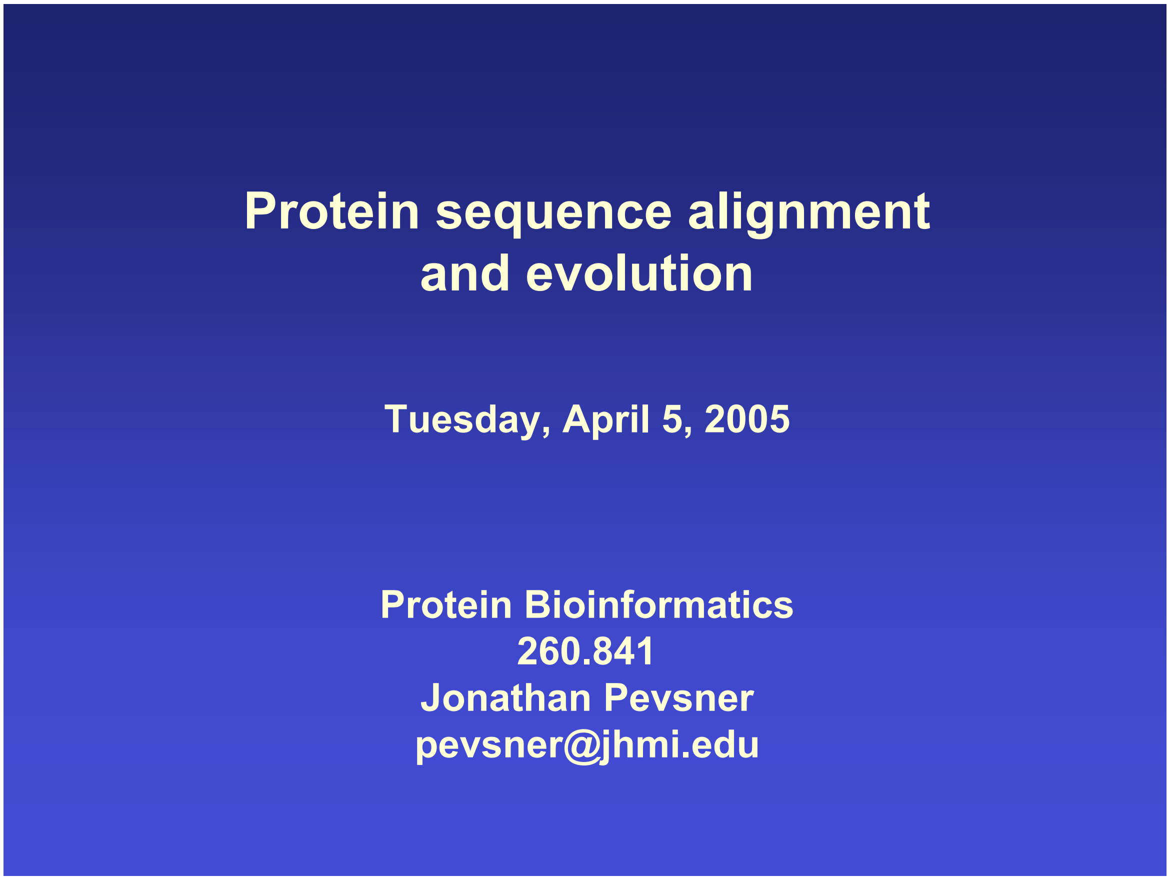
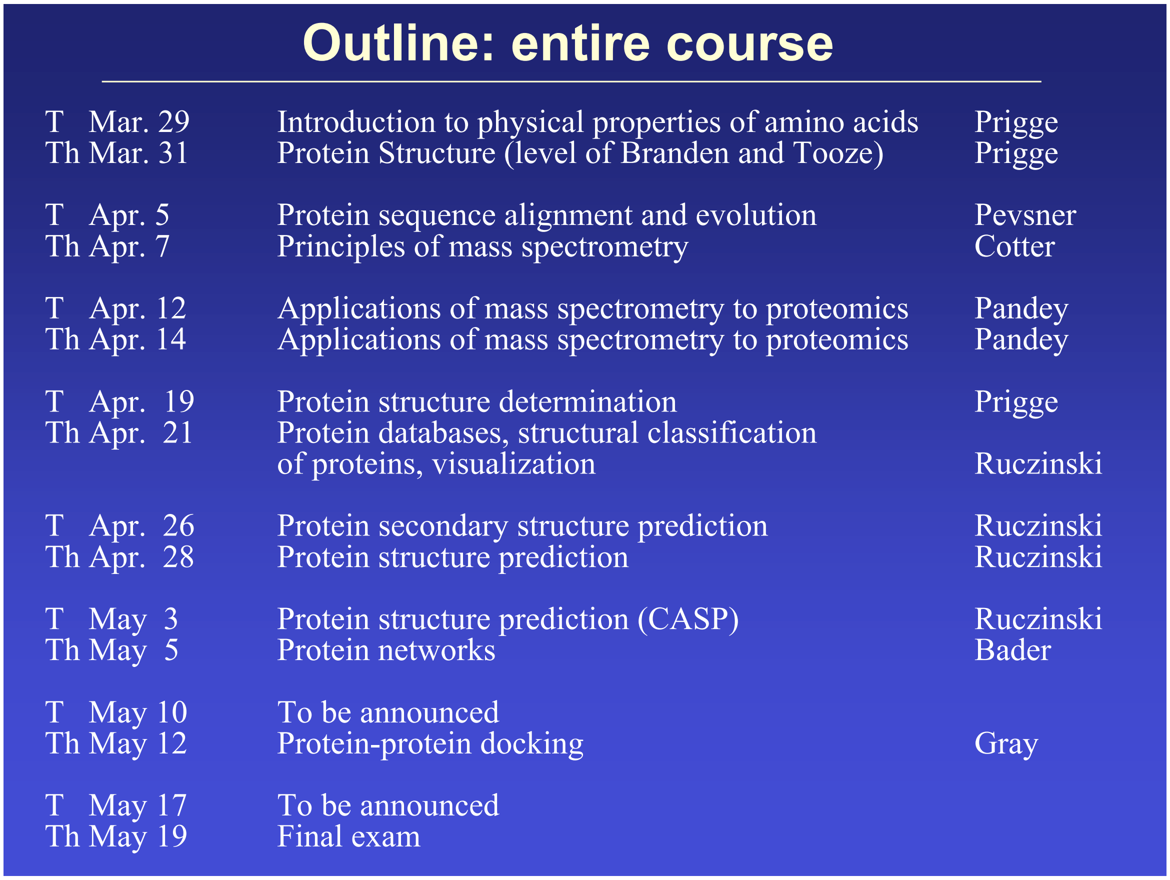


