Sequence alignments (pairwise)

pp1_alignments1

Computational Biology 1 - Protein structure (for Informatics) - TUM summer semester
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

Exercises:
teaching@rostlab.org

CONTACT: Lothar Richter richter@rostlab.org

Announcements

Dmitrij Nechaev

Jonas Reeb
Lothar Richter
Michael Bernhofer
Today: Alignments 1

LAST
- 3D comparison / Alignment chalk board

TODAY
- Alignments and “reach of comparative modeling”

NEXT
- alignment contd
Science is communication

questions are often the first step
Why compare 3D shapes, why not function?
COPS metric

Triangle inequality/transitivity:

protein A ≈ protein B
protein B ≈ protein C

⇒ protein B ≈ protein C (does not imply: protein A similar to protein C)
RMSD is not a metric

- $cRMSD_B = 2.8 \text{ Å} = 0.28 \text{ nm}$
- $cRMSD_C = 2.85 \text{ Å} = 0.285 \text{ nm}$

A similar B
B similar C
NOT implying:
A similar C
COPS metric

Axioms / Definitions:

\[ S_{a,a} = L_a \]
\[ S_{a,b} \geq 0 \]
\[ S_{a,b} = S_{b,a} \]
\[ S_{b,c} \geq S_{a,b} + S_{a,c} - L_a \]
\[ D_{a,b} = L_a + L_b - 2S_{a,b} \]

• Alignment method not so important!
  for COPS: TopMatch

• Metric can reveal alignment problems
  (e.g. via triangle inequality)

3D classification - COPS/TopSearch

COPS
http://cops.services.came.sbg.ac.at
[SJ Suhrer et al. (2009) NAR 37, W539-W44. ]

PDB id 1z6t-A with 2a5y-B

Manfred Sippl
Salzburg
Notation: protein structure 1D, 2D, 3D
Sequence comparisons: pairwise methods
Alignments answer: How similar are proteins?
Alignments

OPA
PAPI
Alignments

OPA..
.PAPI
Alignments

OPA.. .PAPI .PAPA
Alignments

.PAPI
OPA..
.PAPA
.PAPU
.PIPI
Alignments

.PAPI
OPA..
.PAPA
.PAPU
.PIPI

father
Alignments

.PAPI
OPA..
.PAPA
.PAPU
.PIPI

father

something else
“Right” alignment would have been
“Right” is defined by objective function
“Correct” alignment: objective function

- simplest objective function: percentage of letters identical
- more complicated functions describing ‘match’

- but all this ignores “macroscopic similarity” (from level of strings to biology)
  objectives could be:
  - match structure (identify all pairs A,B with RMSD(A,B)<2.5Å)
  - match function (identify all pairs A,B with same function)
  - evolutionary relation (e.g. all A,B with common ancestor)
For time being: let us fix the ‘string matching’ problem
Alignment: Local vs. Global

- **compare:**

- **global: all residues aligned**
  
  - GGQLAKEEAL
  - EGQPVEVLP

- **local: best matches**

  - GGQLAKEEAL
  - GQ..PVEVL
Alignment: Assumptions about meaning

- Biological meaning of insertion ("." )?

- Historical interpretation (reason to start working on the problem):
  - evolutionary relation
  - also implied:
    - similar structure
    - similar function/role

GGQLAKEEAL
GQ..PVEVL
Meaning of “homology”?
phylogeny:
phylo: race/tribe/kind
genesis: birth
genos: origin
Phylogenetic Tree Of Life

© Wikipedia
Phylogenetic Tree Of Life

© Wikipedia
Phylogenetic Tree of Life

Bacteria
- Spirochetes
- Proteobacteria
- Cyanobacteria
- Ranctomyces
- Bacteroides
- Aquifex
- Thermotoga

Archaea
- Methanococci
- Methanosarcina
- Methanobacterium
- Thermoproteus
- Halobacteria
- Entamoeba
- Slime molds
- Animals
- Fungi

