Protein sequence alignment and evolution

Tuesday, April 5, 2005

Protein Bioinformatics
260.841
Jonathan Pevsner
pevsner@jhmi.edu

Outline: entire course

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Instructor</th>
</tr>
</thead>
<tbody>
<tr>
<td>T Mar. 29</td>
<td>Introduction to physical properties of amino acids</td>
<td>Prigge</td>
</tr>
<tr>
<td>Th Mar. 31</td>
<td>Protein Structure (level of Branden and Tooze)</td>
<td>Prigge</td>
</tr>
<tr>
<td>T Apr. 5</td>
<td>Protein sequence alignment and evolution</td>
<td>Pevsner</td>
</tr>
<tr>
<td>Th Apr. 7</td>
<td>Principles of mass spectrometry</td>
<td>Cotter</td>
</tr>
<tr>
<td>T Apr. 12</td>
<td>Applications of mass spectrometry to proteomics</td>
<td>Pandey</td>
</tr>
<tr>
<td>Th Apr. 14</td>
<td>Applications of mass spectrometry to proteomics</td>
<td>Pandey</td>
</tr>
<tr>
<td>T Apr. 19</td>
<td>Protein structure determination</td>
<td>Prigge</td>
</tr>
<tr>
<td>Th Apr. 21</td>
<td>Protein databases, structural classification of proteins, visualization</td>
<td>Ruczinski</td>
</tr>
<tr>
<td>T Apr. 26</td>
<td>Protein secondary structure prediction</td>
<td>Ruczinski</td>
</tr>
<tr>
<td>Th Apr. 28</td>
<td>Protein structure prediction</td>
<td>Ruczinski</td>
</tr>
<tr>
<td>T May 3</td>
<td>Protein structure prediction (CASP)</td>
<td>Ruczinski</td>
</tr>
<tr>
<td>Th May 5</td>
<td>Protein networks</td>
<td>Bader</td>
</tr>
<tr>
<td>T May 10</td>
<td>To be announced</td>
<td>Gray</td>
</tr>
<tr>
<td>Th May 12</td>
<td>Protein-protein docking</td>
<td></td>
</tr>
<tr>
<td>T May 17</td>
<td>To be announced</td>
<td></td>
</tr>
<tr>
<td>Th May 19</td>
<td>Final exam</td>
<td></td>
</tr>
</tbody>
</table>
Outline: entire course

T Mar. 29  Introduction to physical properties of amino acids  Prigge
Th Mar. 31  Protein Structure (level of Branden and Tooze)  Prigge

T Apr. 5  Protein sequence alignment and evolution  Peysner
Th Apr. 7  Principles of mass spectrometry  Cotter

T Apr. 12  Applications of mass spectrometry to proteomics  Pandey
Th Apr. 14  Applications of mass spectrometry to proteomics  Pandey

T Apr. 19  Protein structure determination  Prigge
Th Apr. 21  Protein databases, structural classification of proteins, visualization  Ruczinski

T Apr. 26  Protein secondary structure prediction  Ruczinski
Th Apr. 28  Protein structure prediction  Ruczinski

T May 3  Protein structure prediction (CASP)  Ruczinski
Th May 5  Protein networks  Bader

T May 10  High throughput approaches to proteomics  Boeke
Th May 12  Protein-protein docking  Gray

T May 17  Lab
Th May 19  Final exam

Outline: today’s topic

1. How to access the sequence and structure of a protein at NCBI and the Protein Data Bank (PDB)

2. Overview of databases of all proteins: NCBI and SwissProt

3. How to align the sequences of two proteins: Dayhoff’s evolutionary perspective

4. How to align the sequences of two proteins: pairwise alignment
Many of the powerpoints for today’s lecture are from *Bioinformatics and Functional Genomics* (J. Pevsner, 2003). The powerpoints are available on-line at www.bioinfbook.org

Chapter 2: Access to sequence data
Chapter 3: Pairwise sequence alignment
Chapter 4: Basic Local Alignment Search Tool (BLAST)
Chapter 8: Protein analysis and proteomics
Chapter 9: Protein structure

Outline: today’s topic

1. How to access the sequence and structure of a protein at NCBI and the Protein Data Bank (PDB)

2. Overview of databases of all proteins: NCBI and SwissProt

3. How to align the sequences of two proteins: Dayhoff’s evolutionary perspective

4. How to align the sequences of two proteins: pairwise alignment
### ExPASy Proteomics Server

ExPASy is a web server that provides tools for protein and peptide research. It includes databases and tools for sequence analysis, protein function, and structural information. The server is part of the ExPASy platform, which is hosted by the Swiss Institute of Bioinformatics (SIB).

#### Databases
- Swiss-Protein and TrEMBL - Protein knowledgebase
- PROSITE - Protein families and domains
- SWISS-2DPAGE - Two-dimensional polyacrylamide gel electrophoresis
- ENZYME - Enzyme nomenclature
- SWISS-3DIMAGE - 3D images of proteins and other biological macromolecules
- SWISS-MODEL Repository - Automatically generated protein models
- GeneOnLine - Knowledgebase on gene cell differentiation
- Ashbya Genome Database
- Links to many other molecular biology databases

#### Tools and software packages
- Proteomics and sequence analysis tools
  - ProteinBLAST
  - DNA-Protein
  - Similarity search
  - Neuroblastoma
  - Post-translational modification and topology prediction
  - Primary structure analysis
  - Secondary and tertiary structure prediction
  - Protein structure analysis
  - Superfamily

#### Search in Swiss-Protein and TrEMBL for: amyloid

**Swiss-Protein Release 46.4 of 29-Mar-2005**
**TrEMBL Release 29.4 of 29-Mar-2005**

- Number of sequences found in Swiss-Protein and TrEMBL: 319
- Note that the selected sequences can be saved to a file to be later retrieved; to do so, go to the bottom of this page.
- For more detailed searches, you can use the Sequence Retrieval System (SRS).

