COMPUTATIONAL STRUCTURAL BIOLOGY

STRUCTURE, SIMULATION, FUNCTION & PREDICTION

Lecture 6

Michael Levitt

Structural Biology, Stanford

http://csb.stanford.edu/class
BIOINFORMATICS I

Data in Biology
Statistics of Comparison.
Data Visualization.
Databases.
Web Resources.
Sequence Comparison.
DATA IN BIOLOGY

Strings (1-D).
Sequence.
Relationships (2-D).
Multiple Sequence Alignments.
Data in 3-D.
Sequence Objects.
Structure Objects.
STRINGS

Proteins sequences: AVHTIKHERWTQ

DNA sequence: ATGGCATGACAA

English Text: A CAT SAT ON

Numbers: 123457980123

Digits of π: 3.1415926535 897932384626
RELATIONSHIPS

Trees.
Graphs.
Directed Graphs.
Pathways.

http://www.grt.kyushu-u.ac.jp/spad/pathway/pdgf.html
MULTIPLE SEQUENCE ALIGNMENTS

Initial consensus alignment

1CEL > SACTLQSETHPPLTWQKCSSGGTCTQQTGSVVIDANWRWTHATNSSTNCYDGTNSSTLCPDNETCAK---NCCEL
1EG1 > QPGTSTPEVHPKLYTKTGCGCVAQDTSTVLLDNWYRWMH-DANYNCTVNGGVTNLCP---DEATCGKNCFEE
2A39 > KPGETKEVHPQTTFRCTKRGGCKPATNFIVLSDLSHPIHRAEGGLPGGCDDWGNPPPKDVCPDVESECAKNCTIME
2OVW > TPDKAKEQHPKLETRYRCTKASGCKKQTNYIVADAGIGHGRQNG---AGCGDWGQKPNATACPDEASCAKNCLISSL

Resolved alignment

1CEL > SACTLQSETHPPLTWQKCSSGGTCTQQTGSVVIDANWRWTHATNSSTNCYDGNTNSSTLCPDNETCAKNCCLDG
1EG1 > QPGTSTPEVHPKLYTKTGGCGCAQDTSTVLLDNWYRTMH---ANYNCTVNG---GVNTLCPDEATCGKNCFEE
2A39 > KPGETKEVHPQTTFRCTKRGGCKPATNFIVLSDLSHPIHRA---EGLPGGCCDDWGNPPPKDVCPDVESECAKNCTIME
2OVW > TPDKAKEQHPKLETRYRCTKASGCKKQTNYIVADAGIGHGRQ---KNGACGGGWQKPNATACPDEASCAKNCLISSL

Michael Levitt 04
3-D DATA

Protein Structures.

Cells.

Galaxies.
SEQUENCE OBJECTS

- **Data:**
  - DNA sequences.
  - RNA sequences.
  - Protein sequences.

- **Methods (Biological):**
  - RNA sequence is derived from DNA sequence: Transcription.
  - Protein sequence is derived from RNA sequence: Translation.

- **Methods (Evolutionary):**
  - DNA sequence A is similar to DNA B: Homology.
  - RNA sequence A is similar to DNA B: Gene Finding.

These methods are all sequence to sequence.
STRUCTURE OBJECTS

- **Data:**
  - Organic Molecule.
  - Fibrous Protein.
  - Globular Protein.
  - Membrane Protein.
  - RNA Structure.

- **Methods (Biological):**
  - Organic Molecule is bound to Membrane Protein: Drugs.
  - Globular Protein is transformed to Fibrous Protein: Mad Cow disease.
  - Globular Protein transforms Organic Molecule: Enzymes.

- **Methods (Evolutionary):**
  - Protein structure A is similar to protein structure B: Homology.

These methods are structure to structure.
Statistics of Comparison Concept 6.2
STATISTICS IS IMPORTANT IN BIOINFORMATICS

• All comparisons aim to determine if the observed level of similarity is significant.

• More precisely, what is the probability that the observed level of similarity could have been found between objects that are not--similar?