Eucaryota
- Animals
- Fungi
- Green nonsulfur bacteria
- Actinobacteria
- Gram-positives
- Chlamydiaceae
- Actinobacteria
- Pianctomycetes
- Spirochaetes
- Fusobacteria
- Cyanobacteria (blue-green algae)
- Thermophilic sulfate-reducers
- Crenarchaeota
- Nanoarchaeota
- Euryarchaeota
- Slime mounds
- Plants
- Algae
- Protozoa
- Proteobacteria
- Acidobacteria

© Wikipedia
Phylogenetic Tree of Life

© Wikipedia
homology
homology definition

etymology: Greek:

ομόλογος (homologos):

homo (agreeing/same) + logos (word/reason/opinion)

wikipedia: relation
homology definition

-etymology: Greek: ομόλογος (homologos):
  homo (agreeing/same) + logos (word/reason/opinion)

-wikipedia: relation

-originally (1843 Richard Owen):

“same organ in different animals under every variety of form and function”

1847 diagram by Richard Owen shows his conceptual archetype for all vertebrates
homology definition

- **etymology:** Greek: ὁμόλογος (homologos): homo (agreeing/same) + logos (word/reason/opinion)

- **originally (1843 Richard Owen):** “same organ in different animals under every variety of form and function”

- **Genes:** homologs originate from common ancestor
  - orthologs: speciation event
  - paralogs: duplication event
Evolution: speciation

© http://evolution.berkeley.edu/evosite/evo101/

species=?
Evolution: speciation

© [http://evolution.berkeley.edu/evosite/evo101/](http://evolution.berkeley.edu/evosite/evo101/)

species=mating
Evolution: speciation

© http://evolution.berkeley.edu/evosite/evo101/

happily munching
Evolution: speciation

© http://evolution.berkeley.edu/evosite/evo101/

disaster strikes
Evolution: speciation

© http://evolution.berkeley.edu/evosite/evo101/

populations diverge
Evolution: speciation

© http://evolution.berkeley.edu/evosite/evo101/

rejoined - yet separated
Evolution: speciation


Biology defines species how?
two species mate

> off-spring not passing on genetic information
Evolution: speciation

© http://evolution.berkeley.edu/evosite/evo101/
Evolution: speciation

© http://evolution.berkeley.edu/evosite/evo101/
Evolution: speciation

Happy Face Spider *Theridion grallator*
(same species-interbreed)

Carrion/Hooded crow
(same species?)
homology definition

- etymology: Greek: ομόλογος (homologos):
  homo (agreeing/same) + logos (word/reason/opinion)
  (wikipedia: relation)
- originally (1843 Richard Owen):
  “same organ in different animals under every variety of form and function”
- Genes:
  homologs originate from common ancestor
  - orthologs: speciation event
  - paralogs: duplication event
- misuse of term “homolog structure”
  => similar structure
“Homology” frequently and incorrectly used for genes/proteins
... but, it reminds us of the connection in evolutiona
Alignment: Assumptions about meaning

- Biological meaning of insertion (".")?
  - GGQLAKEEAL
  - GQ..PVEVL

- Historical interpretation
  (reason to start working on the problem):
  - evolutionary relation
  - common ancestor
  - homology
Alignment: Assumptions about meaning

- Biological meaning of insertion (".")?

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

  evolutionary relation

- what we also WANT:
  - similar structure
  - similar function

GGQLAKEEAL
GQ..PVEVL
we compare 1D sequences to find similarity in 3D structure / function
we compare 1D sequences to find similarity in 3D structure / function
Dynamic programming: brute force
How to align sequences?

1. visually

GGQLAKEEAL
GQ..PVEVL
Dynamic programming: Global

GGQLAKEEAL
EGQPVEVL..
Dynamic programming: Global

GGQLAKEEAL
EGQPVEVL

GGQLAKEEAL
. EGQPVEVL .
Dynamic programming: Global

GGQLAKEEAL
EGQPVEVL

GGQLAKEEAL
EGQPVEVL

GGQLAKEEAL
...EGQPVEVL
1. visually


A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins

SAUL B. NEEDLEMAN AND CHRISTIAN D. WUNSCH
Department of Biochemistry, Northwestern University, and Nuclear Medicine Service, V. A. Research Hospital Chicago, Ill. 60611, U.S.A.
(Received 21 July 1969)

A computer adaptable method for finding similarities in the amino acid sequences of two proteins has been developed. From these findings it is possible to determine whether significant homology exists between the proteins. This information is used to trace their possible evolutionary development.