# Protein sequence alignment and evolution

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

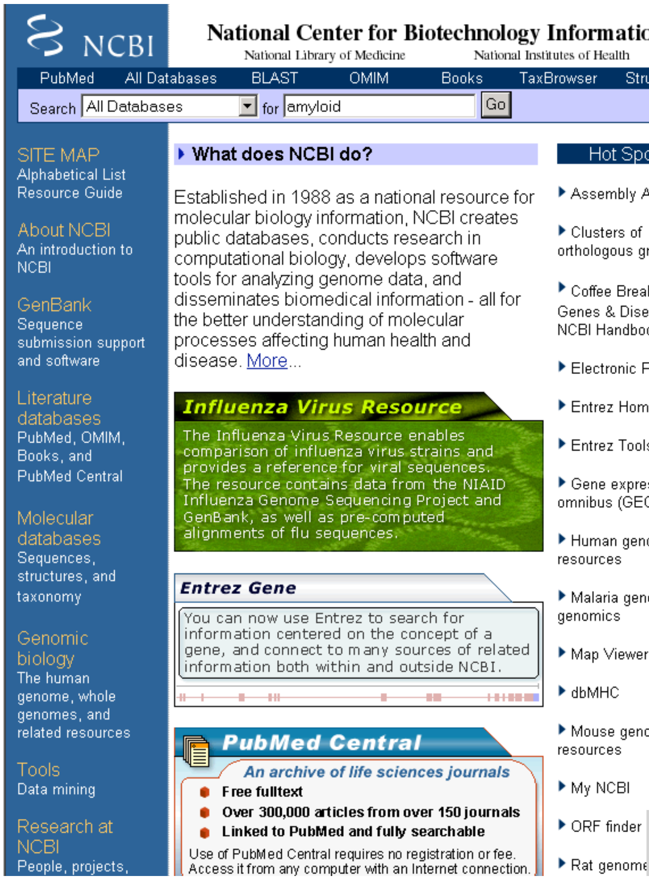
## Outline: today’s topic

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


NCBI
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
PubMed
All Databases
BLAST
OMIM
Books
TaxBrowser
Structure
Search
Go
SITE MAP
Alphabetical List
Resource Guide
About NCBI
An introduction to NCBI
GenBank
Sequence submission support and software
Literature databases
PubMed, OMIM, Books, and PubMed Central
Molecular databases
Sequences, structures, and taxonomy
Genomic biology
The human genome, whole genomes, and related resources
Tools
Data mining
Research at NCBI
People, projects,
Hot Spots
Assembly Archive
Clusters of orthologous groups
Coffee Break, Genes & Disease, NCBI Handbook
Electronic PCR
Entrez Home
Entrez Tools
Gene expression omnibus (GEO)
Human genome resources
Malaria genetics & genomics
Map Viewer
dbMHC
Mouse genome resources
My NCBI
ORF finder
Rat genome

Search All Databases for amyloid

**What does NCBI do?**

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More...

**Influenza Virus Resource**

The Influenza Virus Resource enables comparison of influenza virus strains and provides a reference for viral sequences. The resource contains data from the NIAID Influenza Genome Sequencing Project and GenBank, as well as pre-computed alignments of flu sequences.

**Entrez Gene**

You can now use Entrez to search for information centered on the concept of a gene, and connect to many sources of related information both within and outside NCBI.

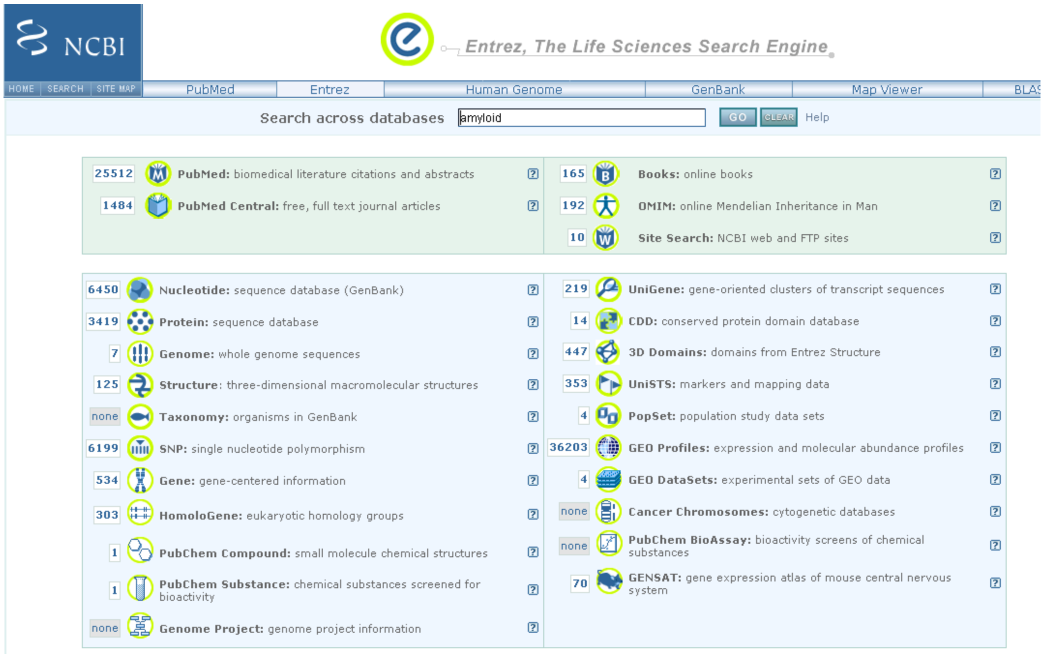
**PubMed Central**

*An archive of life sciences journals*

- Free fulltext
- Over 300,000 articles from over 150 journals
- Linked to PubMed and fully searchable

Use of PubMed Central requires no registration or fee. Access it from any computer with an Internet connection.