#### Search in Swiss-Protein: There are matches to 103 out of 178022 entries

<table>
<thead>
<tr>
<th>Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD-BOVIN (Q19935)</td>
</tr>
<tr>
<td>AD-CABER (Q19936)</td>
</tr>
<tr>
<td>AD-CANEFA (Q19938)</td>
</tr>
<tr>
<td>AD-CAYTO (Q19939)</td>
</tr>
<tr>
<td>AD-TROME (P14599)</td>
</tr>
<tr>
<td>AD-FOOZU (P19937)</td>
</tr>
<tr>
<td>AD-HUMAIA (P19936)</td>
</tr>
</tbody>
</table>

Additional entries are available for further analysis.
Central dogma of molecular biology

DNA → RNA → protein

Central dogma of bioinformatics and genomics

genome → transcriptome → proteome
Accession numbers are labels for sequences

NCBI includes databases (such as GenBank) that contain information on DNA, RNA, or protein sequences. You may want to acquire information beginning with a query such as the name of a protein of interest, or the raw nucleotides comprising a DNA sequence of interest.

DNA sequences and other molecular data are tagged with accession numbers that are used to identify a sequence or other record relevant to molecular data.

What is an accession number?

An accession number is a label that used to identify a sequence. It is a string of letters and/or numbers that corresponds to a molecular sequence.

Examples (all for retinol-binding protein, RBP4):

<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>X02775</td>
<td>GenBank genomic DNA sequence</td>
<td>DNA</td>
</tr>
<tr>
<td>NT_030059</td>
<td>Genomic contig</td>
<td>DNA</td>
</tr>
<tr>
<td>Rs7079946</td>
<td>dbSNP (single nucleotide polymorphism)</td>
<td>DNA</td>
</tr>
<tr>
<td>N91759.1</td>
<td>An expressed sequence tag (1 of 170)</td>
<td>RNA</td>
</tr>
<tr>
<td>NM_006744</td>
<td>RefSeq DNA sequence (from a transcript)</td>
<td>RNA</td>
</tr>
<tr>
<td>NP_007635</td>
<td>RefSeq protein</td>
<td>protein</td>
</tr>
<tr>
<td>AAC02945</td>
<td>GenBank protein</td>
<td>protein</td>
</tr>
<tr>
<td>Q28369</td>
<td>SwissProt protein</td>
<td>protein</td>
</tr>
<tr>
<td>1KT7</td>
<td>Protein Data Bank structure record</td>
<td>protein</td>
</tr>
</tbody>
</table>
NCBI's important RefSeq project:
best representative sequences

RefSeq (accessible via the main page of NCBI) provides an expertly curated accession number that corresponds to the most stable, agreed-upon "reference" version of a sequence.

RefSeq identifiers include the following formats:

- Complete genome: NC_######
- Complete chromosome: NC ########
- Genomic contig: NT ########
- mRNA (DNA format): NM ######## e.g. NM_006744
- Protein: NP ######## e.g. NP_006735

Example: type "amyloid" at NCBI
Click “protein” to find 3419 records for amyloid. Further limit the search to RefSeq only, then to human.
Outline: today’s topic

1. How to access the sequence and structure of a protein at NCBI and the Protein Data Bank (PDB)

2. Overview of databases of all proteins: NCBI and SwissProt

3. How to align the sequences of two proteins: Dayhoff’s evolutionary perspective

4. How to align the sequences of two proteins: pairwise alignment
DNA → RNA → protein → phenotype

cDNA → ESTs → UniGene

DNA → RNA → protein → phenotype

genomic DNA databases → cDNA ESTs UniGene → protein sequence databases

Fig. 2.2
Page 20
**DNA**

<table>
<thead>
<tr>
<th>GenBank</th>
<th>ExPASy</th>
<th>DDBJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCBI</td>
<td>EMBL</td>
<td>DDBJ</td>
</tr>
</tbody>
</table>

**protein**

<table>
<thead>
<tr>
<th>UniProt (<a href="http://www.uniprot.org">www.uniprot.org</a>)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBI</td>
</tr>
<tr>
<td>NCBI</td>
</tr>
<tr>
<td>Protein Data Bank</td>
</tr>
</tbody>
</table>

### Growth of GenBank

Release 146 (Feb 2005) has 46,849,831,226 base pairs

![Growth of GenBank](Fig. 2.1)
The most sequenced organisms in GenBank

<table>
<thead>
<tr>
<th>Organism</th>
<th>Base Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Homo sapiens</em></td>
<td>10.7 billion</td>
</tr>
<tr>
<td><em>Mus musculus</em></td>
<td>6.5b</td>
</tr>
<tr>
<td><em>Rattus norvegicus</em></td>
<td>5.6b</td>
</tr>
<tr>
<td><em>Danio rerio</em></td>
<td>1.7b</td>
</tr>
<tr>
<td><em>Zea mays</em></td>
<td>1.4b</td>
</tr>
<tr>
<td><em>Oryza sativa</em></td>
<td>0.8b</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>0.7b</td>
</tr>
<tr>
<td><em>Gallus gallus</em></td>
<td>0.5b</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>0.5b</td>
</tr>
</tbody>
</table>

Updated 8-12-04
GenBank release 142.0
UniProt (Universal Protein Resource) is the world’s most comprehensive catalog of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR.

UniProt is comprised of three components, each optimized for different uses. The UniProt Knowledgebase (UniProt) is the central access point for extensive curated protein information, including function, classification, and cross-reference. The UniProt Non-redundant Reference (Uniprot) databases combine closely related sequences into a single record to speed searches. The UniProt Archive (UniParc) is a comprehensive repository, reflecting the history of all protein sequences.

The sequences and information in UniProt are accessible via text search, BLAST similarity search, and FTP.

PDB content growth (www.pdb.org)
Outline: today’s topic

1. How to access the sequence and structure of a protein at NCBI and the Protein Data Bank (PDB)

2. Overview of databases of all proteins: NCBI and SwissProt

3. How to align the sequences of two proteins: Dayhoff’s evolutionary perspective

4. How to align the sequences of two proteins: pairwise alignment

Definitions

Signature:
- a protein category such as a domain or motif
Definitions

Signature:
• a protein category such as a domain or motif

Domain:
• a region of a protein that can adopt a 3D structure
• a fold
• a family is a group of proteins that share a domain
• examples: zinc finger domain
             immunoglobulin domain

Motif (or fingerprint):
• a short, conserved region of a protein
• typically 10 to 20 contiguous amino acid residues