• In other words: could such a level of similarity be actually observed for non--similar objects?
**MEAN**

- Roll two die and get total of their values.
- Repeat many times and plot distribution of values.

\[
\text{Mean} = \frac{(2 \times 1 + 3 \times 2 + 4 \times 3 + 5 \times 4 + 6 \times 5 + 7 \times 6 + 8 \times 5 + 9 \times 4 + 10 \times 3 + 11 \times 2 + 12 \times 1)}{(1 + 2 + 3 + 4 + 5 + 6 + 5 + 4 + 3 + 2 + 1)} = \frac{252}{36} = 7
\]

2 = 1+1  
3 = 1+2, 2+1  
4 = 2+2, 1+3, 3+1  
5 = 2+3, 3+2, 1+4, 4+1  
6 = 3+3, 2+4, 4+2, 1+5, 5+1  
7 = 3+4, 4+3, 2+5, 5+2, 1+6, 6+1  
8 = 4+4, 3+5, 5+3, 2+6, 6+2  
9 = 4+5, 5+4, 3+6, 6+3
STANDARD DEVIATION

Mean Squared

\[
\frac{(2^2 \times 1 + 3^2 \times 2 + 4^2 \times 3 + 5^2 \times 4 + 6^2 \times 5 + 7^2 \times 6 + 8^2 \times 5 + 9^2 \times 4 + 10^2 \times 3 + 11^2 \times 2 + 12^2 \times 1)}{(1+2+3+4+5+6+5+4+3+2+1)}
\]

\[= \frac{1974}{36} = 54.83\]

\[\text{Mean Squared (Mean)}^2\]

Variance \[= (54.83 - 7^2)\]

\[= 5.83\]

SD \[= \sqrt{\text{Variance}}\]

\[= \sqrt{5.83}\]

\[= 2.415\]
UNIFORM DISTRIBUTION

\[ P(x) = \frac{1}{\max(x)} \text{ for } 0 < x < \max(x) \]

= 0 otherwise

Mean = min + max / 2

= 5.5

\[ SD^2 = \max - \min / 6 \]

= 1.5
BINOMIAL DISTRIBUTION

\[ P(n) = \frac{N!}{n!(N-n)!} \ p^n(1-p)^{N-n} \]

Mean = \( Np \)

\[ SD^2 = Np(1-p) = \text{Mean} \times (1-p) \]

**Random Pattern (p=0.5) #1's**

<table>
<thead>
<tr>
<th>Pattern</th>
<th>#1's</th>
</tr>
</thead>
<tbody>
<tr>
<td>011110000000000</td>
<td>3</td>
</tr>
<tr>
<td>0111110000010101</td>
<td>6</td>
</tr>
<tr>
<td>110010110000000</td>
<td>4</td>
</tr>
<tr>
<td>11100001000101</td>
<td>5</td>
</tr>
<tr>
<td>11111110101101</td>
<td>8</td>
</tr>
<tr>
<td>111011010000000</td>
<td>5</td>
</tr>
<tr>
<td>00010001111115</td>
<td>5</td>
</tr>
<tr>
<td>10100111100117</td>
<td>7</td>
</tr>
<tr>
<td>1010100001015</td>
<td>5</td>
</tr>
<tr>
<td>00101000011004</td>
<td>4</td>
</tr>
</tbody>
</table>

For \( N=30, \ p=0.5 \),

Mean = 15, \( SD^2 = 7.5 \)

For \( N=100, \ p = 0.15 \),

Mean = 15, \( SD^2 = 15 \times 0.85 \)

If \( p \) small, \( SD = \sqrt{\text{Count}} \)
NORMAL DISTRIBUTION

\[ p(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}} \]

Mean = 0
SD = 1

The Binomial distribution becomes like a Normal Distribution as the sample size increases.
**EXTREME VALUE DISTRIBUTIONS**

- **Normal:**
  \[ P(Z) = \exp(-Z^2) \]
  \[ \log_e(P(Z)) = -Z^2 \]

- **Extreme Value:**
  \[ P(Z) = \exp(-Z - \exp(-Z)) \]
  \[ \log_e(P(Z)) = -Z - \exp(-Z) \]

Where \( Z = (\text{score} - \text{mean}) / \text{SD} \)

The Extreme Value distribution has a long tail.
**EXPECTATION VALUES**

The shaded area is $P(Z > Z_0)$, the probability that a $Z$ score greater than $Z_0$ will occur by chance, i.e. The score $Z_0$ is not significant. This is the $P$-Value.