The maximum match is a number dependent upon the similarity of the sequences. One of its definitions is the largest number of amino acids of one protein that can be matched with those of a second protein allowing for all possible interruptions in either of the sequences. While the interruptions give rise to a very large number of comparisons, the method efficiently excludes from consideration those comparisons that cannot contribute to the maximum match.

Comparisons are made from the smallest unit of significance, a pair of amino acids, one from each protein. All possible pairs are represented by a two-dimensional array, and all possible comparisons are represented by pathways through the array. For this maximum match only certain of the possible pathways must be evaluated. A numerical value, one in this case, is assigned to every cell in the array representing like amino acids. The maximum match is the largest number that would result from summing the cell values of every pathway.
## Dynamic programming concept

<table>
<thead>
<tr>
<th>First letter</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></td>
<td>1</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>
<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>E</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</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>1</td>
<td></td>
<td>1</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></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></td>
<td>1</td>
<td>1</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>1</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>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>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>1</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>

2nd letter:

move on diagonal
## 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></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</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>P</td>
<td></td>
<td></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>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>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>1</td>
<td></td>
<td></td>
<td>1</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>

2nd letter:

- GQP
- GGQ
## Dynamic programming concept

<table>
<thead>
<tr>
<th>2nd letter:</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>1</td>
<td>1</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>GQ</td>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GP</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPQ</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>GPQP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>GPQG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>GPQGQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPQGPL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPQGPLV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPQGPLVE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPQGPLVEV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPQGPLVEL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPQGPLVELP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The 2nd letter is better for: GQP, GGQ, GQ, GP, and P.
Dynamic programming concept no ins

<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>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</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></td>
<td></td>
<td>1</td>
<td>2</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>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>2</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>1</td>
<td>2</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>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
allowing for gaps

no gap
GGQLAKEEAL
GQPVEVL

with gap
GGQLAKEEAL
GQ..PVEVL
Dynamic programming concept ins
Dynamic programming dot plots

© Wikipedia

Dot plot of a human zinc-finger transcription factor (GenBank NM_002383) against itself to show self-similarity
Gap/Insertions: penalty

- linear gap penalty:

  N gaps cost N

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

- linear gap penalty:
  
  N gaps cost N

  EGQ.....PVEVLP  <->  E.G.Q.P.V.EVLP

- idea of BLOCKS:

How to enforce blocks in alignments?

regular alignment score = matches - gaps

$$NW = \sum_{k=l}^{L_{ali}} M_{U_kT_k} - Go \cdot N_{gap}$$
Gap/Insertions: penalty

- **linear gap penalty:**
  
  \[N \text{ gaps cost } N\]

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

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

  \[\text{typical } \text{go} = 10 \times \text{ge}\]
Dynamic programming: optimal alignment

Pair of protein sequences
U: GGQLAKFEAL
T: EGQVEV

Optimal alignment (no gaps)
U: GGQLAKFEAL
T: EV

Optimal alignment (with gaps)
U: GGQLAKFEAL
T: HGQVEV

• 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

\[ SW = \sum_{k=1}^{L_{\text{ali}}} M_{U_k T_k} - Go \cdot N_{\text{gap}} - Ge \cdot (L_{\text{gap}} - N_{\text{gap}}) \]
What is best: local or global?
global clearly worse?

GGQLAKEEAL
G.Q..PVEVL

GGQLAKEEAL
GQ..PVEVL
Is really the total number of matches the best criterion?
Is really the total number of matches the best criterion?

Remember: protein domains are 60-400 residues
what is best?

<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></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</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></td>
<td>1</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></td>
<td>3</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>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>
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>G</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>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td>3</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>4</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>
How to answer the question which is best?