**www.ncbi.nlm.nih.gov**

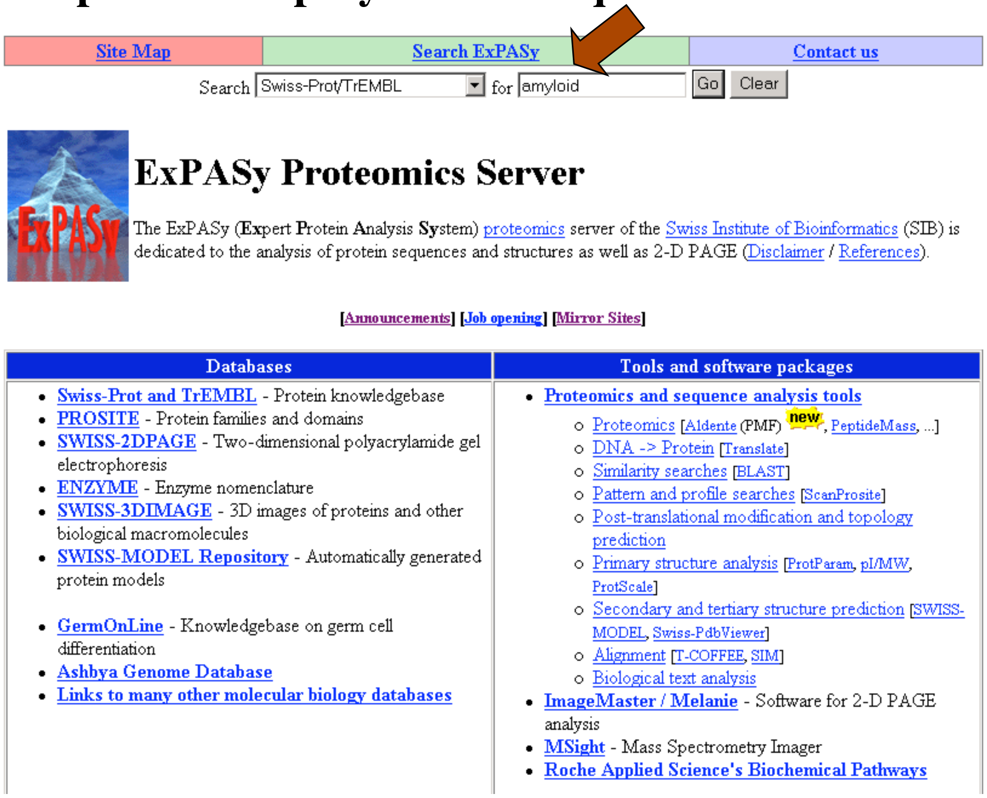


Search across databases amyloid

| | | | |
|---|---|---|---|
| 25512 | **PubMed:** biomedical literature citations and abstracts | 165 | **Books:** online books |
| 1484 | **PubMed Central:** free, full text journal articles | 192 | **OMIM:** online Mendelian Inheritance in Man |
| | | 10 | **Site Search:** NCBI web and FTP sites |

| | | | |
|---|---|---|---|
| 6450 | **Nucleotide:** sequence database (GenBank) | 219 | **UniGene:** gene-oriented clusters of transcript sequences |
| 3419 | **Protein:** sequence database | 14 | **CDD:** conserved protein domain database |
| 7 | **Genome:** whole genome sequences | 447 | **3D Domains:** domains from Entrez Structure |
| 125 | **Structure**: three-dimensional macromolecular structures | 353 | **UniSTS:** markers and mapping data |
| none | **Taxonomy:** organisms in GenBank | 4 | **PopSet:** population study data sets |
| 6199 | **SNP:** single nucleotide polymorphism | 36203 | **GEO Profiles:** expression and molecular abundance profiles |
| 534 | **Gene:** gene-centered information | 4 | **GEO DataSets:** experimental sets of GEO data |
| 303 | **HomoloGene:** eukaryotic homology groups | none | **Cancer Chromosomes:** cytogenetic databases |
| 1 | **PubChem Compound:** small molecule chemical structures | none | **PubChem BioAssay:** bioactivity screens of chemical substances |
| 1 | **PubChem Substance:** chemical substances screened for bioactivity | 70 | **GENSAT:** gene expression atlas of mouse central nervous system |
| none | **Genome Project:** genome project information | | |

