Protein sequence alignment and evolution

Tuesday, April 5, 2005

Protein Bioinformatics
260.841
Jonathan Pevsner
pevsner@jhmi.edu

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	Pevsner
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	To be announced	
Th May 12	Protein-protein docking	Gray
T May 17	To be announced	
Th May 19	Final exam	

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	Pevsner
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

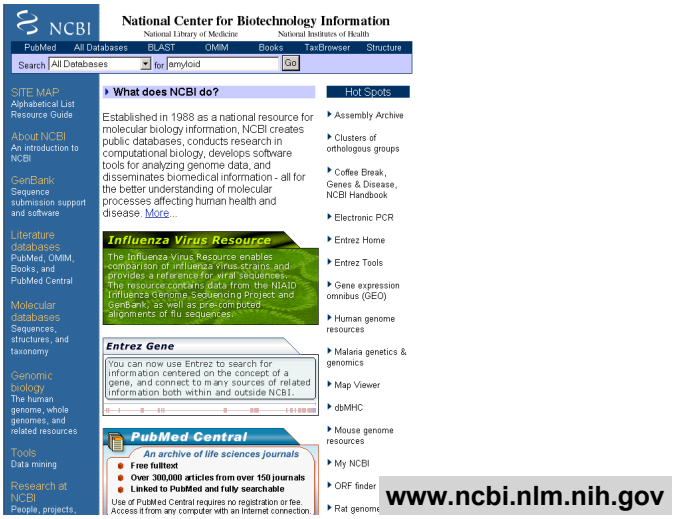
Outline: today's topic

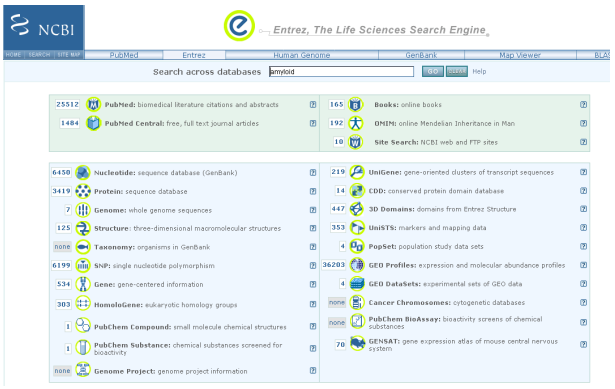
1. How to access the sequence and structure of a protein at NCBI and the Protein Data Bank (PDB)

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

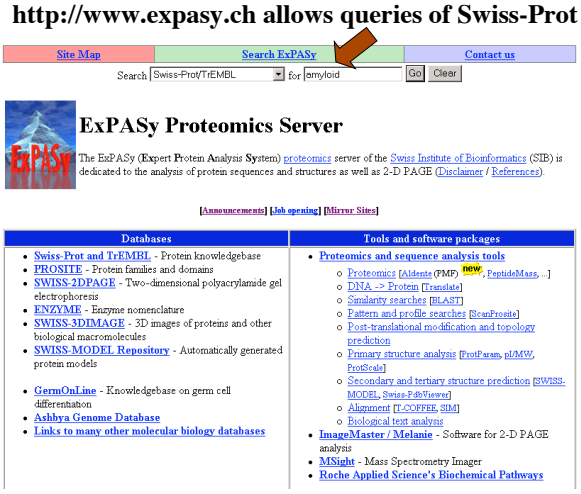
3. How to align the sequences of two proteins: Dayhoff's evolutionary perspective

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





http://www.expasy.ch allows queries of Swiss-Prot



Search in Swiss-Prot and TrEMBL for: amyloid

Swiss-Prot Release 46.4 of 29-Mar-2005
TrEMBL Release 29.4 of 29-Mar-2005

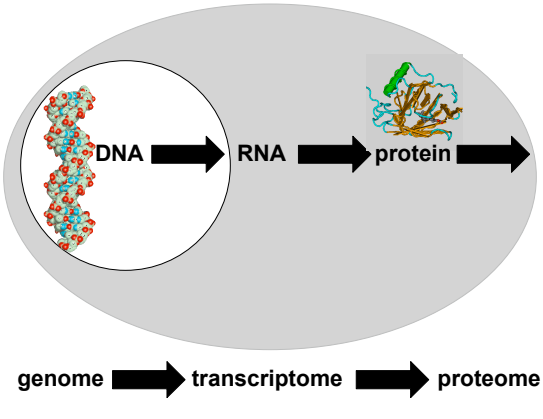
- Number of sequences found in Swiss-Prot(103) and TrEMBL(216): **319**
- Note that the selected sequences can be saved to a file to be later retrieved; to do so, go to the bottom of this page.
- For more directed searches, you can use the Sequence Retrieval System SRS.

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

A4_BOVIN (**Q28053**)
Alzheimer's disease amyloid A4 protein homolog [Contains: Beta-amyloid protein (Beta-APP) (A-beta)] (Fragment). (GENE Name=APP) - Bos taurus (Bovine)
A4_CAEEL (**Q10651**)
Beta-amyloid-like protein precursor. (GENE Name=apl-1; ORFNames=C42D8.8) - Caenorhabditis elegans
A4_CANFA (**Q28280**)
Alzheimer's disease amyloid A4 protein homolog [Contains: Beta-amyloid protein (Beta-APP) (A-beta)] (Fragment). (GENE Name=APP) - Canis familiaris (Dog)
A4_CAVPO (**Q60495**)
Amyloid beta A4 protein precursor (APP) (ABPP) (Alzheimer's disease amyloid protein homolog) [Contains: Soluble APP-alpha (S-APP-alpha); Soluble APP-beta (S-APP-beta); CTF-alpha; CTF-beta; Beta-amyloid protein 42 (Beta-APP42); Beta-amyloid protein 40 (Beta-APP40); P3(42); P3(40); Gamma-CTF(59) (Gamma-secretase C-terminal fragment 59); Gamma-CTF(57) (Gamma-secretase C-terminal fragment 57); C31]. (GENE Name=APP) - Cavia porcellus (Guinea pig)
A4_DROME (**P14599**)
Beta-amyloid-like protein precursor. (GENE Name=Appl; Synonyms=VND; ORFNames=CG7727) - Drosophila melanogaster (Fruit fly)
A4_FUGRU (**O93279**)
Alzheimer's disease amyloid A4 protein homolog precursor [Contains: Beta-amyloid protein (Beta-APP) (A-beta)]. (GENE Name=APP) - Fugu rubripes (Japanese pufferfish) (Takifugu rubripes)
A4_HUMAN (**P05067**)
Amyloid beta A4 protein precursor (APP) (ABPP) (Alzheimer's disease amyloid protein) (Cerebral vascular amyloid peptide) (CVAP) (Protease nexin-II) (PN-II) (APPI) (PreA4) [Contains: Soluble APP-alpha (S-APP-alpha); Soluble APP-beta (S-APP-beta); C99; Beta-amyloid protein 42 (Beta-APP42); Beta-amyloid protein 40 (Beta-APP40); C83; P3(42); P3(40); Gamma-CTF(59) (Gamma-secretase C-terminal fragment 59) (Amyloid intracellular domain 59) (AID(59)); Gamma-CTF(57) (Gamma-secretase C-terminal fragment 57) (Amyloid intracellular domain 57) (AID(57)); Gamma-CTF(50) (Gamma-secretase C-terminal fragment 50) (Amyloid intracellular domain 50) (AID(50)); C31]. (GENE Name=APP; Synonyms=A4, AD1) -