15 most common domains (human)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zn finger, C2H2 type</td>
<td>1093</td>
</tr>
<tr>
<td>Immunoglobulin</td>
<td>1032</td>
</tr>
<tr>
<td>EGF-like</td>
<td>471</td>
</tr>
<tr>
<td>Zn-finger, RING</td>
<td>458</td>
</tr>
<tr>
<td>Homeobox</td>
<td>417</td>
</tr>
<tr>
<td>Pleckstrin-like</td>
<td>405</td>
</tr>
<tr>
<td>RNA-binding region RNP-1</td>
<td>400</td>
</tr>
<tr>
<td>SH3</td>
<td>394</td>
</tr>
<tr>
<td>Calcium-binding EF-hand</td>
<td>392</td>
</tr>
<tr>
<td>Fibronectin, type III</td>
<td>300</td>
</tr>
<tr>
<td>PDZ/DHR/GLGF</td>
<td>280</td>
</tr>
<tr>
<td>Small GTP-binding protein</td>
<td>261</td>
</tr>
<tr>
<td>BTB/POZ</td>
<td>236</td>
</tr>
<tr>
<td>bHLH</td>
<td>226</td>
</tr>
<tr>
<td>Cadherin</td>
<td>226</td>
</tr>
</tbody>
</table>

Source: Integr8 program at www.ebi.ac.uk/proteome/
Pairwise alignments in the 1950s

β-corticotropin (sheep)  
Corticotropin A (pig)  
ala gly glu asp asp glu  
asp gly ala glu asp glu

Oxytocin  
Vasopressin  
CYIQCPLG  
CYFQNCPRG

Early alignments revealed
--differences in amino acid sequences between species
--differences in amino acids responsible for distinct functions

Pairwise sequence alignment is the most fundamental operation of bioinformatics

• It is used to decide if two proteins (or genes) are related structurally or functionally
• It is used to identify domains or motifs that are shared between proteins
• It is the basis of BLAST searching
• It is used in the analysis of genomes
RBP and β-lactoglobulin are homologous proteins that share related three-dimensional structures.
Definitions

Pairwise alignment
The process of lining up two or more sequences to achieve maximal levels of identity (and conservation, in the case of amino acid sequences) for the purpose of assessing the degree of similarity and the possibility of homology.

Definitions

Homology
Similarity attributed to descent from a common ancestor.
Definitions

Homology
Similarity attributed to descent from a common ancestor.

Identity
The extent to which two (nucleotide or amino acid) sequences are invariant.

Orthologs
Homologous sequences in different species that arose from a common ancestral gene during speciation; may or may not be responsible for a similar function.

Paralogs
Homologous sequences within a single species that arose by gene duplication.
Orthologs: members of a gene (protein) family in various organisms.

This tree shows 13 RBP orthologs.

Paralogs: members of a gene (protein) family within a species.

This tree shows 9 human lipocalins.
Pairwise alignment of retinol-binding protein and β-lactoglobulin

1 MKWVWALLLLAANAAEEERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP
  . ||| | | . . | . . | . ||| :||| : :||| :||| :||| :||| :
1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin
51 LFLQDNIVAEFSVDETQMSATAKGRVR.LLNNWD..VCADMVGTFTDTE 97 RBP
 : | | | | : : | . | | | : || | ||
45 ISLLDAQSAPLRV.YVEELKPTPEGDILLQKWENGECQKKIIAEKTK 93 lactoglobulin
98 DPARKMKYWGVASFLQKGNHHIVDHTDYAV...........QYSC 136 RBP
 || || . : |||| | .
94 IPAVFKIDALNKVL.......VLDTDYKKLYLECMENSAEPEQSLAC 135 lactoglobulin
137 RLLNLGTDACSVSVFSRDPNLPEAQKIVRQRQ.EELCLARQYRLIV 185 RBP
 . | | | | : || . | || ||
136 QCLVRTPEVDDEALEKFDKALKPMHILSFNPQLEEQQCHI....... 178 lactoglobulin
Definitions

**Similarity**
The extent to which nucleotide or protein sequences are related. It is based upon identity plus conservation.

**Identity**
The extent to which two sequences are invariant.

**Conservation**
Changes at a specific position of an amino acid or (less commonly, DNA) sequence that preserve the physico-chemical properties of the original residue.

Pairwise alignment of retinol-binding protein and β-lactoglobulin

<table>
<thead>
<tr>
<th>1 MKWVWALLLAAANAAERDCRVSFRVKEFSDKARFSGTWYAMAKKDPEG</th>
<th>50 RBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ...MKCLLLALALTGAQALIVT..QTMKGLIDIQKVAGTWYSLAMAASD.</td>
<td>44 lactoglobulin</td>
</tr>
<tr>
<td>51 LFLQDNIVAEFSVDETQMQSATAKGRVR.LLLNWD..ADMVGTFTDTE</td>
<td>97 RBP</td>
</tr>
<tr>
<td>45 ISLDAOSAPLVR.YEE2KPTPEGDIEILLQKWENNCAQKIIAEKT</td>
<td>93 lactoglobulin</td>
</tr>
<tr>
<td>98 DPARKMKYWVSFLQKGNDDHNIVDDYDTYAV........QYSC</td>
<td>136 RBP</td>
</tr>
<tr>
<td>94 IPAVFKIDNENKVL........VLDTYKYYLLFSENSAEPEQSLAC</td>
<td>135 lactoglobulin</td>
</tr>
<tr>
<td>137 RLLNLGTDTSFSVFSRDPNLPPPEAQKIV........QYRLIV</td>
<td>185 RBP</td>
</tr>
<tr>
<td>136 QCLVTRPEVDEAELAEKFDKALKLPMHILSF........</td>
<td>178 lactoglobulin</td>
</tr>
</tbody>
</table>
Pairwise alignment of retinol-binding protein and β-lactoglobulin

Pairwise alignment

The process of lining up two or more sequences to achieve maximal levels of identity (and conservation, in the case of amino acid sequences) for the purpose of assessing the degree of similarity and the possibility of homology.

Definitions
Pairwise alignment of retinol-binding protein and β-lactoglobulin

1 MKWVWALLLAAANAAERDCRVSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP
  . ||| | . | . | :||:|||:
1 ...MKCLLALALTCAQLAVVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

51 LFLQDNIVAEFSVDQMSATAKGRVR.LLN WD..VCADMVGFTTDTE 97 RBP
  : | | | | :: :|| | || || |
45 ISLDAQSAPLRV.YVEELKFTPEDLEI LQKWENGEC AQQKII A EKT 93 lactoglobulin

98 DPAKFKMKYWGVSFLQGNDDHWIVD TDYAV.........QYSC 136 RBP
  || || | :: ||: ||| |
94 IPAVFKIDALNENKL.........VLDTDKYKLLFCMENSAEPEQSLAC 135 lactoglobulin

137 RLNLNDGTCDSYFSVLPLPEAK IVRQR EELCLARQYRLIV 185 RBP
  . | | : || | || ||
136 QCLVRTPFEVDDEALEKFKK KALPMH IR LSFNP TQLEEQCHI...... 178 lactoglobulin

Gaps

- Positions at which a letter is paired with a null are called gaps.