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

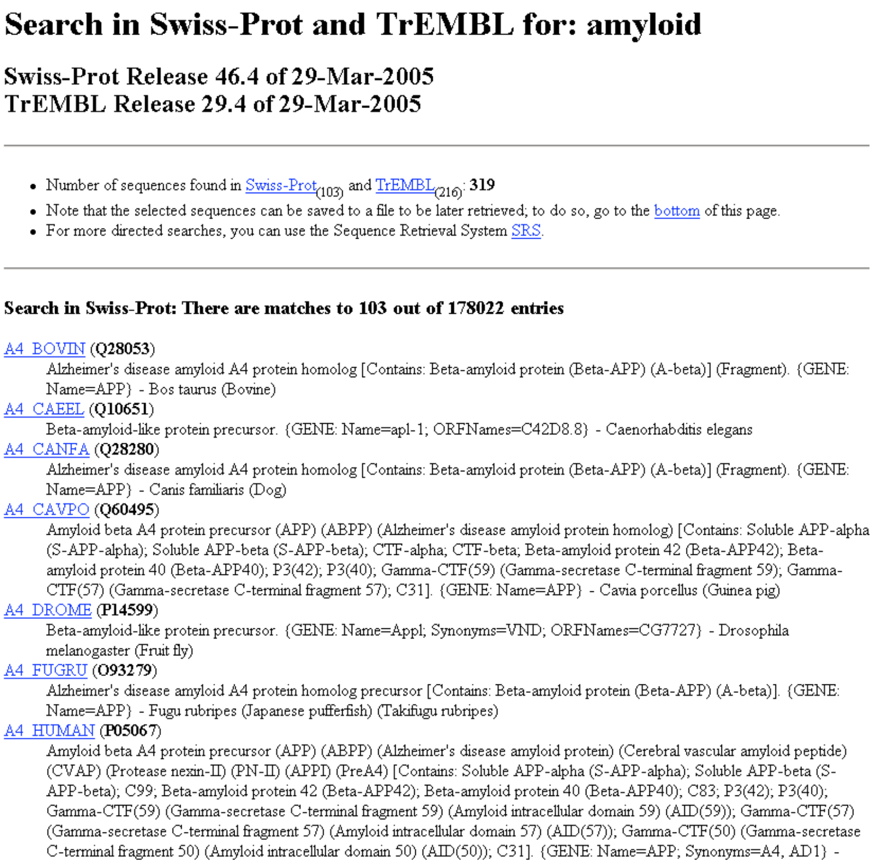


## ExPASy Proteomics Server

The ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE (Disclaimer / References).

[Announcements] [Job opening] [Mirror Sites]

| Databases | Tools and software packages |
|---|---|
| - Swiss-Prot and TrEMBL - Protein knowledgebase<br>- PROSITE - Protein families and domains<br>- SWISS-2DPAGE - Two-dimensional polyacrylamide gel electrophoresis<br>- ENZYME - Enzyme nomenclature<br>- SWISS-3DIMAGE - 3D images of proteins and other biological macromolecules<br>- SWISS-MODEL Repository - Automatically generated protein models<br>- GermOnLine - Knowledgebase on germ cell differentiation<br>- Ashbya Genome Database<br>- Links to many other molecular biology databases | - Proteomics and sequence analysis tools<br>○ Proteomics [Aldente (PMF) new, PeptideMass, ...]<br>○ DNA -> Protein [Translate]<br>○ Similarity searches [BLAST]<br>○ Pattern and profile searches [ScanProsite]<br>○ Post-translational modification and topology prediction<br>○ Primary structure analysis [ProtParam, pI/MW, ProtScale]<br>○ Secondary and tertiary structure prediction [SWISS-MODEL, Swiss-PdbViewer]<br>○ Alignment [T-COFFEE, SIM]<br>○ Biological text analysis<br>- ImageMaster / Melanie - Software for 2-D PAGE analysis<br>- MSight - Mass Spectrometry Imager<br>- Roche Applied Science's Biochemical Pathways |



## 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)

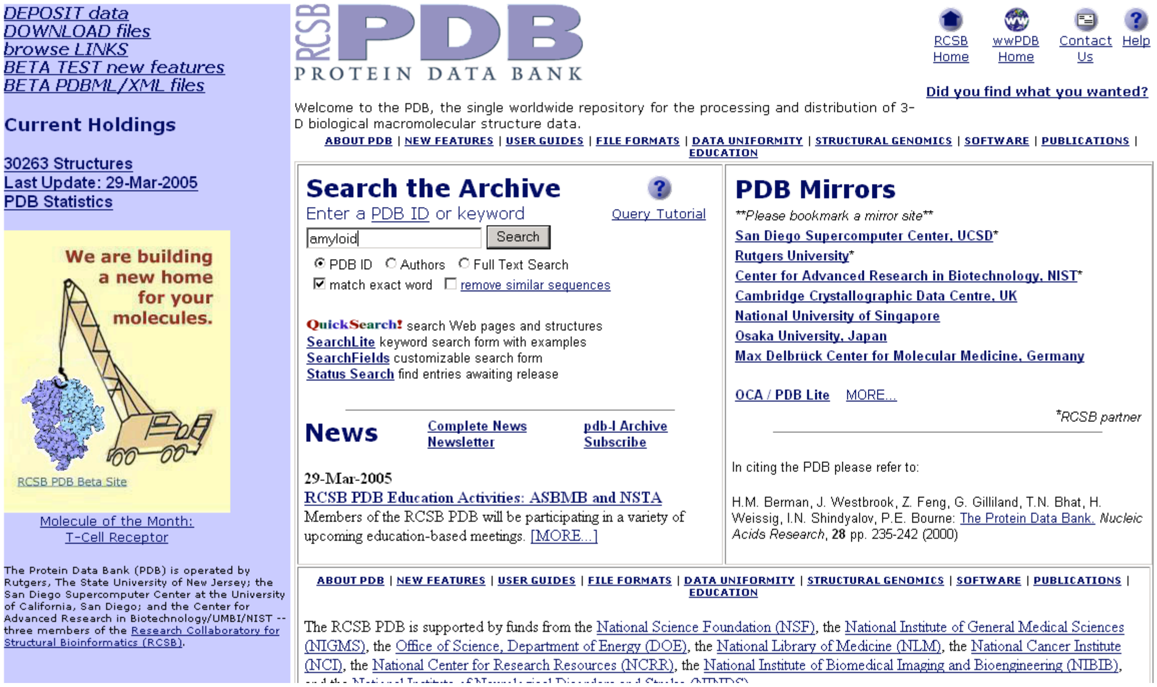
A4_CAVPO (**Q60495**)
Amyloid beta A4 protein precursor (APP) (ABPP) (Alzheimer's disease amyloid protein homolog) [Contains: Soluble APP-alpha (S-APP-alpha); Soluble APP-beta (S-APP-beta); CTF-alpha; CTF-beta; Beta-amyloid protein 42 (Beta-APP42); Beta-amyloid protein 40 (Beta-APP40); P3(42); P3(40); 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)