Protein Data Bank (PDB) (http://www.pdb.org)

Central dogma of molecular biology



Central dogma of bioinformatics and genomics

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):

X02775	GenBank genomic DNA sequence	**DNA**
NT_030059	Genomic contig	
Rs7079946	dbSNP (single nucleotide polymorphism)	
N91759.1	An expressed sequence tag (1 of 170)	**RNA**
NM_006744	RefSeq DNA sequence (from a transcript)	
NP_007635	RefSeq protein	**protein**
AAC02945	GenBank protein	
Q28369	SwissProt protein	
1KT7	Protein Data Bank structure record	

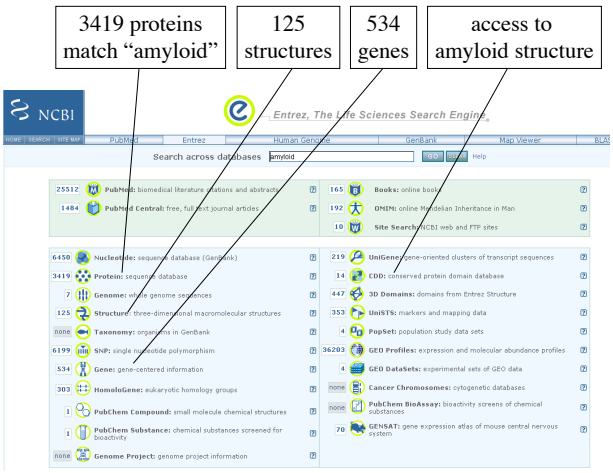
NCBI's important RefSeq project: best representative sequences

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.

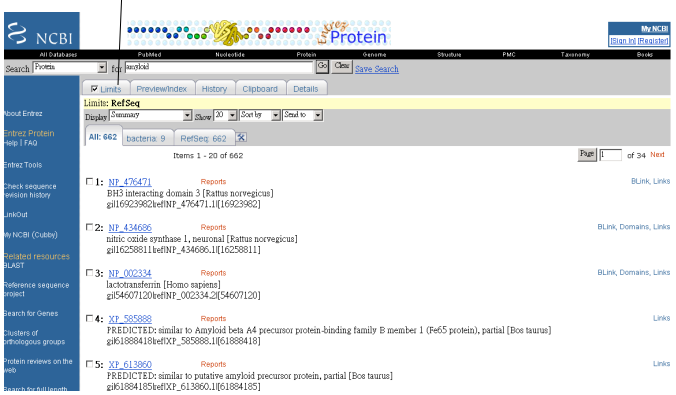
RefSeq identifiers include the following formats:

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.



PDB PROTEIN DATA BANK

Query Result Browser

Your query found 354 structures in the current PDB release and you have selected 0 structures so far. (There are currently 1 structures being processed. can select specific structures by clicking on the checkbox next to their id. If you do not select any structures, certain options will default to all structures. the Explore link!

Pull down to select option: New Search Go

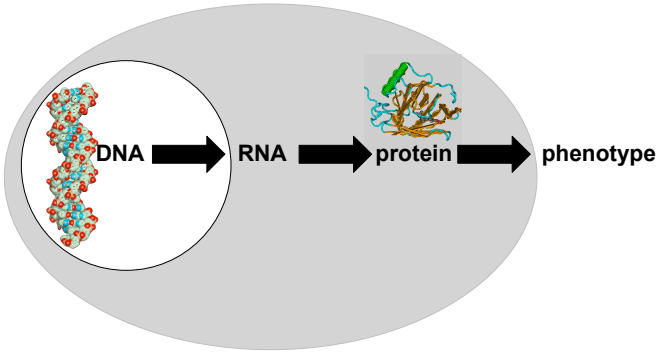
1-20

KEY: = Download compressed (GNU zipped) PDB file = View PDB file = Structure viewing options

133L	*Deposited:* **01-Jun-1993** *Exp. Method:* **X-ray Diffraction** *Resolution:* **1.77 Å**
Title	Role of Arg115 in the catalytic action of human lysozyme. X-ray structure of His115 and Glu115 mutants.
Classification	Hydrolase(O-Glycosyl)
Compound	Lysozyme (E.C. 3.2.1.17) Mutant With Arg 115 Replaced By His (R115H)
134L	*Deposited:* **01-Jun-1993** *Exp. Method:* **X-ray Diffraction** *Resolution:* **1.77 Å**
Title	Role of Arg115 in the catalytic action of human lysozyme. X-ray structure of His115 and Glu115 mutants.
Classification	Hydrolase(O-Glycosyl)
Compound	Lysozyme (E.C. 3.2.1.17) Mutant With Arg 115 Replaced By Glu (R115E)
1AAP	*Deposited:* **14-Sep-1990** *Exp. Method:* **X-ray Diffraction** *Resolution:* **1.50 Å**
Title	X-ray crystal structure of the protease inhibitor domain of Alzheimer's amyloid β-protein precursor.
Classification	Proteinase Inhibitor (Trypsin)
Compound	Protease Inhibitor Domain Of Alzheimer'S Amyloid β-Protein Precursor (APPI)
1AMB	*Deposited:* **21-Oct-1994** *Exp. Method:* **NMR**
Title	Solution structure of residues 1-28 of the amyloid β-peptide.
Classification	Proteinase Inhibitor(Trypsin)
Compound	Alzheimer'S Disease Amyloid β-Peptide (Residues 1 - 28) (E.C. Number Not Assigned) (NMR, Minimized Average Structure)
1AMC	*Deposited:* **14-Nov-1994** *Exp. Method:* **NMR**
Title	Solution structure of residues 1-28 of the amyloid β-peptide.
Classification	Proteinase Inhibitor(Trypsin)

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



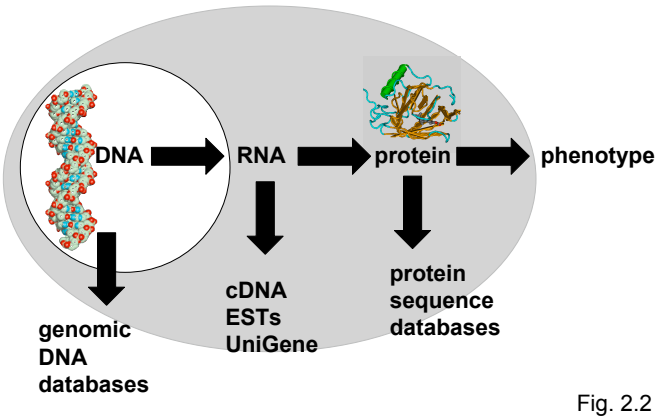
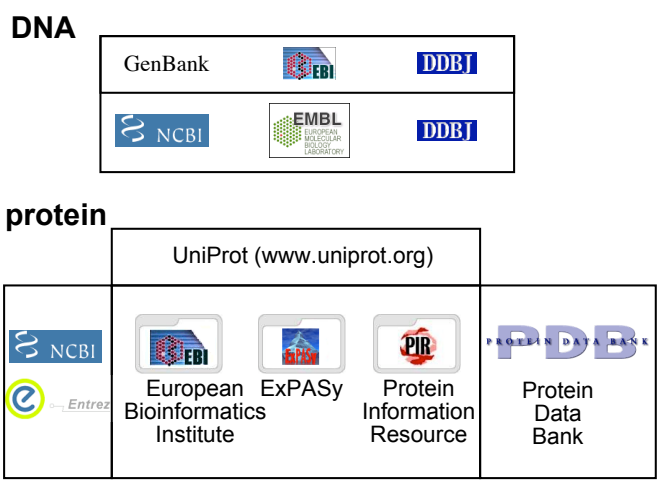