- Gap scores are typically negative.

- Since a single mutational event may cause the insertion or deletion of more than one residue, the presence of a gap is ascribed more significance than the length of the gap.

- In BLAST, it is rarely necessary to change gap values from the default.
Pairwise alignment of retinol-binding protein and β-lactoglobulin

Pairwise alignment of retinol-binding protein from human (top) and rainbow trout (*O. mykiss*)
Multiple sequence alignment of glyceraldehyde 3-phosphate dehydrogenases

fly       GAKKVIISAP SAD.APM..F VCGVNLDAYK PDMKVVSNAS CTTNCLAPLA
human    GAKRVIISAP SAD.APM..F VMGVNHEKYD NSLKIISNAS CTTNCLAPLA
plant     GAKKVIISAP SAD.APM..F VVGVNHTVQ PNMDIVSNAS CTTNCLAPLA
bacterium GAKKVVMTPF SKNTPF..F VKGANFDKY. AGQDIVSNAS CTTNCLAPLA
yeast     GAKKVVTAP SS.TAPM..F VMGVNEEKYT SDLKIVSNAS CTTNCLAPLA
archaeon  GADKVLISAP PKGDEPVQKL VYGVNHDEYD GE.DVSNAS CTNNSITPVA

fly       KVINDNFEIV EGLMTTVHAT TATQKTVDPG SGKLWRDGRG AAQNIIPAST
human    KVIHDNFGIV EGLMTTVHAI TATQKTVDPG SGKLWRDGRG ALQNIIPAST
plant     KVIVEEFGIL EGLMTTVHAT TATQKTVDPG SMKDWRGGRG ASQNIIPSSST
bacterium KVINDNFGII EGLMTTVHAT TATQKTVDPG SHKDWRGGRG ASQNIIPSSST
yeast     KVIDAFGIE EGLMTTVHSL TATQKTVDPG SHKDWRGGRT ASQNIIPSSST
archaeon  KVLDEEFGIN AGQIHTTVHAY TGSQNLMDGP NGKP.RRRRA AEQNIIPSTST

fly       GAAKAVGKVI PALNGKLTGM AFRVPPTNVS VVDLTVRLGK GASYDEIKAK
human    GAAKAVGKVI PEINGKLTGM AFRVPPTANVS VVDLTCRLEK FAKYDIKKV
plant     GAAKAVGKVL PEINQKLTGM AFRVPTSNVS VVDLTCRLEK GASYEKIAAA
bacterium GAAKAVGKVL PEINOKLTGM AFRVPTNSVS VVDLTCRLEK AATTYQIKAA
yeast     GAAQAVGKVL PEIQQKLTGM AFRVPTVDVS VVDLTVKLHK ETYDEIKKV
archaeon  GAAQAAATEVL PELEGKLDGM AIRVPVPNGS ITEPVVSLDD DVTESDVNA

Outline: today’s topic

1. How to access the sequence and structure of a protein at NCBI and the Protein Data Bank (PDB)

2. Overview of databases of all proteins: NCBI and SwissProt

3. How to align the sequences of two proteins: Dayhoff’s evolutionary perspective

4. How to align the sequences of two proteins: pairwise alignment
An early substitution matrix from 1965

Zuckerkandl and Pauling aligned several dozen available globin protein sequences, and derived the following substitution matrix.

![Substitution matrix diagram]

Fig. 3.31
Dayhoff’s 34 protein superfamilies

Dayhoff and colleagues defined “accepted point mutation” (PAM) as a replacement of one amino acid by another residue that has been “accepted” by natural selection.

A PAM occurs when

[1] a gene undergoes a DNA mutation that changes the encoded amino acid

[2] the entire species adopts that change as the predominant form of the protein.

<table>
<thead>
<tr>
<th>Protein</th>
<th>PAMs per 100 million years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ig kappa chain</td>
<td>37</td>
</tr>
<tr>
<td>Kappa casein</td>
<td>33</td>
</tr>
<tr>
<td>Lactalbumin</td>
<td>27</td>
</tr>
<tr>
<td>Hemoglobin α</td>
<td>12</td>
</tr>
<tr>
<td>Myoglobin</td>
<td>8.9</td>
</tr>
<tr>
<td>Insulin</td>
<td>4.4</td>
</tr>
<tr>
<td>Histone H4</td>
<td>0.10</td>
</tr>
<tr>
<td>Ubiquitin</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Page 50
Dayhoff’s numbers of “accepted point mutations”: what amino acid substitutions occur in proteins?

|     | A | A | l | a | R | a | e | g | N | A | s | n | D | A | s | p | C | C | y | s | Q | G | l | n | E | G | l | u | G | l | y | A | R | 3 | 0 | N | 1 | 0 | 9 | 1 | 7 | D |

Dayhoff et al. examined multiple sequence alignments (e.g. glyceraldehyde 3-phosphate dehydrogenases) to generate tables of accepted point mutations