2 in 2

GQ
GQ

or

4 in 7

GQPVE•V•L
GQLAKEEAL
Identity the best criterion?
Substitution matrices
Alignments in brief

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  GSYYLAERGAQPAGAQLITYPRALWWSVETATTTVGYGDLPVTLDWRGCVAVVVMVAGITS 80
        G++  + E  P + + +  ALWW+V TATTVGYGD+ P T  G+ + + VM+ GI++
Sbjct: 147  GAFALYIVFYDPNSSIKSVFDALWWAVVTATTTVGYGDVVPATPIGKVIGIAVMLTGISALTLLIGTVSNMFQKILV

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

Scoring matrix

Algorithm to optimize score
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

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

- BLOSUMn=threshold at n% pairwise sequence identity

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

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>G</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>U</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>P</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>N</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Q</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>R</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>M</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>V</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>W</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Scoring Matrix


© Burkhard Rost

ROSTLAB.

TUM
### Alignments: scoring matrix

**Scoring Matrix**

![Scoring Matrix Diagram]

**Reference:**
### Alignments: scoring matrix

<table>
<thead>
<tr>
<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>C</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U</td>
<td></td>
<td></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></td>
<td></td>
<td></td>
<td>1</td>
<td>3</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>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>W</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>


© Burkhard Rost

ROSTLAB

TUM
**Alignments: scoring matrix**

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>S</th>
<th>U</th>
<th>S</th>
<th>P</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>D</th>
<th>E</th>
<th>O</th>
<th>Q</th>
<th>H</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>I</th>
<th>V</th>
<th>F</th>
<th>G</th>
<th>U</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

[@Henikoff1992](https://www.pnas.org/content/89/24/10915)
Many more substitution matrices exist today
Many more substitution matrices exist today

Most use: BLOSUM62
Many more substitution matrices exist today. Most use: BLOSUM62

$$S_{i,j} = \log \frac{p_i \cdot M_{i,j}}{p_i \cdot p_j} = \log \frac{M_{i,j}}{p_j} = \log \frac{\text{observed frequency}}{\text{expected frequency}}$$
Interactive software tool

Ignacio Ibarra & Francisco Melo:

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

Bioinformatics 2010, 26:1664-5

http:/melolab.org/sat
dynamic programming
optimal, but ...
Issues: Dynamic programming

- time used?
- order of length^2 -> long
### Big data - big CPU

<table>
<thead>
<tr>
<th>Year</th>
<th>Proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009/09</td>
<td>$8 \cdot 10^6$</td>
</tr>
<tr>
<td>2014/04</td>
<td>$55 \cdot 10^6$</td>
</tr>
</tbody>
</table>

- **6.9 fold**
- **55 months**

*double every 2 years*
### Big data - big CPU

<table>
<thead>
<tr>
<th></th>
<th>Proteins</th>
<th>comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009/09</td>
<td>$8 \cdot 10^6$</td>
<td></td>
</tr>
<tr>
<td>2014/04</td>
<td>$55 \cdot 10^6$</td>
<td></td>
</tr>
</tbody>
</table>

- **Cytochrome P450 - Mentha piperita - peppermint**
  - MELLQLWSALIIIVTTYTSLLINQWRKPQPKPGFPKGFLPLIGHLHLLWGKLPOHAL
  - ASVAKEGPVAVHQLGEVFSVLLSSREATKEAMKLVDPACANRFESIGTRIMWYDNEDII

- **Cytochrome P450 - Mentha spicata - spearmint**
  - MELDDLSAIIILVATYIVSSLINQWRSKOS/QSLLPPPKLPGVHIIHFLWGGLPQHFR
  - SIAQKYGPVAVHQLGEVYSVLLSSAEAKQAMKVLDPNFADRFDGIGSRTMWDYDKIIIF

- double every 2 years
# Big data - big CPU

<table>
<thead>
<tr>
<th>Year</th>
<th>Proteins</th>
<th>Comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009/09</td>
<td>$8 \cdot 10^6$</td>
<td>$64 \cdot 10^{12}$</td>
</tr>
<tr>
<td>2014/04</td>
<td>$55 \cdot 10^6$</td>
<td>$3025 \cdot 10^{12}$</td>
</tr>
</tbody>
</table>

- Proteins: double every 2 years
- Comparisons: >20-fold every 2 years

47 fold every 55 months
## Big data - big CPU

<table>
<thead>
<tr>
<th></th>
<th>Proteins</th>
<th>comparisons</th>
<th>we need</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009/09</td>
<td>$8 \cdot 10^6$</td>
<td>$64 \cdot 10^{12}$</td>
<td>5 TB</td>
</tr>
<tr>
<td>2014/04</td>
<td>$55 \cdot 10^6$</td>
<td>$3025 \cdot 10^{12}$</td>
<td>300 TB</td>
</tr>
</tbody>
</table>