DEPOSIT data
DOWNLOAD files
browse LINKS
BETA TEST new features
BETA PDBML/XML files
RCSB PDB PROTEIN DATA BANK
RCSB Home
wwPDB Home
Contact Us
Help
Did you find what you wanted?
ABOUT PDB | NEW FEATURES | USER GUIDES | FILE FORMATS | DATA UNIFORMITY | STRUCTURAL GENOMICS | SOFTWARE | PUBLICATIONS | EDUCATION
Search
We are building a new home for your molecules.
RCSB PDB Beta Site

Current Holdings

30263 Structures
Last Update: 29-Mar-2005
PDB Statistics

Molecule of the Month:
T-Cell Receptor

The Protein Data Bank (PDB) is operated by Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the Center for Advanced Research in Biotechnology/UMBI/NIST -- three members of the Research Collaboratory for Structural Bioinformatics (RCSB).

Welcome to the PDB, the single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data.

Search the Archive

Enter a PDB ID or keyword  Query Tutorial

amyloid

● PDB ID ○ Authors ○ Full Text Search
☑ match exact word ☐ remove similar sequences

QuickSearch! search Web pages and structures
SearchLite keyword search form with examples
SearchFields customizable search form
Status Search find entries awaiting release

News  Complete News Newsletter  pdb-l Archive Subscribe

29-Mar-2005
RCSB PDB Education Activities: ASBMB and NSTA
Members of the RCSB PDB will be participating in a variety of upcoming education-based meetings. [MORE...]

PDB Mirrors

**Please bookmark a mirror site**

San Diego Supercomputer Center, UCSD*
Rutgers University*
Center for Advanced Research in Biotechnology, NIST*
Cambridge Crystallographic Data Centre, UK
National University of Singapore
Osaka University, Japan
Max Delbrück Center for Molecular Medicine, Germany

OCA / PDB Lite   MORE...

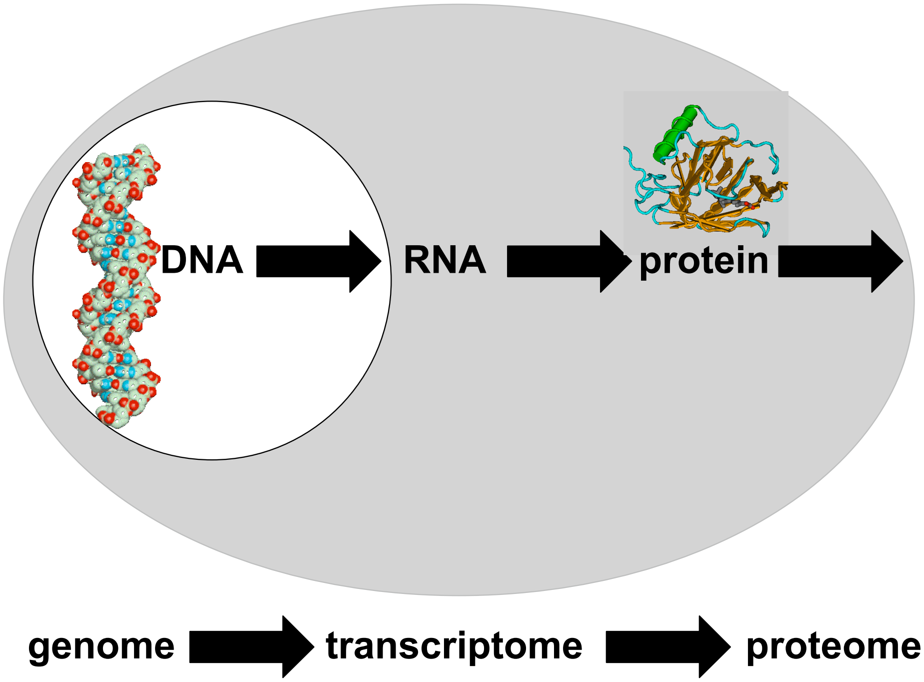
*RCSB partner

In citing the PDB please refer to:

H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The Protein Data Bank. *Nucleic Acids Research*, **28** pp. 235-242 (2000)

The RCSB PDB is supported by funds from the National Science Foundation (NSF), the National Institute of General Medical Sciences (NIGMS), the Office of Science, Department of Energy (DOE), the National Library of Medicine (NLM), the National Cancer Institute (NCI), the National Center for Research Resources (NCRR), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and the National Institute of Neurological Disorders and Stroke (NINDS).