<table>
<thead>
<tr>
<th></th>
<th>fly</th>
<th>GACKVIISAP</th>
<th>SAD.APM..F</th>
<th>VCGVNLADAYK</th>
<th>PDMKVSNAS</th>
<th>CTTNCLAPLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>GAKRVIISAP</td>
<td>SAD.APM..F</td>
<td>VGVMNHEKYD</td>
<td>NSLKIISNAS</td>
<td>CTTNCLAPLA</td>
<td></td>
</tr>
<tr>
<td>plant</td>
<td>GAKVIIISAP</td>
<td>SAD.APM..F</td>
<td>VGVNEHTYQ</td>
<td>PNMDIVNAS</td>
<td>CTTNCLAPLA</td>
<td></td>
</tr>
<tr>
<td>bacterium</td>
<td>GAKKVMTGP</td>
<td>SKNTPM..F</td>
<td>VCGANFDKY.</td>
<td>AGQDIVNAS</td>
<td>CTTNCLAPLA</td>
<td></td>
</tr>
<tr>
<td>yeast</td>
<td>GAKKVITAP</td>
<td>SS.TAPM..F</td>
<td>VGMVEEKTYS</td>
<td>SDLKIVNAS</td>
<td>CTTNCLAPLA</td>
<td></td>
</tr>
<tr>
<td>archaeon</td>
<td>GADKVLISAP</td>
<td>PKGDVPQVKL</td>
<td>VGYVHDEYDG</td>
<td>GDVSNAS</td>
<td>CTTSITPVA</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>fly</th>
<th>KVINDNFEIV</th>
<th>EGLMTVHAT</th>
<th>TATQKVTDGQ</th>
<th>SGKLWRDGRG</th>
<th>AAQNIIPAST</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>KVHNDNGIV</td>
<td>EGLMTVHAI</td>
<td>TATQKVTDGQ</td>
<td>SGKLWRDGRG</td>
<td>ALQNIIPAST</td>
<td></td>
</tr>
<tr>
<td>plant</td>
<td>KUVHEEFGIL</td>
<td>ELMMTVHAT</td>
<td>TATQKVTDGQ</td>
<td>SMKDWRGRGG</td>
<td>ASQNIIPSST</td>
<td></td>
</tr>
<tr>
<td>bacterium</td>
<td>KVINDFPGII</td>
<td>EGLMTVHAT</td>
<td>TATQKVTDGQ</td>
<td>SHKDWRGRGG</td>
<td>ASQNIIPSST</td>
<td></td>
</tr>
<tr>
<td>yeast</td>
<td>KVINDAPGIE</td>
<td>ELMNVTLSL</td>
<td>TATQKVTDGQ</td>
<td>SHKDWRGGRG</td>
<td>ASQNIIPSST</td>
<td></td>
</tr>
<tr>
<td>archaeon</td>
<td>KVLDEFPGIN</td>
<td>AGQLTVTHAY</td>
<td>TCEQMLMDGP</td>
<td>NGKP.RRRRA</td>
<td>AAENIIPSTT</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>fly</th>
<th>GAAKAVKVL</th>
<th>FALNGKLTMG</th>
<th>AFRVPTPNSV</th>
<th>VVDLTVRLGK</th>
<th>GASYDEIKAK</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>GAAKAVGKVL</td>
<td>FELNGKLTMG</td>
<td>AFRVPTANVS</td>
<td>VVDLTCRLEK</td>
<td>PAKKYDIKKV</td>
<td></td>
</tr>
<tr>
<td>plant</td>
<td>GAAKAVGKVL</td>
<td>FELNGKLTMG</td>
<td>AFRVPTSNVS</td>
<td>VVDLTCRLEK</td>
<td>GASYEDVKA</td>
<td></td>
</tr>
<tr>
<td>bacterium</td>
<td>GAAKAVGKVL</td>
<td>FELNGKLTMG</td>
<td>AFRVPTPNSV</td>
<td>VVDLTVRLEK</td>
<td>AATYEQIKA</td>
<td></td>
</tr>
<tr>
<td>yeast</td>
<td>GAAKAVGKVL</td>
<td>FELQGLKTGM</td>
<td>AFRVPTVGDV</td>
<td>VVDLTVKLN K</td>
<td>ETTYDEIKKV</td>
<td></td>
</tr>
<tr>
<td>archaeon</td>
<td>GAAQAAEITVL</td>
<td>FELHKLDGM</td>
<td>AIRVVPVNGS</td>
<td>ITEFVVDLD</td>
<td>DVTESDVNA</td>
<td></td>
</tr>
</tbody>
</table>
Dayhoff et al. estimated the relative mutability of amino acids

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Asn</td>
<td>134</td>
<td>His</td>
</tr>
<tr>
<td>Ser</td>
<td>120</td>
<td>Arg</td>
</tr>
<tr>
<td>Asp</td>
<td>106</td>
<td>Lys</td>
</tr>
<tr>
<td>Glu</td>
<td>102</td>
<td>Pro</td>
</tr>
<tr>
<td>Ala</td>
<td>100</td>
<td>Gly</td>
</tr>
<tr>
<td>Thr</td>
<td>97</td>
<td>Tyr</td>
</tr>
<tr>
<td>Ile</td>
<td>96</td>
<td>Phe</td>
</tr>
<tr>
<td>Met</td>
<td>94</td>
<td>Leu</td>
</tr>
<tr>
<td>Gln</td>
<td>93</td>
<td>Cys</td>
</tr>
<tr>
<td>Val</td>
<td>74</td>
<td>Trp</td>
</tr>
</tbody>
</table>

Table 3.1

Page 53

Normalized frequencies of amino acids: variations in frequency of occurrence

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Gly</td>
<td>8.9%</td>
<td>Arg</td>
</tr>
<tr>
<td>Ala</td>
<td>8.7%</td>
<td>Asn</td>
</tr>
<tr>
<td>Leu</td>
<td>8.5%</td>
<td>Phe</td>
</tr>
<tr>
<td>Lys</td>
<td>8.1%</td>
<td>Gln</td>
</tr>
<tr>
<td>Ser</td>
<td>7.0%</td>
<td>Ile</td>
</tr>
<tr>
<td>Val</td>
<td>6.5%</td>
<td>His</td>
</tr>
<tr>
<td>Thr</td>
<td>5.8%</td>
<td>Cys</td>
</tr>
<tr>
<td>Pro</td>
<td>5.1%</td>
<td>Tyr</td>
</tr>
<tr>
<td>Glu</td>
<td>5.0%</td>
<td>Met</td>
</tr>
<tr>
<td>Asp</td>
<td>4.7%</td>
<td>Trp</td>
</tr>
</tbody>
</table>

blue=6 codons; red=1 codon

Page 53
Dayhoff's numbers of "accepted point mutations": what amino acid substitutions occur in proteins?

A|A|a|R|a|R|g|N|a|s|N|a|D|a|s|p|c|C|y|s|Q|I|n|E|G|l|u|G|I|y|A|R\^{30}N\^{10917}
Dayhoff’s PAM1 mutation probability matrix

- All the PAM data come from alignments of closely related proteins (>85% amino acid identity)

- PAM matrices are based on global sequence alignments.

- The PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence.

- Each element of the matrix shows the probability that an original amino acid (columns) will be replaced by another amino acid (rows) over an evolutionary interval.