$Z_0$

\[ P(Z > Z_0) = 1 - \exp(-\exp(-Z_0)) \]

The expectation value, $E(Z > Z_0)$ is the expected number of errors. It is $E(Z > Z_0) = N_{db} P(Z > Z_0)$, where $N_{db}$ is number of queries.

©Michael Levitt 04
EXPECTATION VALUES

- Database searches use the expectation value, \( E(Z > Z_0) \), to indicate whether the score, \( S_0 \), is significant.

Note, \( Z_0 = \frac{(S_0 - \text{Mean})}{\text{SD}} \), where SD is the standard deviation.

- Typically one requires that the expectation value be less than \( 10^{-20} \) for a sequence search and less than \( 10^{-4} \) for a structural match.

- The expectation value depends on the size of the database: A score of 100 might be best when there are 1000 comparisons but would be much less good when there are 1,000,000 comparisons.
Data Visualization
Concept 6.3
DATA VISUALIZATION

Hierarchical Clustering.
K-Means Clustering.
ROC Curves.
Views of Structure Space.
Multidimensional Scaling.
INTUITIVE CLUSTERING

- Many possibilities.
- Not so easy.
- What is best clustering?

- Clustering seems easy and intuitive but it is actually very hard. Is there a solution?
HIERARCHICAL CLUSTERING

- Link the closest pairs. Keep going until no more close pairs.
- Single linkage clustering. Bad as can have distant members in same cluster.
K-MEANS CLUSTERING

- Select K points at random.
- Associate all points with K point nearest it.
- Calculate a new mid point (K).
- Repeat till no change.

This can fail badly if some regions are very dense.
ROC ANALYSIS

(Receiver Operating Characteristic)

- Can method sort items as correct and incorrect (positive or negative)?

- Compare the results with a “Gold Standard”.

- Predict all below threshold score to be positive. Increase threshold.
**TRUE OR FALSE**

- Have a "Gold Standard" classification that indicates the **Positive** and **Negative** outcomes.
- Predict which outcomes are **Positive** or **Negative**.

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Predict Positive; this is True.</td>
<td>Predict Positive; this is False.</td>
</tr>
<tr>
<td>Negative</td>
<td>Predict Negative; this is False.</td>
<td>Predict Negative; this is True.</td>
</tr>
</tbody>
</table>
**ROC CURVE EXAMPLES**

- The best classification has the largest area under the curve.
- Very sensitive to errors in the “gold standard” classification.
VIEWS OF STRUCTURE SPACE

(A:) Chimeric hemoglobin beta-alpha [Synthetic, based on ]

How can one see 5000 structure space?

Erik Sandelin

Cluster center. Size indicates cluster size
Cluster member
Cluster center of this cluster
VIEWS OF STRUCTURE SPACE

(A:) Mating type protein A1 Homeodomain {Baker's yeast (Saccharomyces cerevisiae)}

Highly similar: $P < -5$ or $Z > 5$
Similar: $P < -3$ or $Z > 4$

- Cluster center. Size indicates cluster size
- Cluster member
- Cluster center of this cluster

(A:) Aspartate receptor, ligand-binding domain {Salmonella typhimurium}

Make distorted 2-D view that changes as you move over it.

©Michael Levitt 04
Databases
Concept 6.4
DATABASES

Protein Data-Base (PDB).
RNA Database.
Membrane Database.
Small Molecule Databases.
Pathway Databases.
PROTEIN DATABASE RCSB

http://www.rcsb.org/pdb

6,400 Backward Links
PROTEIN DATABASE PDBLITE

www.pdblite.org

PDB Lite for Searching and Downloading Macromolecules

PDB Lite is designed for nonspecialists who search for atomic coordinate ("PDB") files at the Protein Data Bank on an occasional basis. It is especially targeted towards students and educators. See PDB Lite: What and Why? See also Nature of 3D Structural Data.

These sites were last tested, and these lists updated on 08/12/2003.