Fig. 2.2
Page 20



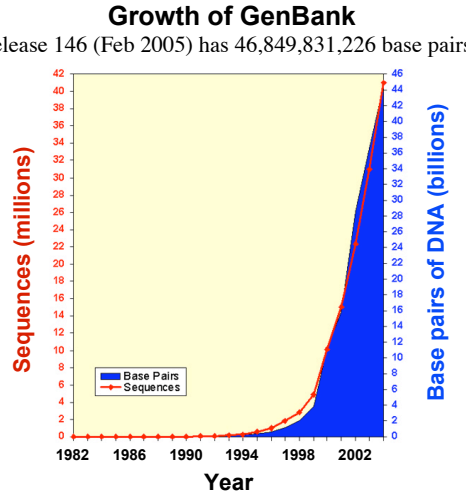
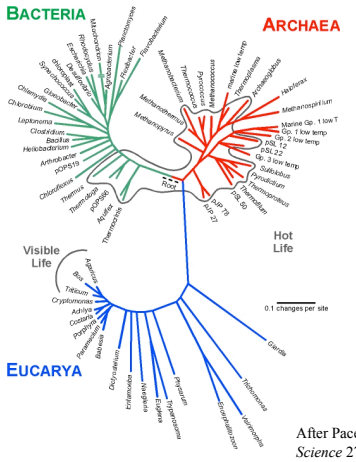


Fig. 2.1
Page 17



After Pace NR (1997)
Science 276:734

The most sequenced organisms in GenBank

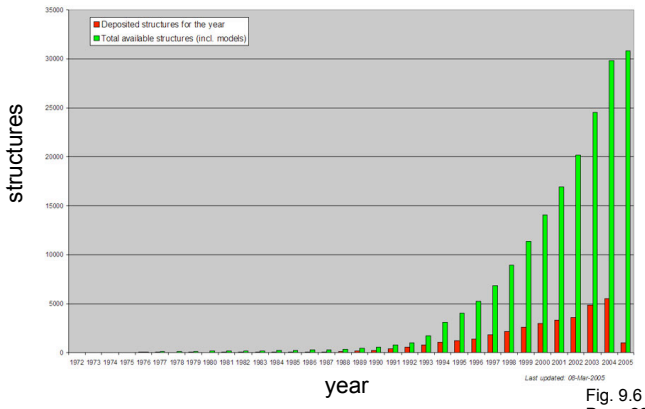
Homo sapiens	10.7 billion bases
Mus musculus	6.5b
Rattus norvegicus	5.6b
Danio rerio	1.7b
Zea mays	1.4b
Oryza sativa	0.8b
Drosophila melanogaster	0.7b
Gallus gallus	0.5b
Arabidopsis thaliana	0.5b

Updated 8-12-04
GenBank release 142.0

www.uniprot.org

SwissProt: 178,022 entries
TrEMBL: 1,647,645 entries
3-29-05 update

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

Page 225

15 most common domains (human)

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

Source: Integr8 program at www.ebi.ac.uk/proteome/

Table 8-3
Page 227

Pairwise alignments in the 1950s

β-corticotropin (sheep)	`ala gly glu asp asp glu`
Corticotropin A (pig)	`asp gly ala glu asp glu`
Oxytocin	`CYIQNCPLG`
Vasopressin	`CYFQNCPRG`

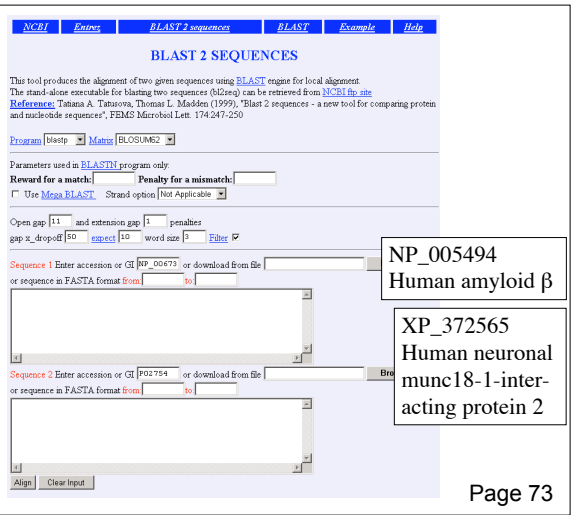
Early alignments revealed
--differences in amino acid sequences between species
--differences in amino acids responsible for distinct functions

Page 40

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

Page 41

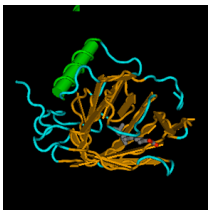


NP_005494
Human amyloid β

XP_372565
Human neuronal munc18-1-inter-acting protein 2

Page 73

RBP and β-lactoglobulin are homologous proteins that share related three-dimensional structures



retinol-binding protein
(NP_006735)

β-lactoglobulin
(P02754)

Figure 3.1
Page 42

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.

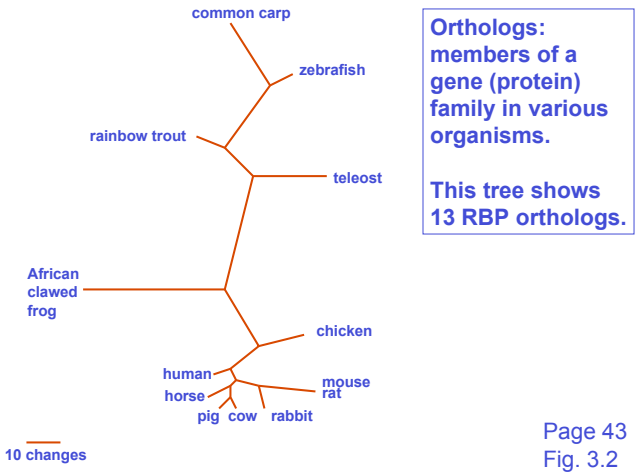
Identity
The extent to which two (nucleotide or amino acid)
sequences are invariant.

```
RBP          26  RVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWD- 84
                  +K++ +++   GTW++MA   +    L +   A   V  T  +         +L+ W+
glycodelin   23  QTKQDLELPKLAGTWHSMAMA-TNNISLMATLKAPLRVHITSLLPTPEDNLEIVLHRWEN 81
```


Definitions: two types of homology

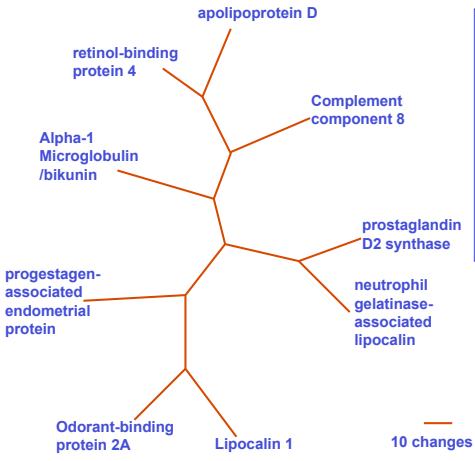
Orthologs
Homologous sequences in different species
that arose from a common ancestral gene
during speciation; may or may not be responsible
for a similar function.