- For the PAM1 matrix, that interval is 1% amino acid divergence; note that the interval is not in units of time.

Page 53
A substitution matrix contains values proportional to the probability that amino acid \( i \) mutates into amino acid \( j \) for all pairs of amino acids.

Substitution matrices are constructed by assembling a large and diverse sample of verified pairwise alignments (or multiple sequence alignments) of amino acids.

Substitution matrices should reflect the true probabilities of mutations occurring through a period of evolution.

The two major types of substitution matrices are PAM and BLOSUM.

---

**PAM matrices: Point-accepted mutations**

PAM matrices are based on global alignments of closely related proteins.

The PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence.

Other PAM matrices are extrapolated from PAM1.

All the PAM data come from closely related proteins (>85% amino acid identity)
PAM0 and PAM∞ mutation probability matrices

Consider a PAM0 matrix. No amino acids have changed, so the values on the diagonal are 100%.

Consider a PAM2000 (nearly infinite) matrix. The values approach the background frequencies of the amino acids (given in Table 3-2).

Dayhoff’s PAM1 mutation probability matrix

A A l a R a g N A s n D A s p C C y s Q G l n E G l u G G l y H H i s I I l e A
### Dayhoff’s PAM0 mutation probability matrix: the rules for extremely slowly evolving proteins

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>R</th>
<th>N</th>
<th>D</th>
<th>C</th>
<th>Q</th>
<th>E</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>100%</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>R</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>N</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>D</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>C</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Q</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>E</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>G</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Top: original amino acid  
Side: replacement amino acid  
Fig. 3.12  
Page 56

### Dayhoff’s PAM2000 mutation probability matrix: the rules for very distantly related proteins

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>R</th>
<th>N</th>
<th>D</th>
<th>C</th>
<th>Q</th>
<th>E</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>8.7%</td>
<td>8.7%</td>
<td>8.7%</td>
<td>8.7%</td>
<td>8.7%</td>
<td>8.7%</td>
<td>8.7%</td>
<td>8.7%</td>
</tr>
<tr>
<td>R</td>
<td>4.1%</td>
<td>4.1%</td>
<td>4.1%</td>
<td>4.1%</td>
<td>4.1%</td>
<td>4.1%</td>
<td>4.1%</td>
<td>4.1%</td>
</tr>
<tr>
<td>N</td>
<td>4.0%</td>
<td>4.0%</td>
<td>4.0%</td>
<td>4.0%</td>
<td>4.0%</td>
<td>4.0%</td>
<td>4.0%</td>
<td>4.0%</td>
</tr>
<tr>
<td>D</td>
<td>4.7%</td>
<td>4.7%</td>
<td>4.7%</td>
<td>4.7%</td>
<td>4.7%</td>
<td>4.7%</td>
<td>4.7%</td>
<td>4.7%</td>
</tr>
<tr>
<td>C</td>
<td>3.3%</td>
<td>3.3%</td>
<td>3.3%</td>
<td>3.3%</td>
<td>3.3%</td>
<td>3.3%</td>
<td>3.3%</td>
<td>3.3%</td>
</tr>
<tr>
<td>Q</td>
<td>3.8%</td>
<td>3.8%</td>
<td>3.8%</td>
<td>3.8%</td>
<td>3.8%</td>
<td>3.8%</td>
<td>3.8%</td>
<td>3.8%</td>
</tr>
<tr>
<td>E</td>
<td>5.0%</td>
<td>5.0%</td>
<td>5.0%</td>
<td>5.0%</td>
<td>5.0%</td>
<td>5.0%</td>
<td>5.0%</td>
<td>5.0%</td>
</tr>
<tr>
<td>G</td>
<td>8.9%</td>
<td>8.9%</td>
<td>8.9%</td>
<td>8.9%</td>
<td>8.9%</td>
<td>8.9%</td>
<td>8.9%</td>
<td>8.9%</td>
</tr>
</tbody>
</table>

Top: original amino acid  
Side: replacement amino acid  
Fig. 3.12  
Page 56
The PAM250 matrix is of particular interest because it corresponds to an evolutionary distance of about 20% amino acid identity (the approximate limit of detection for the comparison of most proteins).

Note the loss of information content along the main diagonal, relative to the PAM1 matrix.
Why do we go from a mutation probability matrix to a log odds matrix?

- We want a scoring matrix so that when we do a pairwise alignment (or a BLAST search) we know what score to assign to two aligned amino acid residues.

- Logarithms are easier to use for a scoring system. They allow us to sum the scores of aligned residues (rather than having to multiply them).
How do we go from a mutation probability matrix to a log odds matrix?

• The cells in a log odds matrix consist of an “odds ratio”:

   the probability that an alignment is authentic
   the probability that the alignment was random

The score $S$ for an alignment of residues $a,b$ is given by:

$$S(a,b) = 10 \log_{10} \left( \frac{M_{ab}}{p_b} \right)$$

As an example, for tryptophan,

$$S(a, \text{tryptophan}) = 10 \log_{10} \left( \frac{0.55}{0.010} \right) = 17.4$$

What do the numbers mean in a log odds matrix?

$$S(a, \text{tryptophan}) = 10 \log_{10} \left( \frac{0.55}{0.010} \right) = 17.4$$

A score of +17 for tryptophan means that this alignment is 50 times more likely than a chance alignment of two Trp residues.

$$S(a,b) = 17$$

Probability of replacement ($M_{ab}/p_b$) = $x$

Then

$$17 = 10 \log_{10} x$$

$$1.7 = \log_{10} x$$

$$10^{1.7} = x = 50$$
What do the numbers mean in a log odds matrix?

A score of +2 indicates that the amino acid replacement occurs 1.6 times as frequently as expected by chance.

A score of 0 is neutral.

A score of –10 indicates that the correspondence of two amino acids in an alignment that accurately represents homology (evolutionary descent) is one tenth as frequent as the chance alignment of these amino acids.

---

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

PAM250 log odds scoring matrix

---

Fig. 3.14

Page 58
PAM10 log odds scoring matrix

Note that penalties for mismatches are far more severe than for PAM250; e.g. W→T –19 vs. –5.

RAT versus mouse RBP

RAT versus bacterial lipocalin
Comparing two proteins with a PAM1 matrix gives completely different results than PAM250!