## Central dogma of molecular biology



DNA → RNA → protein →

genome → transcriptome → proteome

## 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

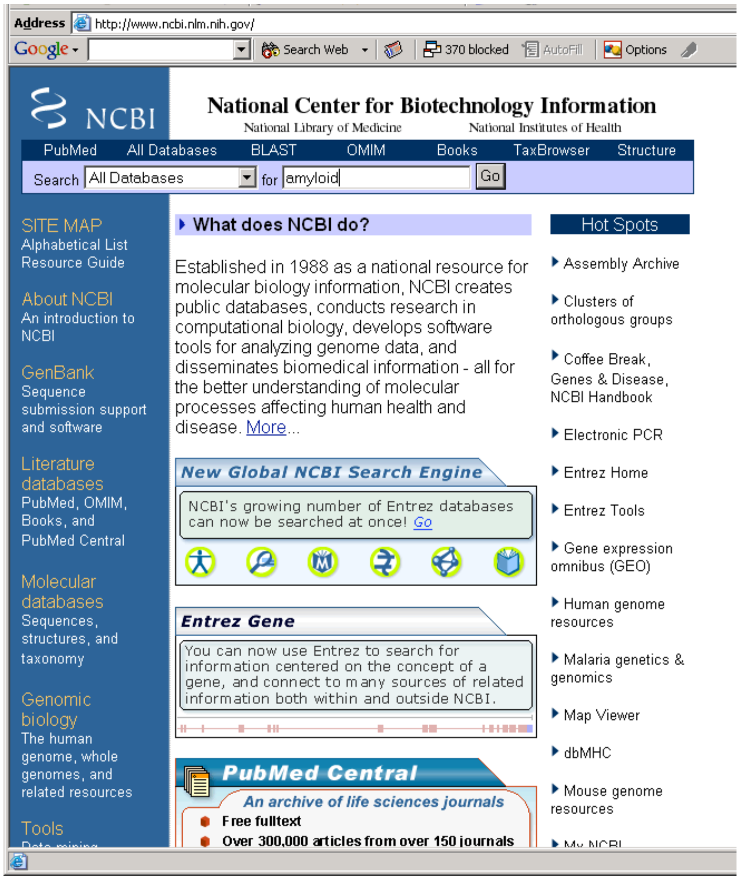
RefSeq (accessible via the main page of NCBI) provides an expertly curated accession number that corresponds to the most stable, agreed-upon "reference" version of a sequence.

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 |

Page 29-30

Example: type "amyloid" at NCBI



Address
http://www.ncbi.nlm.nih.gov/
Google
Search Web
370 blocked
AutoFill
Options
NCBI
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
PubMed
All Databases
BLAST
OMIM
Books
TaxBrowser
Structure
Search
All Databases
for
amyloid
Go
SITE MAP
Alphabetical List
Resource Guide
About NCBI
An introduction to NCBI
GenBank
Sequence submission support and software
Literature databases
PubMed, OMIM, Books, and PubMed Central
Molecular databases
Sequences, structures, and taxonomy
Genomic biology
The human genome, whole genomes, and related resources
Tools
Hot Spots
Assembly Archive
Clusters of orthologous groups
Coffee Break, Genes & Disease, NCBI Handbook
Electronic PCR
Entrez Home
Entrez Tools
Gene expression omnibus (GEO)
Human genome resources
Malaria genetics & genomics
Map Viewer
dbMHC
Mouse genome resources

What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More...

New Global NCBI Search Engine

NCBI's growing number of Entrez databases can now be searched at once! Go

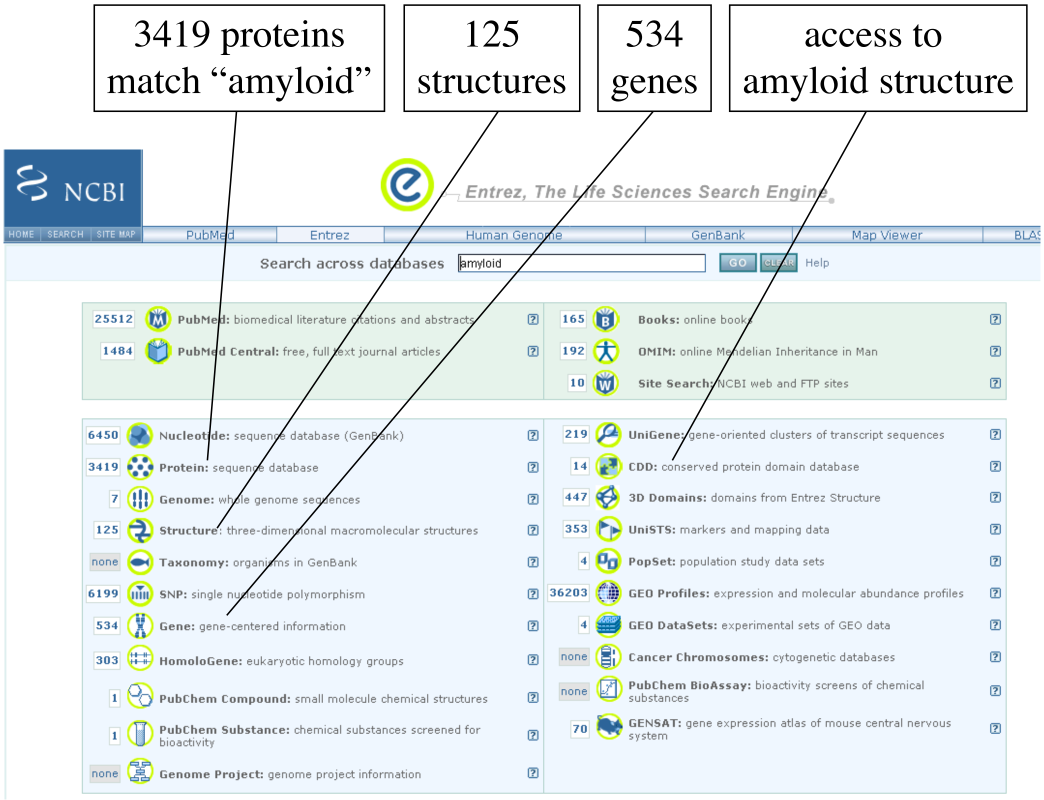
Entrez Gene

You can now use Entrez to search for information centered on the concept of a gene, and connect to many sources of related information both within and outside NCBI.