PDB Lite is available, including direct links to Protein Explorer 2 Beta, updated weekly with new entries from RCSB, from:

- Australia (The Walter and Eliza Hall Institute of Medical Research, Melbourne)
- India (Bioinformatics Centre, University of Pune)
- Israel (Bioinformatics, Weizmann Institute of Science, Rehovot)
- Israel (Tel Aviv University)
- Poland (Interdisciplinary Centre for Modeling, Warsaw University)
- United Kingdom (Cambridge Crystallographic Data Centre)
- United Kingdom (EMBL Outstation, European Bioinformatics Institute, Hinxton)
- USA (BioMolecular Engineering Research Center, Boston U)

For advanced searches, see Jaime Prilusky's as an alternative to SearchFields at RCSB. OCA can find some things that SearchFields cannot. For example, OCA has query fields for Kingdom, Gene, Disease, and Function. On the other hand, SearchFields can find some things more easily than OCA. For example it can limit searches to entries that contain coordinates for RNA but neither protein nor DNA, and it can find "phospholipase C" while OCA ignores the "C".

The above PDB Lite sites were keeping their databases up to date with new releases when checked on 08/12/2003. Other former mirror sites of the former Protein Data Bank (PDB) at Brookhaven National Laboratory (now closed) or OCA mirror sites were out of date, or were no longer offering PDB Lite (at least in fully-functional form).
RNA STRUCTURE DATABASE

RNABase.org
The RNA Structure Database

Listing of RNABase Entries
- Complete
  - A listing of all entries in RNABase with links to detailed records for each entry.
  - All Entries
- Technique
  - A listing of all entries in RNABase by experimental technique.
    - x-ray crystallography
    - NMR spectroscopy
    - all other methods
- Category
  - A listing of all entries in RNABase by structural or functional category.
    - transfer RNAs
    - ribosomal RNAs
    - messenger RNAs
    - transcription-related RNAs
    - introns
    - splicing-related RNAs
    - signal recognition particle RNAs
    - ribozymes
    - RNase P
    - aptamers
    - pseudoknots
    - tetraloops
    - bulges
    - DNA-RNA hybrids
    - PNA-RNA hybrids
    - drug-RNA complexes
    - viral & phage RNAs
- Outlier Rate
  - A tabulation of error rate for each structure in RNABase organized by technique and category.
  - All Entries

Search RNABase
- Basic Search
  - Find the entry you are looking for by PDB or NDB code, author name, classification, experimental technique, resolution, or keywords.
- Advanced Search
  - For those seeking more precise search capabilities.

http://www.rnabase.org
ORGANIC MOLECULE STRUCTURES

Library of 3-D Molecular Structures

Molecules of the Month

- Mar 2002 - Astemizole

If you are using CosmaPlayer click on the above image of a water dimer

[ About the Database ]

To enter the library, click on the appropriate buttons below:

Water and Ice
Carbons
Hydrocarbons
Molecules of Life
Drugs

http://www.chem.ox.ac.uk/mom

http://www.nyu.edu/pages/mathmol/library

212 Backward Links
52 Backward Links
SMALL MOLECULE DATABASE

NIST Chemistry WebBook

NIST Standard Reference Database Number 69 - March, 2003 Release

View: Search Options, Models and Tools, Documentation, Notes

Show Credits

NIST reserves the right to charge for access to this database in the future.

---

Search Options top

- General Searches
  - Formula
  - Name
  - CAS registry number
  - Reaction
  - Author
  - Structure

- Physical Property Based Searches
  - Ion energetics properties
  - Vibrational and electronic energies
  - Molecular weight

---

Models and Tools top

- Thermophysical Properties of Fluid Systems High accuracy data for a select group of fluids.
- Group Additivity Based Estimates Estimates of gas phase thermodynamic properties based on a submitted structure.

---

Documentation top

- Frequently asked questions

---

1,210 Backward Links

http://webbook.nist.gov/chemistry
PATHWAY DATABASES

Signaling Pathway Database

The Signaling Pathway Database (SPAD) is an integrated database for genetic information and signal transduction systems.

There are multiple signal transduction pathways: cascade of information from plasma membrane to nucleus in response to an extracellular stimulus in living organisms. Extracellular signal molecule binds specific intracellular receptor, and initiates the signaling pathway. Now, there is a large amount of information about the signaling pathway which controls the gene expression and cellular proliferation. We have developed an integrated database SPAD to understand the overview of signaling transduction. SPAD is divided into four categories based on extracellular signal molecules (Growth factor, Cytokine, and Hormone) and stress, that initiate the intracellular signaling pathway. SPAD is compiled in order to describe information on interaction between protein and protein, protein and DNA as well as information on sequences of DNA and proteins.