Consider two distantly related proteins. A PAM40 matrix is not forgiving of mismatches, and penalizes them severely. Using this matrix you can find no real match.

```
hsrbp, 136 CRLLNLGDTC
btlact,  3 CLLLALALTC
          * * * *
```

A PAM250 matrix is very tolerant of mismatches.

```
hsrbp,  86 --CADMVGTFTDTEDPAFKM
btlact, 80 GECQKKIIAETKIPAVFKI
          **    *  ** **
```

PAM matrices: Point-accepted mutations

PAM matrices are based on global alignments of closely related proteins.

The PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence.

Other PAM matrices are extrapolated from PAM1.

All the PAM data come from closely related proteins (>85% amino acid identity)
Two randomly diverging protein sequences change in a negatively exponential fashion

At PAM1, two proteins are 99% identical
At PAM10.7, there are 10 differences per 100 residues
At PAM80, there are 50 differences per 100 residues
At PAM250, there are 80 differences per 100 residues
PAM matrices reflect different degrees of divergence

- Two proteins with 50% identity may have 80 changes per 100 residues. (Why? Because any residue can be subject to back mutations.)

- Proteins with 20% to 25% identity are in the “twilight zone” and may be statistically significantly related.

- PAM or “accepted point mutation” refers to the “hits” or matches between two sequences (Dayhoff & Eck, 1968)
Ancestral sequence

ACCCCTAC

- A no change
- C single substitution
- C multiple substitutions
- C --> G coincidental substitutions
- T --> A parallel substitutions
- A --> C --> T convergent substitutions
- C back substitution

Sequence 1: ACCGATC
Sequence 2: AATAATC

Li (1997) p.70

Fig. 11.11
Page 374

Percent identity between two proteins:
What percent is significant?

100%
80%
65%
30%
23%
19%

Page 62
Outline: today’s topic

1. How to access the sequence and structure of a protein at NCBI and the Protein Data Bank (PDB)

2. Overview of databases of all proteins: NCBI and SwissProt

3. How to align the sequences of two proteins: Dayhoff’s evolutionary perspective

4. How to align the sequences of two proteins: pairwise alignment

General approach to pairwise alignment

• Choose two sequences
• Select an algorithm that generates a score
• Allow gaps (insertions, deletions)
• Score reflects degree of similarity
• Alignments can be global or local
• Estimate probability that the alignment occurred by chance
An alignment scoring system is required to evaluate how good an alignment is

- positive and negative values assigned
- gap creation and extension penalties
- positive score for identities
- some partial positive score for conservative substitutions
- global versus local alignment
- use of a substitution matrix

Calculation of an alignment score

\[ S = \sum (\text{identities, mismatches}) - \sum (\text{gap penalties}) \]

\[ \text{Score} = \text{Max}(S) \]

We will first consider the global alignment algorithm of Needleman and Wunsch (1970).

We will then explore the local alignment algorithm of Smith and Waterman (1981).

Finally, we will consider BLAST, a heuristic version of Smith-Waterman.

Global alignment with the algorithm of Needleman and Wunsch (1970)

• Two sequences can be compared in a matrix along x- and y-axes.

• If they are identical, a path along a diagonal can be drawn

• Find the optimal subpaths, and add them up to achieve the best score. This involves
  -- adding gaps when needed
  -- allowing for conservative substitutions
  -- choosing a scoring system (simple or complicated)

• N-W is guaranteed to find optimal alignment(s)
Three steps to global alignment with the Needleman-Wunsch algorithm

[1] set up a matrix
[2] score the matrix
[3] identify the optimal alignment(s)

Four possible outcomes in aligning two sequences

[1] identity (stay along a diagonal)
[2] mismatch (stay along a diagonal)
[3] gap in one sequence (move vertically!)
[4] gap in the other sequence (move horizontally!)
Start Needleman-Wunsch with an identity matrix
Start Needleman-Wunsch with an identity matrix

Fill in the matrix starting from the bottom right
Fig. 3.21
Page 65
Fig. 3.21
Page 65

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>N</th>
<th>J</th>
<th>R</th>
<th>Q</th>
<th>C</th>
<th>L</th>
<th>R</th>
<th>P</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Fig. 3.22
Page 66

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>N</th>
<th>J</th>
<th>R</th>
<th>Q</th>
<th>C</th>
<th>L</th>
<th>R</th>
<th>P</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Rule for assigning score in position $i, j$: 

$$s_{i,j} = \max \begin{cases} s_{i-1,j-1} + s(a_i, b_j) \\ s_{i-x,j} \text{ (i.e. add a gap of length } x) \\ s_{i,j-x} \text{ (i.e. add a gap of length } x) \end{cases}$$
After you’ve filled in the matrix, find the optimal path(s) by a “traceback” procedure.

Fig. 3.22
Page 66
Needleman-Wunsch: dynamic programming

N-W is guaranteed to find optimal alignments, although the algorithm does not search all possible alignments.

It is an example of a dynamic programming algorithm: an optimal path (alignment) is identified by incrementally extending optimal subpaths. Thus, a series of decisions is made at each step of the alignment to find the pair of residues with the best score.

Fig. 3.23
Page 68
hsrbp.pep x btlacto.pep

1 MKWWALLLLLARAWAAAEDCAUSSFRAKENFDKAAFSGTWYARMKAKDPEG 50
1 ...MKCLLLLALALTCQAPQLIUT...QTMKGLDIQKVQGTWYSLAMAR50 44
51 LFLQDHIVAEFSUDEGQMSATKGRUR.LLNNWD...UCADMUGTFTDE97
45 ISLLDAQSAFLRVEEELKPTPEGOELLQWENGECARQKIIAEKT093
98 DPAAFKMKYUGBASFLQGNNDDHIVDNYDTYAV.........QYSC136
45 ISLLDAQSAFLRVEEELKPTPEGOELLQWENGECARQKIIAEKT093
94 IPARFKIDALHENKUL...........QYSC136
137 ALLNLDGTCADSYSFVSSPDNLPLPERQKIVQRAQ.EECLLARQYAL105
136 QCLURTPEUDEALEYKDIAKLPMBHMALSFMNTOLLEEOCH170

?> bestfit

Bestfit makes an optimal alignment of the best segment of similarity
between two sequences. Optimal alignments are found by inserting gaps
to maximize the number of matches using the local homology algorithm of
Smith and Waterman.

BESTFIT of what sequence 1 ? hsrbp.pep

Begin (* 1 *)?
End (* 199 *)?

to what sequence 2 (* hsrbp.pep *) ? btlacto.pep

Begin (* 1 *)?
End (* 178 *)?