- Double every 2 years
- 20-fold every 2 years
- >5-fold every 2 years
Issues: Dynamic programming

- time used?
  order of length^2 => long

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

TTYKLIILNGKTLKGETTTEAVDAATAEKEVFQYANDNGVDGEWTYDDATKTFTVTEK
TTYKLIILLLLLLLLLLLLLLLLLLAWTVEKAFKTFAAAATATATVEKAFKTFAAAAA

#1 seed=3

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

TTYKLI\_LNGKTLKGETTTEAVDAATAEKFVKQYANDNGVDGEWTYDDATKFTVTEK
TTYKLI\_LLL\_LLLLLLLLLLLLLLLAWTVEKAFKTFFAAAAAAAAWTVEKAFKTF

#1 seed=3

TTYKLIL  AATAEKFVKQYA  WTYDDATKTF
TTYKLIL  AWTVEKAFKTF  WTVEKAFKTF

#2 extend

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

TTYKLILNGKTLKGETTTEAVDAATAEKEVFKQYANDNGVDGEWTYDDATKTFTVTEK
TTYKLILLLLLLLLLLLLLLLLLAWTVEKAFKTFFAAAAAAAAAAWTVEKAFKTF

---

#1 seed=3

---

#2 extend

TTYKLIL
TTYKLIL

TTYKLIL

AATAEKVFKQYA
AWTVEKAFKTF

TTYKLIL

WTYDDATKTF
WTVEKAFKTF

Default “word” size for “seeds” = 3
Minds behind hashing

- FASTA (extension of FASTN/FASTP)

- 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’

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

- BLAST
  - pre-compile distribution for entire database

William R Pearson

David Lipman

Stephen Altschul
How accurate are pairwise alignments?
Annotation transfer

similar sequence \(\rightarrow\) similar function

similar sequence \(\rightarrow\) similar structure
Sequence -> Structure

Sequence folds into unique structure

S -> T
Sequence -> Structure

- Sequence folds into unique structure
  \[ S \rightarrow T \]
- Similar sequences fold into similar structures
  \[ S + S' \rightarrow T \]
Sequence -> Structure

- Sequence folds into unique structure
  \[ S \rightarrow T \]
- Similar sequences fold into similar structures
  \[ S + S' \rightarrow T \]
- Most sequences don’t fold, at all
  \[ S \rightarrow \text{no } T \]
Zones

Zones

Zones

Schematic: structural similarity in Twilight zone

True Positives = pairs of proteins with similar structure

Percentage: accuracy - or - specificity

sketch illustrating concept of twilight zone - NOT real data
Schematic: true and false in Twilight zone

True Positives = pairs of proteins with *similar* structure
False Positives = pairs of proteins with *DIFFERENT* structure

Number of pairs:

*sketch illustrating concept of twilight zone - NOT real data*
How to assess alignment accuracy?
All-vs-all: PDB

3D = structural alignment

1D = sequence alignment

proteins of known 3D structure (PDB)

<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!
Original HSSP curve

HSSP curve revised with 10x data

\[ \text{PIDE} = \text{Percentage pairwise sequence identity} \]

\[ \text{Number of residues aligned} \]

\[ \text{old Sander & Schneider} \]

\[ \text{new} \quad 480^*L^{-0.32}(1 + 1/e^L/1000) \]


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

C Sander & R Schneider 1991 *Proteins* 9:56-69
B Rost 1999 *Prot Engin* 12, 85-94
Structure prediction using homology

Score = 83.2 bits (205), Expect = 9e-17
Identities = 18/101 (17%), Positives = 36/101 (35%), Gaps = 2/101 (2%)

Query: MEDLVSVGITHKEAEVELEKARFESDEAVRDIVESFGLSGS
VLLQTRSVGTVASAGALRDAEELGLDHYLDMVAVKGSAVRHLFRVAGLHSGVTGEGSILRQVKKAYDRAARLGTLDEALKIV
FRAILNGKRAEETGSEGAVSI