There are two methods for retrieving this database. Please select one of the two items.

- Extracellular Signal Molecules
  
  This WWW service 'SPAD' is still under development.

mail to: archive@  
Molecular Gene Techniques  
Hakozaki Higashi-ku,  
Fukuoka, 812-8581, Japan  
Graduate School of Genetic Resources Technology  
Kyushu University

Last Update Oct 13, 1998

http://www.grt.kyushu-u.ac.jp/spad/pathway/pdgf.html

87 Backward Links
PATHWAY DATABASES

KEGG - Table of Contents

1. KEGG Databases

<table>
<thead>
<tr>
<th>Category</th>
<th>Database</th>
<th>Search &amp; Compute</th>
<th>DBGET Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pathway information</td>
<td>KEGG PATHWAY Database</td>
<td>XML Search objects in KEGG pathways</td>
<td>DBGET PATHWAY</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Color objects in KEGG pathways Generate possible reaction paths</td>
<td></td>
</tr>
<tr>
<td>Genomic information</td>
<td>KEGG GENES Database</td>
<td>KO Search similar GENES sequences Search similar GENOME sequences</td>
<td>DBGET KO GENOME</td>
</tr>
<tr>
<td>Chemical information</td>
<td>KEGG LIGAND Database</td>
<td>RC Search similar compound structures Search similar glycan structures Search similar reactions</td>
<td>DBGET RC LIGAND ENZYME</td>
</tr>
</tbody>
</table>

2. KEGG Gene Catalogs

2.1 Genomes in KEGG

<table>
<thead>
<tr>
<th>Category</th>
<th>Genome</th>
<th>DBGET Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organism</td>
<td>Complete genomes in KEGG Complete genomes (taxonomy)</td>
<td>GENES</td>
</tr>
<tr>
<td>Virus</td>
<td>Complete viral genomes</td>
<td>VGENES</td>
</tr>
<tr>
<td>Organelle</td>
<td>Complete mitochondrial genomes Complete plastid genomes Complete nucleomorph genomes</td>
<td>OGENES</td>
</tr>
</tbody>
</table>

Caprolactam degradation - Reference pathway

Web Resources
Concept 6.5
WEB RESOURCES

EBI (European Bioinformatics Institute).

NCBI (National Center for Biotechnology Information).

NCBI for Sequences.

EBI for Tools.
ENSEMBL EBI SANGER GENOME VIEWER

http://www.ensembl.org

Current Release 19.34a.1
This release is based on the NCBI 34 assembly of the human genome.
View the status history of the human assemblies.
Last Update: 08-01-2004

Ensembl gene predictions: 23531
(incl. 1744 pseudogenes)
Genscan gene predictions: 65010
Ensembl gene exons: 225897
Ensembl gene transcripts: 31609
Contigs: 26614
Clones: 26614
Base Pairs: 3201762515
Golden Path Length: 2841366484

@Michael Levitt '04
ENSEMBL EBI SANGER GENOME VIEWER

- Huge amount of data
- Amazing detail
NCBI GENOME VIEWER

Homo sapiens Map View
Build 34 Version 2
BLAST The Human Genome
Chromosome: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X [Y]
Master Map: Genes On Sequence
Total Genes On Chromosome: 247
Region Displayed: 8,877K-8,977K bp
Download/View Sequence/Evidence
Genes Labeled: 5 Total Genes in Region: 5

Symbol LinkOut Cyto Description
LOC392580 sv pr dl ev mm Y Yp11.2 testis specific protein, Y-linked

Jump from Ensemble
Same position on Y.
NCBI GENOME VIEWER

Start with Chromosome 17

It has 1359 genes

NCBI IS THE PLACE FOR SEQUENCE

- PubMed Central (book text)
- Entrez
- Map Viewer
- 17,300 Backward Links

EBI IS THE PLACE FOR TOOLS

- FASTA
- ClustalW
- Expression Profiler
- DALI
- 7,200 Backward Links

http://www.ebi.ac.uk/services/index.html
**Sequence Comparison Concept 6.6**
SEQUENCE COMPARISON

Ungapped Comparison.
Gapped Comparison.
Scores and Penalties.
Advanced Methods.
IDENTICAL COMPARISON

Sequence A: YGTPWRSAAAQ
Sequence B: YGTPWRSAAAQ

- Assume that score is 1 for a match and 0 otherwise.
- Mark all the matches.
IDENTICAL TRACE

Sequence A: Y G T P W R S A A Q
Sequence B: Y G T P W R S A A Q

- Start at the top left and move down the diagonal from high-scoring match to high-scoring match.
- Add the scores along the path.
- Find the path that goes from top left to bottom right that collects the highest score.