PubMed Central

An archive of life sciences journals

- Free fulltext
- Over 300,000 articles from over 150 journals



3419 proteins match “amyloid”

125 structures

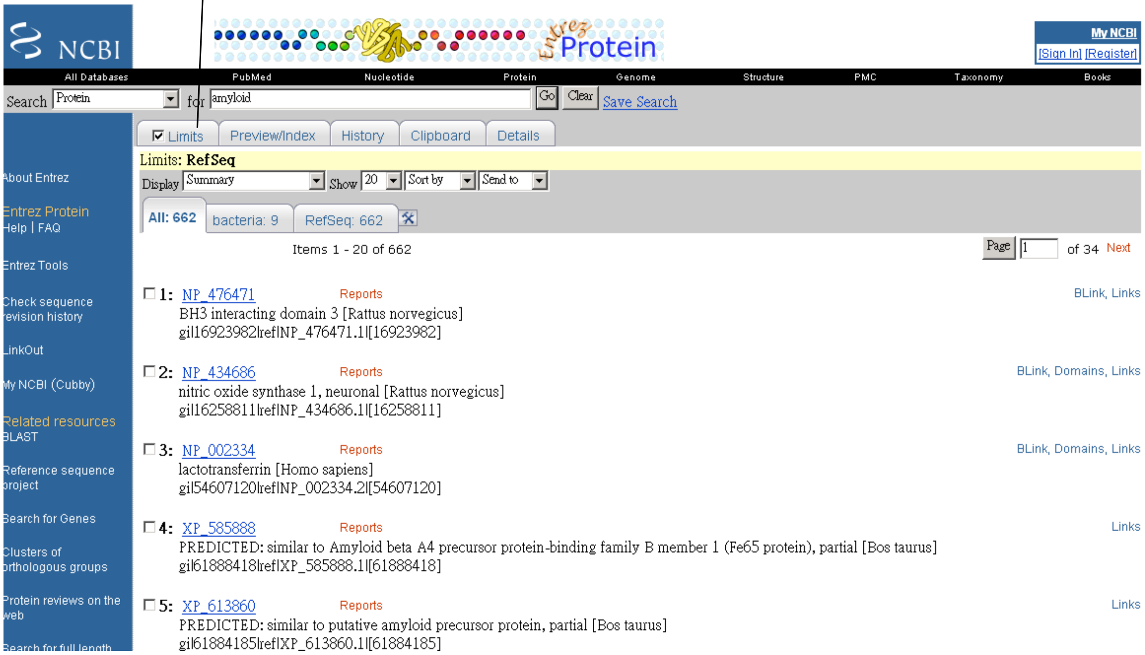
534 genes

access to amyloid structure

Search across databases amyloid GO CLEAR Help

| | | | |
|---|---|---|---|
| 25512 | **PubMed:** biomedical literature citations and abstracts | 165 | **Books:** online books |
| 1484 | **PubMed Central:** free, full text journal articles | 192 | **OMIM:** online Mendelian Inheritance in Man |
| | | 10 | **Site Search:** NCBI web and FTP sites |
| 6450 | **Nucleotide:** sequence database (GenBank) | 219 | **UniGene:** gene-oriented clusters of transcript sequences |
| 3419 | **Protein:** sequence database | 14 | **CDD:** conserved protein domain database |
| 7 | **Genome:** whole genome sequences | 447 | **3D Domains:** domains from Entrez Structure |
| 125 | **Structure:** three-dimensional macromolecular structures | 353 | **UniSTS:** markers and mapping data |
| none | **Taxonomy:** organisms in GenBank | 4 | **PopSet:** population study data sets |
| 6199 | **SNP:** single nucleotide polymorphism | 36203 | **GEO Profiles:** expression and molecular abundance profiles |
| 534 | **Gene:** gene-centered information | 4 | **GEO DataSets:** experimental sets of GEO data |
| 303 | **HomoloGene:** eukaryotic homology groups | none | **Cancer Chromosomes:** cytogenetic databases |
| 1 | **PubChem Compound:** small molecule chemical structures | none | **PubChem BioAssay:** bioactivity screens of chemical substances |
| 1 | **PubChem Substance:** chemical substances screened for bioactivity | 70 | **GENSAT:** gene expression atlas of mouse central nervous system |
| none | **Genome Project:** genome project information | | |

Click “protein” to find 3419 records for amyloid.
Further limit the search to RefSeq only, then to human.



Search Protein for amyloid

Limits: **RefSeq**

Display Summary Show 20 Sort by Send to

All: 662 bacteria: 9 RefSeq: 662

Items 1 - 20 of 662 Page 1 of 34 Next

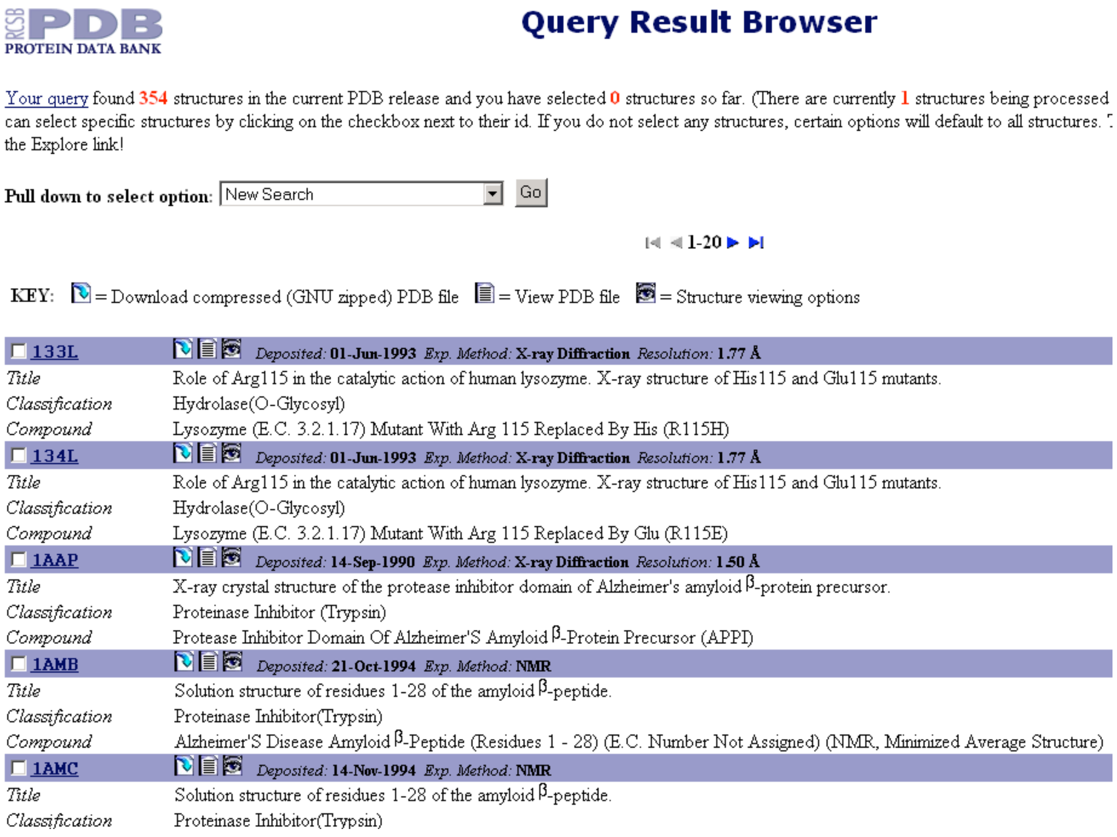
□ 1: NP_476471 Reports BLink, Links
BH3 interacting domain 3 [Rattus norvegicus]
gi|16923982|ref|NP_476471.1|[16923982]