Sbjct: METLILTQEEVESLISMDAMNAVEEAFLYLAFLGAQMPKVPVLIFEGDOLRMPAHLMGAYGLKWNHSPNPDKGLPTVMALN
ILNSPETGFLPAVMDATYYTSLRTAGGGIAAKYLGGEAVRFDE-LVDHLARSDVVVSAAPHPVIHDDVREAL

C Sander & R Schneider 1991 *Proteins* 9:56-69
Structure prediction using homology

C Sander & R Schneider 1991 *Proteins* 9:56-69
Raw data: density? lesson learned?

C Sander & R Schneider 1991 *Proteins* 9:56-69
B Rost 1999 *Prot Engin* 12, 85-94
How to estimate performance from the curves?
Distance from new HSSP-curve

C Sander & R Schneider 1991 Proteins 9:56-69
B Rost 1999 Prot Engin 12, 85-94

© Burkhard Rost

ROSTLAB. TUM
Twilight zone = true positives explode
Twilight zone = false positives bazoom!!

Percentage sequence identity

Number of protein pairs

Distance from HSSP threshold

B Rost 1999 Prot Engin 12, 85-94
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
Success through sequence space hopping

intermediate sequences

Percentage sequence identity

Cumulative % of true pairs

Distance from threshold

Percentage of cumulative true positives

Distance from threshold
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)
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
Dynamic programming?
for 3 sequences: $O(N1 \times N2 \times N3)$
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>
</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>

Multiple alignment:

- **A**: GGQLAKEEAL
- **B**: GGQLAKDEAL
- **C**: GGQIAKDEAL
- **D**: GGQIAKDEAI
## 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>

GGQLAKEEAL  
GGQLAKDEAL  
GGQIAKDEAL  
GGQIAKDEAI

---

Step 1

GGQLAKEEAL  
GGQLAKDEAL  
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
- ggqiaakdeal
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
- gqqiakdeal

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

---

**Step 1**

- GGQLAKEEAL
- GGQLAKDEAL
- 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:
- GGQIAKDEAL
- GGQIAKDEAI
- ggqlakeeal
- 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**
- GGQIAKDEAI

**Step 3**
- GGQIAKDEAI
Lecture plan (PP1 structure)

- **01**: 04/25 Tue: no lecture
- **02**: 04/27 Thu: no lecture
- **03**: 05/02 Tue: Organization of lecture: intro into cells & biology
- **04**: 05/04 Thu: Intro 2: amino acids, protein structure/domains
- **05**: 05/09 Tue: No lecture
- **06**: 05/11 Thu: Alignment 1
- **07**: 05/16 Tue: Alignment 2
- **08**: 05/18 Thu: Comparative modeling & experimental structure determination & secondary structure assignment
- **09**: 05/23 Tue: SKIP: student assembly (SVV)
- **10**: 05/25 Thu: SKIP: Ascension Day
- **11**: 05/30 Tue: SKIP: Whitsun holiday (05/15-17)
- **12**: 06/01 Thu: 1D: Secondary structure prediction 1
- **13**: 06/06 Tue: SKIP: Whitsun holiday (06/03-06)
- **14**: 06/08 Thu: 1D: Secondary structure prediction 2
- **15**: 06/13 Tue: 1D: Secondary structure prediction 3 / Transmembrane structure prediction 1
- **16**: 06/15 Thu: SKIP: Corpus Christi
- **17**: 06/20 Tue: 1D: Transmembrane structure prediction 2 / Solvent accessibility prediction
- **18**: 06/22 Thu: 1D: Disorder prediction
- **19**: 06/27 Tue: 2D prediction
- **20**: 06/29 Thu: 3D prediction / Nobel prize symposium
- **21**: 07/04 Tue: TBA
- **22**: 07/06 Thu: recap 1
- **23**: 07/11 Tue: recap 2
- **24**: 07/12 Thu: examen
- **25**: 07/13 Tue: TBA
- **26**: 07/18 Thu: TBA
- **27**: 07/20 Tue: TBA
- **28**: 07/22 Thu: TBA
- **29**: 07/25 Thu: TBA