Total score is 10
MISMATCH COMPARISON

Sequence A: YGTPWRSAAQ
Sequence B: YGPTWRSAQA

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>T</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>P</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>W</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>R</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>S</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Q</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Similarity Matrix

- The sequences are the same length.
- There are four differences or mismatches.
MISMATCH TRACE

Sequence A: YGTPWRSAAQ
Sequence B: YGPTWRSAAQA

- As the gap is two positions wide, it costs more than a single width gap.
- Assume the gap cost is 1.
- Include it in the count after skipping the gap.

Total score is 5.
DELETION COMPARISON

Sequence A: YGTPWRSAAQ
Sequence B: YGTWRSAAAQ

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>G</th>
<th>T</th>
<th>W</th>
<th>R</th>
<th>S</th>
<th>A</th>
<th>A</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</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></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</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></td>
<td></td>
<td></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></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></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></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Sequence B is now shorter than sequence A.
DELETION TRACE

Sequence A: Y G T P W R S A A Q
Sequence B: Y G T W R S A A Q

- Assume that the gap costs 0.5.
- Include this cost in the total score after the crossing the gap.

Sum Matrix

Total score is 8.5.
**DELETION/INSERTION COMPARISON**

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>G</th>
<th>W</th>
<th>R</th>
<th>S</th>
<th>Y</th>
<th>G</th>
<th>A</th>
<th>A</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Q</strong></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><strong>A</strong></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><strong>S</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>R</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>W</strong></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><strong>G</strong></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><strong>Y</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>P</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>T</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>G</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>A</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>A</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Q</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Deletion and insertion in B**

- The two sequences are the same length.
- There appears to be a long mis-match but it is really a deletion followed by an insertion.
DELETION/INSERTION TRACE

Sequence A: Y G T P W R S − − A A Q
Sequence B: Y G − − W R S Y G A A Q

Deletion and insertion in B

- Each gap skips over two positions and is assumed to cost 1.
- Add this cost into the score after crossing each gap.

Total score is 6.
**Scoring Matrix**

Matrix entry is the score of amino acid i matched to amino j.

- **Symmetric.**
- **Strongest matches:**
  
  \( C..C = 9 \)
  
  \( W..W = 11 \)

- **Weakest matches:**
  
  \( C..E = -4 \)
  
  \( W..D = -4 \)

- **Similar matches:**
  
  \( F..Y = 3 \)
  
  \( K..R = 2 \)
  
  \( D..E = 2 \)
  
  \( I..L = 2 \)

