage sequence identity

Number of protein pairs

Distance from HSSP threshold

B Rost 1999 Prot Engin 12, 85-94

© Burkhard Rost (TU Munich)
Twilight zone = false positives explode!!

Percentage sequence identity

Number of protein pairs

Distance from HSSP threshold

B Rost 1999 Prot Engin 12, 85-94
Detecting true hits in Twilight zone

Percentage of cumulative true positives

Distance from threshold

old HSSP

sim

ide

10%

similarity-larger-than-identity

they-dont-know-what-they-do only sequence identity

B Rost 1999 Prot Engin 12, 85-94

© Burkhard Rost (TU Munich)

Tuesday June 7, 2011
Finding similar structures in Twilight zone

B Rost 1999 Prot Engin 12, 85-94
Accuracy vs. Coverage

ACCURACY = how many of the correct proteins were found?

COVERAGE = how many of the proteins found are correct?
BLAST is not enough ...

they-dont-know-what-they-do
only sequence identity

B Rost 1999 Prot Engin 12, 85-94
Threshold in percentage sequence identity

- PIDE
- >100, >30%
- new HSSP
Threshold in percentage sequence identity
So far: pairwise - anything more 2 do?
Triangle beyond reach

- protein A
- protein B
- protein C
Sequence Space Hopping

protein A
seq_x
seq_y
sel_x

protein B
anb_x
unk_x
unk_y

protein C
cal_x
cal_y

© Burkhard Rost (TU Munich)
Success through sequence space hopping

![Graph showing percentage sequence identity and cumulative % of true pairs against distance from threshold.]

- The x-axis represents the distance from the threshold.
- The y-axis shows the percentage sequence identity.
- The graph includes two sequences: 'ide' and 'old', with their respective plots.
- The cumulative % of true pairs increases as the distance from the threshold decreases.

Tuesday June 7, 2011
Sequence comparisons:
multiple alignment
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)

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>G</th>
<th>Q</th>
<th>L</th>
<th>A</th>
<th>K</th>
<th>E</th>
<th>E</th>
<th>A</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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
# Multiple alignment: progressive 1

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>GGQLAKEEAL</th>
<th>GGQLAKDEAL</th>
<th>GGQIAKDEAL</th>
<th>GGQIAKDEAI</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>90</td>
<td>80</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>90</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td>90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Multiple alignment: progressive

### Step 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
- GGQIAKDEAL
- GGQIAKDEAI

- ggqlakeeal
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></td>
<td>90</td>
<td>80</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**
- GGQIAKDEAL
- GGQIAKDEAI
- ggqiakdeal
## 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
- ggqiakdeai

**Step 3**
- GGQIAKDEAL
- GGQIAKDEAI
- ggqiakdeal

---

Tuesday June 7, 2011
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></td>
<td>90</td>
<td>80</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
- GGQLAKDEAL
- GQQIAKDEAI
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></td>
<td>90</td>
<td>80</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
- GGQIAKDEAI
Announcements

Videos: SciVee
www.rostlab.org

THANKS:
Tim Karl + Haitam Sohby

NO lectures:
Tue Jun 28
Thu Jun 28?

LAST lecture: Jul 7

Examen: Jul 12 (?), 10:30 (likely this room)
• Makeup: likely: October 13 - morning

CONTACT: Marlena Drabik assistant@rostlab.org

© Burkhard Rost (TU Munich)