Paralogs
Homologous sequences within a single species
that arose by gene duplication.



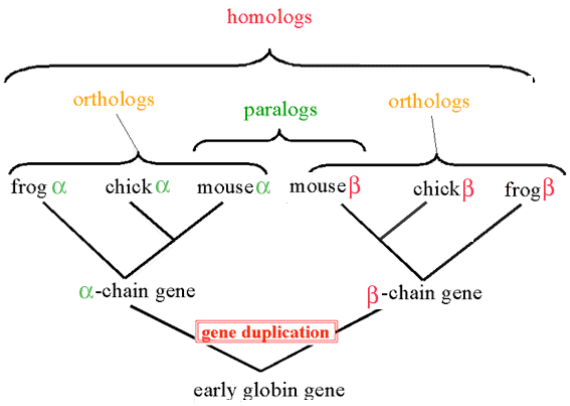
Orthologs: members of a gene (protein) family in various organisms.

This tree shows 13 RBP orthologs.



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

This tree shows 9 human lipocalins.



http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Orthology.html

Pairwise alignment of retinol-binding protein and β-lactoglobulin

```
  1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP
      .  |||  |     .   |.  .  .  |  : .||||.:|    :
  1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

 51 LFLQDNIVAEFSVDETGQMSATAKGRVR.LLNNWD..VCADMVGTFTDTE 97 RBP
    : | |   |   |    ::  | .| .  ||  |:   ||        |.
 45 ISLLDAQSAPLRV.YVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTK 93 lactoglobulin

 98 DPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAV...........QYSC 136 RBP
    || ||.        |         :.||||  | .           .|
 94 IPAVFKIDALNENKVL.........VLDTDYKKYLLFCMENSAEPEQSLAC 135 lactoglobulin

137 RLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQ.EELCLARQYRLIV 185 RBP
    . |       |     | :    ||   .      | || |
136 QCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI....... 178 lactoglobulin
```

Page 46
Fig. 3.5

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.

Page 47

Pairwise alignment of retinol-binding protein and β-lactoglobulin

```
  1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP
      .  |||  |     .   |.  .  .  |  : .||||.:|    :
  1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

 51 LFLQDNIVAEFSVDETGQMSATAKGRVR.LLNNWD..VCADMVGTFTDTE 97 RBP
    : | |   |   |    ::  | .| .  ||  |:   ||        |.
 45 ISLLDAQSAPLRV.YVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTK 93 lactoglobulin

 98 DPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAV...........QYSC 136 RBP
    || ||.        |         :.||||  | .           .|
 94 IPAVFKIDALNENKVL.........VLDTDYKKYLLFCMENSAEPEQSLAC 135 lactoglobulin

137 RLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQ.EELCLARQYRLIV 185 RBP
    . |       |     | :    ||   .      | || |
136 QCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI....... 178 lactoglobulin
```

Identity (bar)

Page 46
Fig. 3.5

Pairwise alignment of retinol-binding protein and β-lactoglobulin

```
  1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP
      .  |||  |     .   |.  .  .  |  : .||||.:|    :
  1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

 51 LFLQDNIVAEFSVDETGQMSATAKGRVR.LLNNWD..VCADMVGTFTDTE 97 RBP
    : | |   |   |    ::  | .| .  ||  |:   ||        |.
 45 ISLLDAQSAPLRV.YVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTK 93 lactoglobulin

 98 DPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAV...........QYSC 136 RBP
    || ||.        |         :.||||  | .           .|
 94 IPAVFKIDALNENKVL.........VLDTDYKKYLLFCMENSAEPEQSLAC 135 lactoglobulin

137 RLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQ.EELCLARQYRLIV 185 RBP
    . |       |     | :    ||   .      | || |
136 QCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI....... 178 lactoglobulin
```

Somewhat similar (one dot)

Very similar (two dots)

Page 46
Fig. 3.5

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.

Page 47

Pairwise alignment of retinol-binding protein and β-lactoglobulin

```
  1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP
      .  |||  |     .   |.  .  .  |  : .||||.:|    :
  1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

 51 LFLQDNIVAEFSVDETGQMSATAKGRVR.LLNNWD..VCADMVGTFTDTE 97 RBP
    : | |   |   |    ::  | .| .  ||  |:   ||        |.
 45 ISLLDAQSAPLRV.YVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTK 93 lactoglobulin

 98 DPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAV...........QYSC 136 RBP
    || ||.        |        :.||||  | .           .|
 94 IPAVFKIDALNENKVL........VLDTDYKKYLLFCMENSAEPEQSLAC 135 lactoglobulin

137 RLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQ.EELCLARQYRLIV 185 RBP
    . |       |     | :    ||   .      | || |
136 QCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI..... 178 lactoglobulin
```

Internal gap

Terminal gap

Page 46
Fig. 3.5

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.

Page 47

Pairwise alignment of retinol-binding protein and β-lactoglobulin

```
  1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP
      .  |||  |     .   |.  .  .  |  : .||||.:|    :
  1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

 51 LFLQDNIVAEFSVDETGQMSATAKGRVR.LLNNWD..VCADMVGTFTDTE 97 RBP
    : | |   |   |    ::  | .| .  ||  |:   ||        |.
 45 ISLLDAQSAPLRV.YVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTK 93 lactoglobulin

 98 DPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAV...........QYSC 136 RBP
    || ||.        |        :.||||  | .           .|
 94 IPAVFKIDALNENKVL........VLDTDYKKYLLFCMENSAEPEQSLAC 135 lactoglobulin

137 RLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQ.EELCLARQYRLIV 185 RBP
    . |       |     | :    ||   .      | || |
136 QCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI..... 178 lactoglobulin
```

Page 46
Fig. 3.5

Pairwise alignment of retinol-binding protein from human (top) and rainbow trout (*O. mykiss*)

```
  1 .MKWVWALLLLA.AWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDP 48
     ::   ||  ||  ||   .||.||. .| :|||:.|:.| |||.|||||
  1 MLRICVALCALATCWA...QDCQVSNIQVMQNFDRSRYTGRWYAVAKKDP 47
          .        .        .        .        .
 49 EGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTED 98
     |||| ||:||:|||||.|.|.|||| ||| :|||||:.||.| ||| || |
 48 VGLFLLDNVVAQFSVDESGKMTATAHGRVIILNNWEMCANMFGTFEDTPD 97
          .        .        .        .        .
 99 PAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRLLNLDGTCADS 148
    ||||||:||| ||:|| |||||||::||||| ||: |||| ..||||| |
 98 PAKFKMRYWGAASYLQTGNDDHWVIDTDYDNYAIHYSCREVDLDGTCLDG 147
          .        .        .        .        .
149 YSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL 199
    |||:||| | || || ||||  :..|:|   .|| : | |:|:
148 YSFIFSRHPTGLRPEDQKIVTDKKKEICFLGKYRRVGHTGFCESS..... 192
```