---

**BLOSUM62 score matrix (Henikoff)**

|   | A | B | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | X | Y | Z |
| A | 4 | -2 | 0 | -2 | -1 | -2 | 0 | -2 | -1 | -2 | 1 | 0 | 0 | -3 | -1 | -2 | -1 |   |   |   |   |   |   |   |
| B | -2 | 6 | -3 | 6 | 2 | -3 | 1 | -1 | 0 | -2 | 0 | -1 | -3 | 4 | -1 | -3 | 2 |   |   |   |   |   |   |   |
| C | 0 | -3 | -3 | -2 | -3 | -1 | -3 | -1 | -3 | -3 | -3 | -1 | -1 | -1 | -2 | -1 | -2 | -4 |   |   |   |   |   |   |   |
| D | -2 | 6 | -3 | 6 | 2 | -3 | 1 | -1 | 0 | -2 | 0 | -1 | -3 | 4 | -1 | -3 | 2 |   |   |   |   |   |   |   |
| E | -1 | 2 | -4 | 2 | 5 | -3 | -2 | 0 | -1 | 2 | 0 | -1 | -2 | -3 | -1 | -2 | 5 |   |   |   |   |   |   |   |
| F | -2 | 6 | -3 | 6 | 2 | -3 | 1 | -1 | 0 | -2 | 0 | -1 | -3 | 4 | -1 | -3 | 2 |   |   |   |   |   |   |   |
| G | 0 | -3 | -1 | -2 | -3 | 6 | -2 | 4 | -2 | 4 | -3 | 0 | -2 | -2 | 2 | 0 | -3 | 3 | -1 |   |   |   |   |   |
| H | -2 | 1 | -1 | -3 | -1 | 0 | -1 | 2 | 8 | 3 | -3 | 1 | 1 | -1 | -2 | 1 | -2 | 0 |   |   |   |   |   |   |
| I | -2 | 1 | -3 | -1 | 0 | -1 | 2 | 8 | 3 | -3 | 1 | 1 | -1 | -2 | 1 | -2 | 0 |   |   |   |   |   |   |
| J | 0 | -3 | -3 | -3 | -2 | 0 | -3 | 4 | -3 | -2 | 2 | 1 | 3 | -3 | 1 | -1 | 1 | -3 |   |   |   |   |   |   |
| K | -2 | 1 | -3 | 1 | 1 | 3 | -1 | 2 | 3 | 1 | 1 | 3 | -1 | 1 | 1 | 1 | 1 | 1 |   |   |   |   |   |   |
| L | -1 | 4 | -4 | -3 | 0 | -4 | -3 | 2 | 2 | 4 | 2 | -3 | -2 | -2 | -1 | 1 | -2 | 1 |   |   |   |   |   |   |
| M | -1 | 3 | -1 | -3 | 1 | 0 | -3 | 2 | 1 | 1 | 2 | 5 | 2 | -2 | 0 | -3 | 1 | -1 | 1 | -1 | -1 |   |
| N | -2 | 1 | -1 | -3 | 1 | 3 | 0 | -3 | 2 | 6 | 2 | 0 | 0 | 0 | 1 | 0 | -3 | 4 | -1 | -2 | 0 |   |
| O | -2 | 1 | -3 | 0 | -3 | 2 | 0 | -3 | 1 | 2 | 0 | 1 | 2 | 1 | 2 | 0 | 1 | -2 | -1 | -1 | 1 | -2 |
| P | -1 | 2 | -3 | 2 | 0 | -3 | 2 | 0 | 3 | 2 | 2 | -2 | -1 | 1 | 0 | -2 | 1 | 5 | 1 | -1 | -3 | -1 | -2 |
| Q | 0 | -3 | -1 | -1 | 1 | -2 | 2 | 7 | 1 | -2 | 1 | 2 | 4 | -1 | -3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| R | -1 | 2 | -3 | 2 | 0 | -3 | 2 | 0 | 3 | 2 | 2 | -2 | -1 | 1 | 0 | -2 | 1 | 5 | 1 | -1 | -3 | -1 | -2 |
| S | 1 | 0 | -1 | 0 | 0 | -2 | 0 | -2 | 0 | 2 | 1 | 2 | 2 | -2 | 0 | 1 | 1 | 4 | 1 | 2 | 3 | -1 | -2 |
| T | 0 | -3 | -1 | -3 | -1 | 2 | 1 | 1 | 3 | -1 | -2 | 2 | 1 | -1 | 1 | 1 | 5 | 0 | -2 | -1 | -2 | 1 |
| U | 0 | -2 | -3 | -1 | -3 | 3 | 2 | 1 | 1 | 3 | -2 | 3 | 2 | 0 | -4 | 3 | -1 | -1 | 1 | 2 | 3 | 1 | 1 |
| V | 0 | -3 | -2 | -4 | -3 | 1 | 2 | 2 | -3 | -2 | -1 | 0 | 4 | 4 | -3 | 1 | -1 | 1 | 2 | 1 | -2 | 1 | 1 |
| W | 0 | -3 | -2 | -4 | -3 | 1 | 2 | 2 | -3 | -2 | -1 | 0 | 4 | 4 | -3 | 1 | -1 | 1 | 2 | 1 | -2 | 1 | 1 |
| X | -2 | 3 | -2 | -3 | 3 | 2 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Y | -3 | 3 | -2 | -3 | 3 | 2 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Z | -1 | 2 | -4 | 2 | 5 | -3 | -2 | 0 | -1 | 2 | 0 | 0 | 1 | -2 | 3 | -1 | 2 | 5 |   |   |   |   |   |   |
GAP PENALTIES