What is the gap creation penalty (* 8 *)?
What is the gap extension penalty (* 2 *)?
What should I call the paired output display file (* hsrbp.pair *)?

Aligning ..........-
Aligning ..........-

Gaps: 5
Quality: 59
Quality Ratio: 0.631
% Similarity: 39.138
Length: 105
Global alignment versus local alignment

Global alignment (Needleman-Wunsch) extends from one end of each sequence to the other.

Local alignment finds optimally matching regions within two sequences (“subsequences”).

Local alignment is almost always used for database searches such as BLAST. It is useful to find domains (or limited regions of homology) within sequences.

Smith and Waterman (1981) solved the problem of performing optimal local sequence alignment. Other methods (BLAST, FASTA) are faster but less thorough.
How the Smith-Waterman algorithm works

Set up a matrix between two proteins (size m+1, n+1)

No values in the scoring matrix can be negative! $S \geq 0$

The score in each cell is the maximum of four values:
1. $s(i-1, j-1) + \text{the new score at } [i,j]$ (a match or mismatch)
2. $s(i, j-1) - \text{gap penalty}$
3. $s(i-1, j) - \text{gap penalty}$
4. zero

Smith-Waterman local alignment algorithm

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>A</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>U</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>G</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7</td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.6</td>
<td>1.3</td>
<td>0.3</td>
<td>1.0</td>
<td>0.3</td>
<td>2.0</td>
<td>0.7</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>1.0</td>
<td>0.7</td>
<td>0.0</td>
<td>1.6</td>
<td>3.0</td>
<td>1.7</td>
<td>1.3</td>
<td>1.6</td>
<td>1.3</td>
<td>1.7</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.0</td>
<td>2.0</td>
<td>0.7</td>
<td>0.3</td>
<td>1.7</td>
<td>2.7</td>
<td>1.3</td>
<td>1.6</td>
<td>0.7</td>
<td>1.0</td>
<td>1.3</td>
<td>1.3</td>
</tr>
<tr>
<td>U</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7</td>
<td>1.7</td>
<td>0.3</td>
<td>1.3</td>
<td>2.3</td>
<td>1.0</td>
<td>0.7</td>
<td>1.7</td>
<td>2.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>U</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.3</td>
<td>1.3</td>
<td>1.0</td>
<td>2.3</td>
<td>2.3</td>
<td>2.0</td>
<td>0.7</td>
<td>1.7</td>
<td>2.7</td>
<td>1.7</td>
</tr>
<tr>
<td>G</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.3</td>
<td>0.0</td>
<td>1.0</td>
<td>1.6</td>
<td>2.0</td>
<td>3.3</td>
<td>2.0</td>
<td>1.7</td>
<td>1.3</td>
<td>2.3</td>
</tr>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>1.6</td>
<td>0.3</td>
<td>0.7</td>
<td>0.7</td>
<td>2.6</td>
<td>3.0</td>
<td>1.7</td>
<td>1.3</td>
<td>2.3</td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.7</td>
<td>1.0</td>
<td>2.0</td>
<td>0.7</td>
<td>1.7</td>
<td>1.3</td>
<td>2.7</td>
<td>1.3</td>
<td>1.0</td>
<td>2.0</td>
</tr>
<tr>
<td>G</td>
<td>0.0</td>
<td>0.0</td>
<td>1.7</td>
<td>0.0</td>
<td>0.3</td>
<td>0.7</td>
<td>1.7</td>
<td>0.3</td>
<td>2.7</td>
<td>1.7</td>
<td>2.7</td>
<td>2.3</td>
<td>1.0</td>
</tr>
<tr>
<td>G</td>
<td>0.0</td>
<td>0.0</td>
<td>1.7</td>
<td>0.7</td>
<td>0.3</td>
<td>0.3</td>
<td>1.3</td>
<td>1.3</td>
<td>2.3</td>
<td>1.3</td>
<td>2.3</td>
<td>2.0</td>
<td>2.0</td>
</tr>
</tbody>
</table>
Rapid, heuristic versions of Smith-Waterman: FASTA and BLAST

Smith-Waterman is very rigorous and it is guaranteed to find an optimal alignment.

But Smith-Waterman is slow. It requires computer space and time proportional to the product of the two sequences being aligned (or the product of a query against an entire database).

Gotoh (1982) and Myers and Miller (1988) improved the algorithms so both global and local alignment require less time and space.

FASTA and BLAST provide rapid alternatives to S-W.

• Go to http://www.ncbi.nlm.nih.gov/BLAST
• Choose BLAST 2 sequences
• In the program,
  [1] choose blastp or blastn
  [2] paste in your accession numbers
     (or use FASTA format)
  [3] select optional parameters
     --3 BLOSUM and 3 PAM matrices
     --gap creation and extension penalties
     --filtering
     --word size
  [4] click “align”
BLAST 2 SEQUENCES

This tool produces the alignment of two given sequences using BLAST engine for local alignment.

The stand-alone executable for blasting two sequences (Blastp) can be retrieved from NCBI in the
Reference: Tatiana A. Tatusova, Thomas L. Madden (1999), "Blast 2 sequences - a new tool for comparing protein
and nucleotide sequences", TIGR Microbial Lign. 15:247-250

Fig. 3.27

Page 73

BLAST 2 SEQUENCES RESULTS VERSION BLASTP 2.2.1 [Jul 12 2001]

<table>
<thead>
<tr>
<th>Matrix</th>
<th>PAM50</th>
<th>gap open</th>
<th>gap extension</th>
<th>x-dropoff</th>
<th>Freq</th>
<th>Expect</th>
<th>Wordsize</th>
<th>Filter</th>
<th>Alignt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence 1</td>
<td>580112</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length</td>
<td>199</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequence 2</td>
<td>129910</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length</td>
<td>178</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE: The statistics (bit score and expect value) are calculated based on the size of nr database.

Query: Lipocalin / Cynthiaus fatty-acid
retinal binding protein
26
26
------------------------------------------------------------------------------------------------------------------

Object: lipocalin / Cynthiaus fatty-acid
retinal binding protein
68
68

Fig. 3.28

Page 74
Fig. 3.29
Page 76
True positives | False positives
---|---
False negatives | True negatives

**Sensitivity:** ability to find true positives

**Specificity:** ability to minimize false positives