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 VVGVNEHTYQ PNMDIVSNAS CTTNCLAPLA
bacterium GAKKVVMTGP SKDNTPM..F VKGANFDKY. AGQDIVSNAS CTTNCLAPLA
yeast     GAKKVVITAP SS.TAPM..F VMGVNEEKYT SDLKIVSNAS CTTNCLAPLA
archaeon  GADKVLISAP PKGDEPVKQL VYGVNHDEYD GE.DVVSNAS CTTNSITPVA

fly       KVINDNFEIV EGLMTTVHAT TATQKTVDGP SGKLWRDGRG AAQNIIPAST
human     KVIHDNFGIV EGLMTTVHAI TATQKTVDGP SGKLWRDGRG ALQNIIPAST
plant     KVVHEEFGIL EGLMTTVHAT TATQKTVDGP SMKDWRGGRG ASQNIIPSST
bacterium KVINDNFGII EGLMTTVHAT TATQKTVDGP SHKDWRGGRG ASQNIIPSST
yeast     KVINDAFGIE EGLMTTVHSL TATQKTVDGP SHKDWRGGRT ASGNIIPSST
archaeon  KVLDEEFGIN AGQLTTVHAY TGSQNLMDGP NGKP.RRRRA AAENIIPTST

fly       GAAKAVGKVI PALNGKLTGM AFRVPTPNVS VVDLTVRLGK GASYDEIKAK
human     GAAKAVGKVI PELNGKLTGM AFRVPTANVS VVDLTCRLEK PAKYDDIKKV
plant     GAAKAVGKVL PELNGKLTGM AFRVPTSNVS VVDLTCRLEK GASYEDVKAA
bacterium GAAKAVGKVL PELNGKLTGM AFRVPTPNVS VVDLTVRLEK AATYEQIKAA
yeast     GAAKAVGKVL PELQGKLTGM AFRVPTVDVS VVDLTVKLNK ETTYDEIKKV
archaeon  GAAQAATEVL PELEGKLDGM AIRVPVPNGS ITEFVVDLDD DVTESDVNAA
```

Page 48
Fig. 3.7

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.

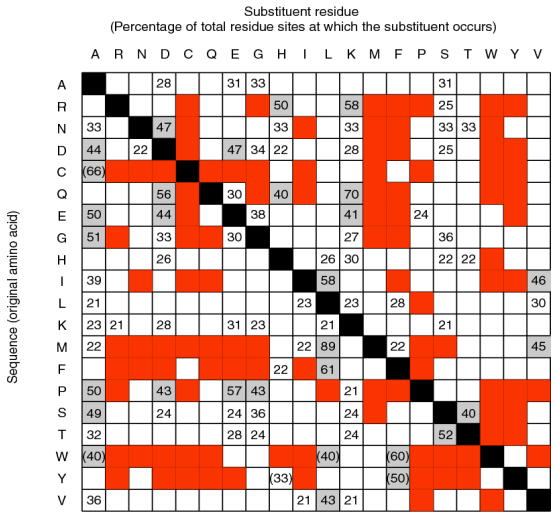


Fig. 3.31
Page 80

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.

Dayhoff's 34 protein superfamilies

Protein	PAMs per 100 million years
Ig kappa chain	37
Kappa casein	33
Lactalbumin	27
Hemoglobin α	12
Myoglobin	8.9
Insulin	4.4
Histone H4	0.10
Ubiquitin	0.00

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

AAlaRArgNAsnDAspCCysQGlnEGluGGlyAR^{30}N^{10917}D^{1}

Fig. 3.10
Page 52

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

```
fly       GAKKVIISAP SAD.APM..F VCGVNLDAYK PDMKVVSNAS CTTNCLAPLA
human     GAKRVIISAP SAD.APM..F VMGVNHEKYD NSLKIISNAS CTTNCLAPLA
plant     GAKKVIISAP SAD.APM..F VVGVNEHTYQ PNMDIVSNAS CTTNCLAPLA
bacterium GAKKVVMTGP SKDNTPM..F VKGANFDKY. AGQDIVSNAS CTTNCLAPLA
yeast     GAKKVVITAP SS.TAPM..F VMGVNEEKYT SDLKIVSNAS CTTNCLAPLA
archaeon  GADKVLISAP PKGDEPVKQL VYGVNHDEYD GE.DVVSNAS CTTNSITPVA

fly       KVINDNFEIV EGLMTTVHAT TATQKTVDGP SGKLWRDGRG AAQNIIPAST
human     KVIHDNFGIV EGLMTTVHAI TATQKTVDGP SGKLWRDGRG ALQNIIPAST
plant     KVVHEEFGIL EGLMTTVHAT TATQKTVDGP SMKDWRGGRG ASQNIIPSST
bacterium KVINDNFGII EGLMTTVHAT TATQKTVDGP SHKDWRGGRG ASQNIIPSST
yeast     KVINDAFGIE EGLMTTVHSL TATQKTVDGP SHKDWRGGRT ASGNIIPSST
archaeon  KVLDEEFGIN AGQLTTVHAY TGSQNLMDGP NGKP.RRRRA AAENIIPTST