The gap penalty is usually proportional to the number of positions skipped as follows:

<table>
<thead>
<tr>
<th>GAP</th>
<th>PENALTY</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>a</td>
</tr>
<tr>
<td>2</td>
<td>a + b</td>
</tr>
<tr>
<td>3</td>
<td>a + 2b</td>
</tr>
<tr>
<td>n</td>
<td>a + (n-1)b</td>
</tr>
</tbody>
</table>

- The optimum gap value depends on the Scoring Matrix.
- For the Blosum62 matrix, the best values are: 
  \[ a = -10, \quad b = -2 \]
  
Non-linear gaps are possible but make calculation much slower.
DYNAMIC PROGRAMMING

Best To I,J = Score at I,J + Best in block + Cost of gaps.

Similarity Matrix
DYNAMIC PROGRAMMING

In local alignment, score can never be less than 0.

<table>
<thead>
<tr>
<th>Y</th>
<th>G</th>
<th>W</th>
<th>R</th>
<th>S</th>
<th>Y</th>
<th>G</th>
<th>A</th>
<th>A</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>A</td>
<td>S</td>
<td>R</td>
<td>W</td>
<td>P</td>
<td>T</td>
<td>G</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Similarity Matrix**

<table>
<thead>
<tr>
<th>Y</th>
<th>G</th>
<th>W</th>
<th>R</th>
<th>S</th>
<th>Y</th>
<th>G</th>
<th>A</th>
<th>A</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.5</td>
<td>2</td>
<td>1.5</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1.5</td>
<td>2</td>
<td>1.5</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1.5</td>
<td>2</td>
<td></td>
<td>2</td>
<td></td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

**Sum Matrix**
DYNAMIC PROGRAMMING

\[
\text{BestTo}[i][j] = \text{Score}[i][j] + \max \{ \text{BestTo}[i][j] + \text{gap}_{IJ \to ij} \}
\]

where \( \text{gap}_{IJ \to ij} = \max\{|I-i-1,J-j-1\}/2 \), say.

**Similarity Matrix**

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>G</th>
<th>W</th>
<th>R</th>
<th>S</th>
<th>Y</th>
<th>G</th>
<th>A</th>
<th>A</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</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></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></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>1</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>1</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></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></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Sum Matrix**

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>G</th>
<th>W</th>
<th>R</th>
<th>S</th>
<th>Y</th>
<th>G</th>
<th>A</th>
<th>A</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>5</td>
<td>2</td>
<td>1.5</td>
<td>1</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1.5</td>
<td>2</td>
<td>1.5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>1.5</td>
<td>2</td>
<td>1.5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>1.5</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
ADVANCED METHODS

• Dynamic programming is slow as one needs to calculate a score for every cell of the similarity and score matrix. This is $O(n^2)$ at best.

• It can be speeded up by only looking in regions of the similarity matrix where there are high scores.
**FASTA**

Look for identities of single amino acids or pairs.

Mark every single identity.

Sum scores along diagonals with identities.

**Similarity Matrix**

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>G</th>
<th>W</th>
<th>R</th>
<th>S</th>
<th>Y</th>
<th>G</th>
<th>A</th>
<th>A</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</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></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</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>R</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>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></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></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</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>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

**Sum Matrix**

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>G</th>
<th>W</th>
<th>R</th>
<th>S</th>
<th>Y</th>
<th>G</th>
<th>A</th>
<th>A</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</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></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</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>R</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>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></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></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</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>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

*Pearson & Lipman, PNAS, 85, 2444 (1988).*
BLAST AND PSI-BLAST

- Look for triplets that have high match scores. For example with Blosum62: Y K D is a good match to F R E with a score of 7.

- Mark these on the Similarity Matrix.

- Extend these diagonals in the Sum Matrix.

- Merge separate fragments.

PSI-BLAST works with a profile rather than a sequence (more later). It is very, very clever.
BLAST: COMPARE SEQUENCES

Blast is an amazing resource. Play with it. This is the only way to learn.

THE END

of Lecture 6