fly       GAAKAVGKVI PALNGKLTGM AFRVPTPNVS VVDLTVRLGK GASYDEIKAK
human     GAAKAVGKVI PELNGKLTGM AFRVPTANVS VVDLTCRLEK PAKYDDIKKV
plant     GAAKAVGKVL PELNGKLTGM AFRVPTSNVS VVDLTCRLEK GASYEDVKAA
bacterium GAAKAVGKVL PELNGKLTGM AFRVPTPNVS VVDLTVRLEK AATYEQIKAA
yeast     GAAKAVGKVL PELQGKLTGM AFRVPTVDVS VVDLTVKLNK ETTYDEIKKV
archaeon  GAAQAATEVL PELEGKLDGM AIRVPVPNGS ITEFVVDLDD DVTESDVNAA
```

Dayhoff et al. estimated the relative mutability of amino acids

Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18

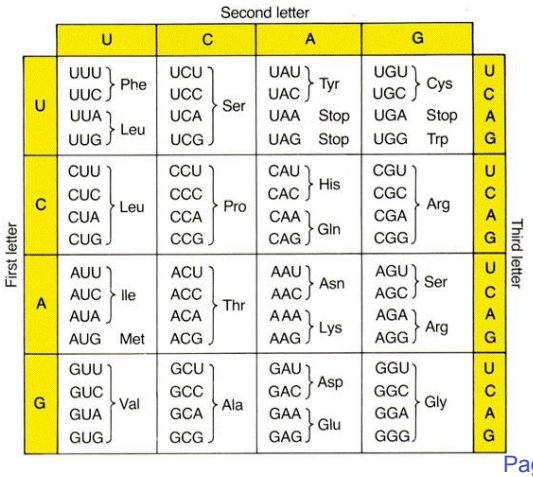
Table 3.1
Page 53

Normalized frequencies of amino acids: variations in frequency of occurrence

Gly	8.9%	Arg	4.1%
Ala	8.7%	Asn	4.0%
Leu	8.5%	Phe	4.0%
Lys	8.1%	Gln	3.8%
Ser	7.0%	Ile	3.7%
Val	6.5%	His	3.4%
Thr	5.8%	Cys	3.3%
Pro	5.1%	Tyr	3.0%
Glu	5.0%	Met	1.5%
Asp	4.7%	Trp	1.0%

blue=6 codons; red=1 codon

Page 53



Page 54

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

A	A	l	a	R	a	r	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	A	R	30	N	10	9	17	L

Page 52

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

Dayhoff's PAM1 mutation probability matrix

Original amino acid

A	A	l	a	R	a	r	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	l	e	A	98	67	29	10	3

Each element of the matrix shows the probability that an amino acid (top) will be replaced by another residue (side)

Fig. 3.11
Page 55

Substitution Matrix

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.

Page 53

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).

Page 55-56

Dayhoff's PAM1 mutation probability matrix

AAlaR	ArgN	AsnD	AspC	CysQ	GlnE	GluG	GlyH	HisI	IleA	98 6 5 7 2 9 10 3

Page 55

Dayhoff's PAM0 mutation probability matrix: the rules for extremely slowly evolving proteins

PAM0	AAlaR	ArgN	AsnD	AspC	CysQ	GlnE	GluG	GlyA	100%	0%	0%	0%	0%	0%	R

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

PAM∞	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly
A	8.7%	8.7%	8.7%	8.7%	8.7%	8.7%	8.7%	8.7%
R	4.1%	4.1%	4.1%	4.1%	4.1%	4.1%	4.1%	4.1%
N	4.0%	4.0%	4.0%	4.0%	4.0%	4.0%	4.0%	4.0%
D	4.7%	4.7%	4.7%	4.7%	4.7%	4.7%	4.7%	4.7%
C	3.3%	3.3%	3.3%	3.3%	3.3%	3.3%	3.3%	3.3%
Q	3.8%	3.8%	3.8%	3.8%	3.8%	3.8%	3.8%	3.8%
E	5.0%	5.0%	5.0%	5.0%	5.0%	5.0%	5.0%	5.0%
G	8.9%	8.9%	8.9%	8.9%	8.9%	8.9%	8.9%	8.9%

Top: original amino acid
Side: replacement amino acid

Fig. 3.12
Page 56

The PAM250 mutation probability matrix

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.

Page 56-57

PAM250 mutation probability matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

Top: original amino acid
Side: replacement amino acid

Fig. 3.13
Page 57

PAM250 log odds scoring matrix

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

Fig. 3.14
Page 58

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).

Page 57

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":

$$\frac{\text{the probability that an alignment is authentic}}{\text{the probability that the alignment was random}}$$

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

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

As an example, for tryptophan,

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

Page 57

What do the numbers mean in a log odds matrix?

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

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
$17 = 10 \log_{10} x$
$1.7 = \log_{10} x$
$10^{1.7} = x = 50$

Page 58

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.

Page 58

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

PAM250 log odds scoring matrix

Fig. 3.14
Page 58

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	7																			
R	-10	9																		
N	-7	-9	9																	
D	-6	-17	-1	8																
C	-10	-11	-17	-21	10															
Q	-7	-4	-7	-6	-20	9														
E	-5	-15	-5	0	-20	-1	8													
G	-4	-13	-6	-6	-13	-10	-7	7												
H	-11	-4	-2	-7	-10	-2	-9	-13	10											
I	-8	-8	-8	-11	-9	-11	-8	-17	-13	9										
L	-9	-12	-10	-19	-21	-8	-13	-14	-9	-4	7									
K	-10	-2	-4	-8	-20	-6	-7	-10	-10	-9	-11	7								
M	-8	-7	-15	-17	-20	-7	-10	-12	-17	-3	-2	-4	12							
F	-12	-12	-12	-21	-19	-19	-20	-12	-9	-5	-5	-20	-7	9						
P	-4	-7	-9	-12	-11	-6	-9	-10	-7	-12	-10	-10	-11	-13	8					
S	-3	-6	-2	-7	-6	-8	-7	-4	-9	-10	-12	-7	-8	-9	-4	7				
T	-3	-10	-5	-8	-11	-9	-9	-10	-11	-5	-10	-6	-7	-12	-7	-2	8			
W	-20	-5	-11	-21	-22	-19	-23	-21	-10	-20	-9	-18	-19	-7	-20	-8	-19	13		
Y	-11	-14	-7	-17	-7	-18	-11	-20	-6	-9	-10	-12	-17	-1	-20	-10	-9	-8	10	
V	-5	-11	-12	-11	-9	-10	-10	-9	-9	-1	-5	-13	-4	-12	-9	-10	-6	-22	-10	8

PAM10 log odds scoring matrix

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

Fig. 3.15
Page 59

BLOSUM90 PAM30 | BLOSUM80 PAM120 | BLOSUM62 PAM180 | BLOSUM45 PAM240

BLOSUM 80 PAM 1 | BLOSUM 62 PAM 120 | BLOSUM 45 PAM 250

Less divergent ←→ *More divergent*

Rat versus mouse RBP

Rat versus bacterial lipocalin

Fig. 3.18
Page 61

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 CRLLNLDGTC
btlact,   3 CLLLALALTC
              * ** *  **
```

A PAM250 matrix is very tolerant of mismatches.

```
24.7% identity in 81 residues overlap; Score: 77.0; Gap frequency: 3.7%
 hsrbp, 26 RVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDV
btlact, 21 QTMKGLDIQKVAGTWYSLAMAASD-ISLLDAQSAPLRVYVEELKPTPEGDLEILLQKWEN
                *     ****  *        * *       *               **  *

hsrbp,  86 --CADMVGTFTDTEDPAKFKM
btlact, 80 GECAQKKIIAEKTKIPAVFKI
             **        *  ** **
```

Page 60

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

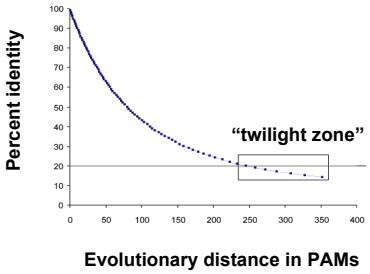


Fig. 3.19
Page 62

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

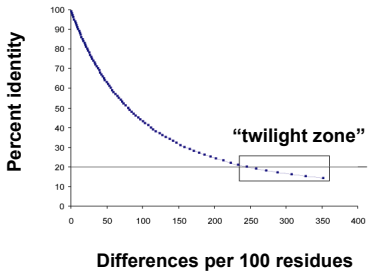
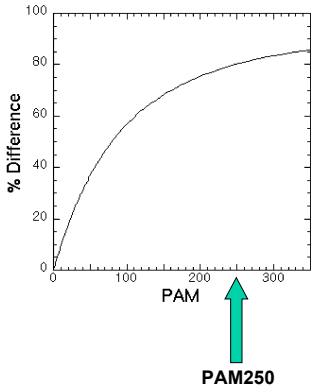


Fig. 3.19
Page 62

PAM matrices reflect different degrees of divergence



PAM: "Accepted point mutation"

- 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)

Page 62

Ancestral sequence

ACCCTAC

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

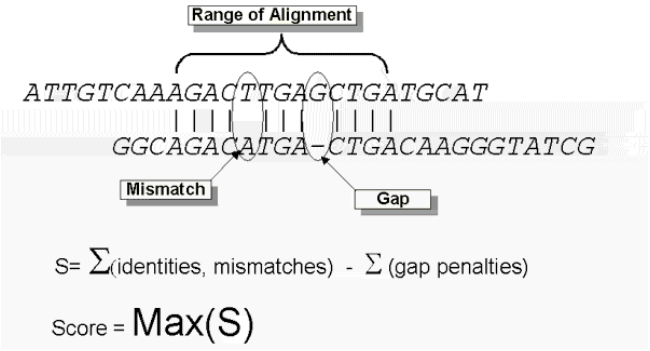
General approach to pairwise alignment

- 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



http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Alignment_Scores2.html

Two kinds of sequence alignment: global and local

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)

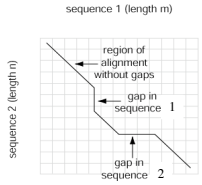
Three steps to global alignment with the Needleman-Wunsch algorithm

[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!)

Fig. 3.20
Page 64

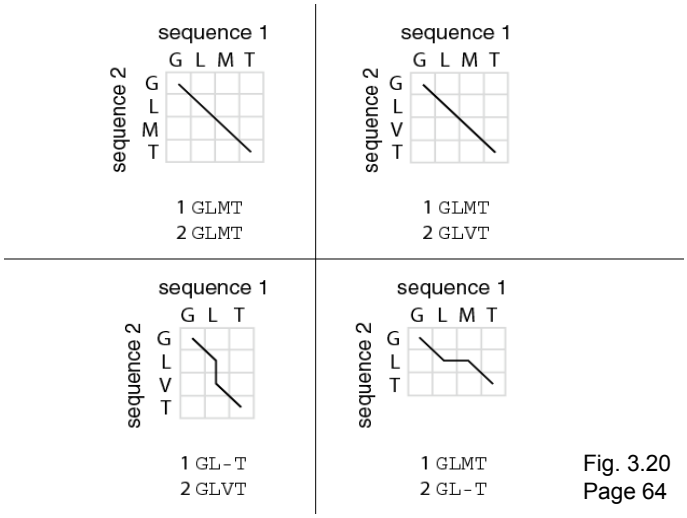


Fig. 3.20
Page 64

Start Needleman-Wunsch with an identity matrix

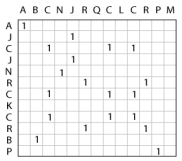
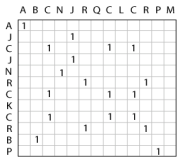


Fig. 3.21
Page 65

Start Needleman-Wunsch with an identity matrix

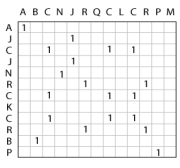


sequence 1 **ABCNJ-RQCLCR-PM**
sequence 2 **AJC-JNR-CKCRBP-**

sequence 1 **ABC-NJRQCLCR-PM**
sequence 2 **AJCJN-R-CKCRBP-**

Fig. 3.21
Page 65

Fill in the matrix starting from the bottom right



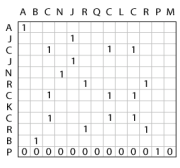
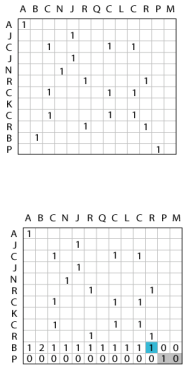


Fig. 3.21
Page 65



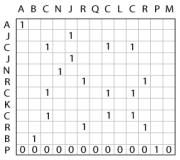
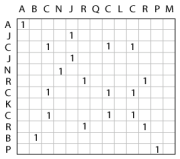
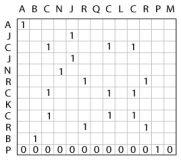
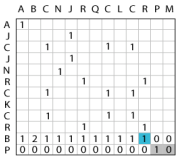


Fig. 3.21
Page 65







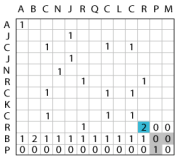
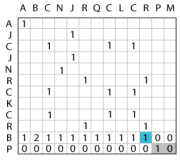


Fig. 3.21
Page 65



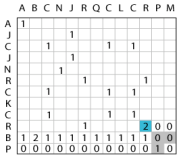
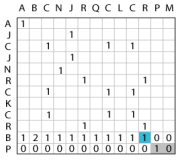
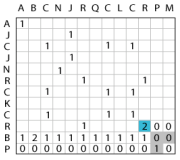


Fig. 3.21
Page 65





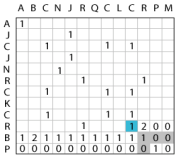
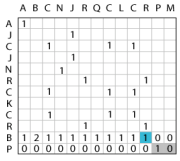
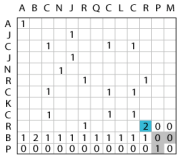
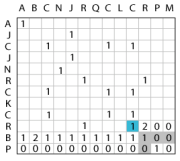


Fig. 3.22
Page 66







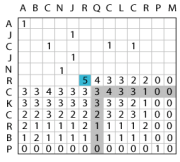
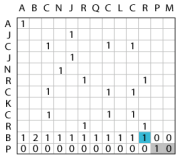
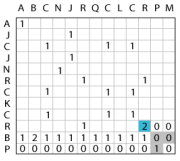
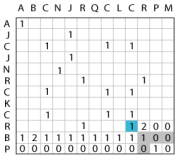


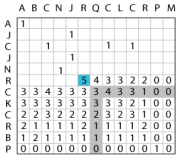
Fig. 3.22
Page 66







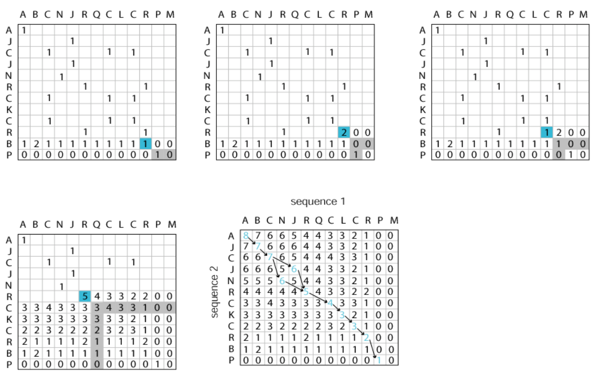
Rule for assigning score in position i, j:



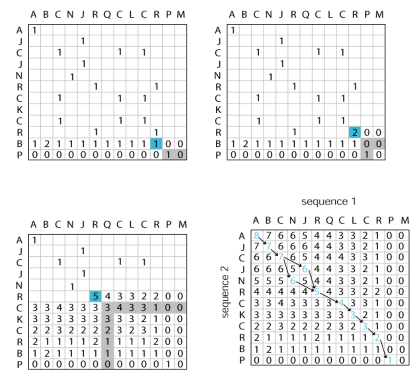
$s_{i,j} = \max$

- $s_{i-1,j-1} + s(a_i b_j)$
- $s_{i-x,j}$ (i.e. add a gap of length x)
- $s_{i,j-x}$ (i.e. add a gap of length x)

Fig. 3.22
Page 66



After you've filled in the matrix, find the optimal path(s) by a "traceback" procedure



sequence 1 **ABCNJ-RQCLCR-PM**
sequence 2 **AJC-JNR-CKCRBP-**

sequence 1 **ABC-NJRQCLCR-PM**
sequence 2 **AJCJN-R-CKCRBP-**

Fig. 3.22

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.


```
> gap

Gap uses the algorithm of Needleman and Wunsch to find the alignment of
two complete sequences that maximizes the number of matches and minimizes
the number of gaps.

 GAP 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:     8
       Quality:    37
 Quality Ratio: 0.208
  % Similarity: 31.902
        Length:   214
```

Fig. 3.23


```
         Gap Weight:      8      Average Match:  2.912
      Length Weight:      2   Average Mismatch: -2.003

            Quality:     37             Length:    214
              Ratio:  0.208               Gaps:      8
Percent Similarity: 31.902   Percent Identity: 26.380

       Match display thresholds for the alignment(s):
                   ¦ = IDENTITY
                   : =   2
                   . =   1

hsrbp.pep x btlacto.pep   July 16, 2001 14:45  ..


                  .         .         .         .         .
      1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50
           .  ¦¦¦  ¦      .    ¦.   .   .   ¦  :  .¦¦¦¦.:¦    :
      1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44

                  .         .         .         .         .
     51 LFLQDNIVAEFSVDETGQMSATAKGRVR.LLNNWD..VCADMVGTFTDTE 97
        : ¦ ¦    ¦    ¦     ::  ¦ .¦ .  ¦¦  ¦:   ¦¦         ¦.
     45 ISLLDAQSAPLRV.YVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTK 93

                  .         .         .         .         .
     98 DPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAV..........QYSC 136
         ¦¦ ¦¦.         ¦         :.¦¦¦¦  ¦ .           .¦
     94 IPAVFKIDALNENKVL........VLDTDYKKYLLFCMENSAEPEQSLAC 135

                  .         .         .         .         .
    137 RLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQ.EELCLARQYRLIV 185
         . ¦       ¦     ¦ :    ¦¦    .       ¦ ¦¦ ¦
    136 QCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI....... 178
                                    .
```

Fig. 3.24


```
> 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: 0.621
  % Similarity: 39.130
        Length:   105
```

Fig. 3.26

```
       Gap Weight:      8       Average Match:  2.912
    Length Weight:      2   Average Mismatch: -2.003

          Quality:     59              Length:    105
            Ratio:  0.621                Gaps:      5
Percent Similarity: 39.130   Percent Identity: 30.435

      Match display thresholds for the alignment(s):
                | = IDENTITY
                : =   2
                . =   1

hsrbp.pep x btlacto.pep    July 16, 2001 14:41   ..

    29 ENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVR 78
    24 KGLDIQKVAGTWYSLAMAASD.ISLLDAQSAPLRV.YVEELKPTPEGDLE 71

    79 .LLNNWD..VCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDT 125
    72 ILLQKWENGECAQKKIIAEKTKIPAVFKIDALNENKVL........VLDT 113

   126 DYDTY 130
   114 DYKKY 118
```

Fig. 3.26
Page 71

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.

Page 69

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) + the new score at [i,j] (a match or mismatch)
[2] s(i,j-1) – gap penalty
[3] s(i-1,j) – gap penalty
[4] zero

Page 69

Smith-Waterman local alignment algorithm

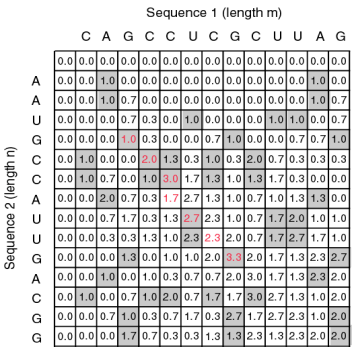


Fig. 3.25
Page 70

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

Page 71

Pairwise alignment: BLAST 2 sequences

- 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”

Page 72

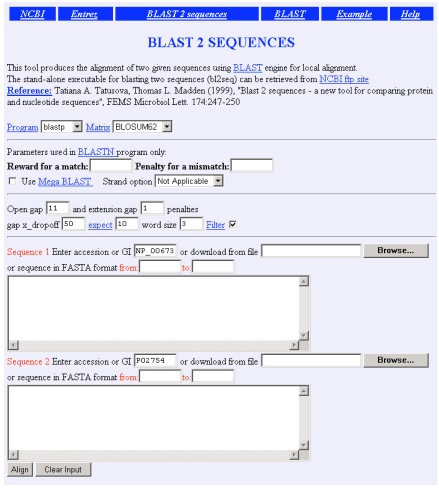


Fig. 3.27
Page 73

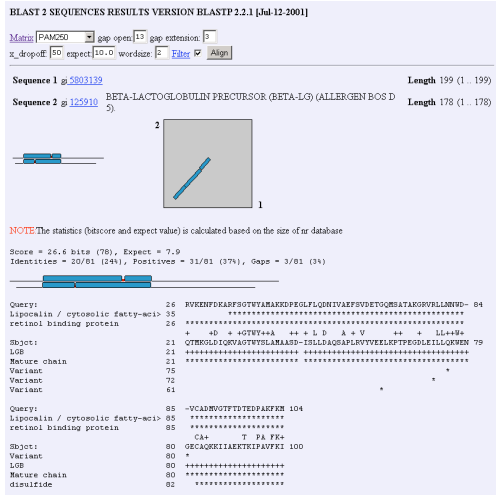


Fig. 3.28
Page 74

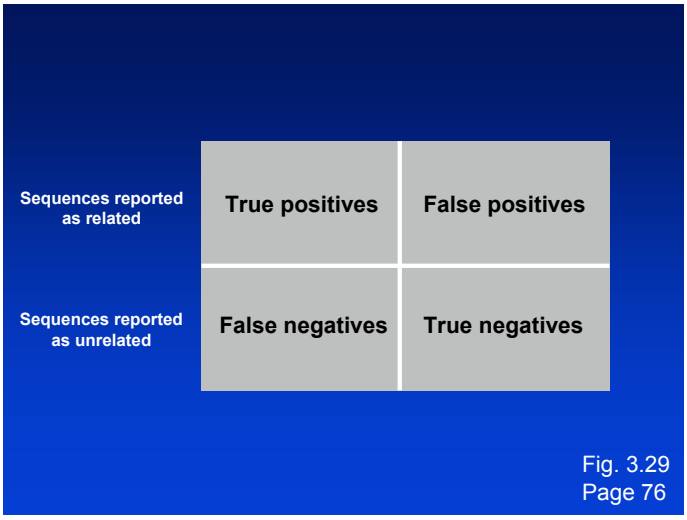


Fig. 3.29
Page 76

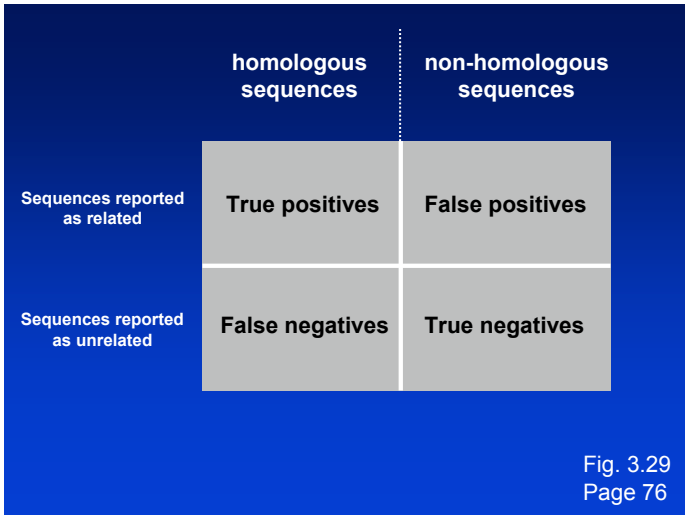


Fig. 3.29
Page 76