□ 2: NP_434686 Reports BLink, Domains, Links
nitric oxide synthase 1, neuronal [Rattus norvegicus]
gi|16258811|ref|NP_434686.1|[16258811]

□ 3: NP_002334 Reports BLink, Domains, Links
lactotransferrin [Homo sapiens]
gi|54607120|ref|NP_002334.2|[54607120]

□ 4: XP_585888 Reports Links
PREDICTED: similar to Amyloid beta A4 precursor protein-binding family B member 1 (Fe65 protein), partial [Bos taurus]
gi|61888418|ref|XP_585888.1|[61888418]

□ 5: XP_613860 Reports Links
PREDICTED: similar to putative amyloid precursor protein, partial [Bos taurus]
gi|61884185|ref|XP_613860.1|[61884185]



RCSB PDB PROTEIN DATA BANK

## Query Result Browser

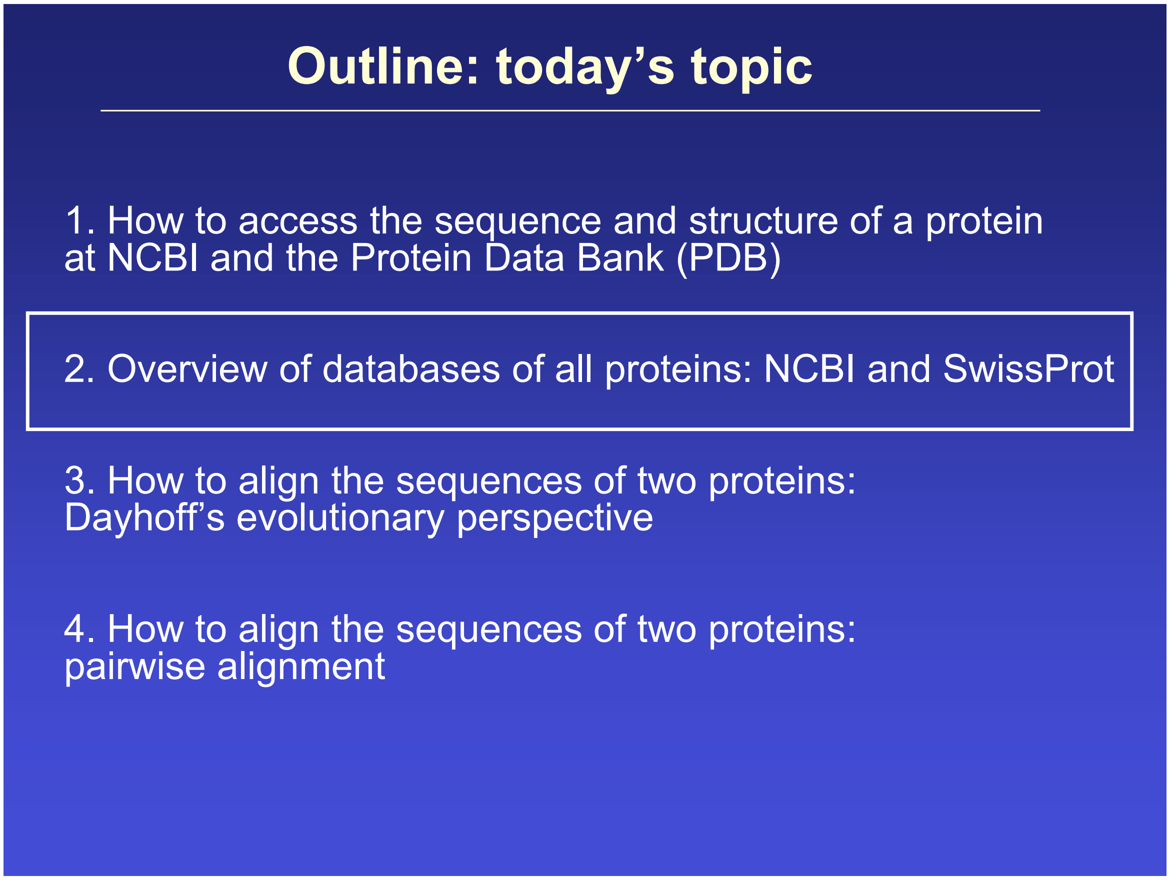
Your query found 354 structures in the current PDB release and you have selected 0 structures so far. (There are currently 1 structures being processed can select specific structures by clicking on the checkbox next to their id. If you do not select any structures, certain options will default to all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



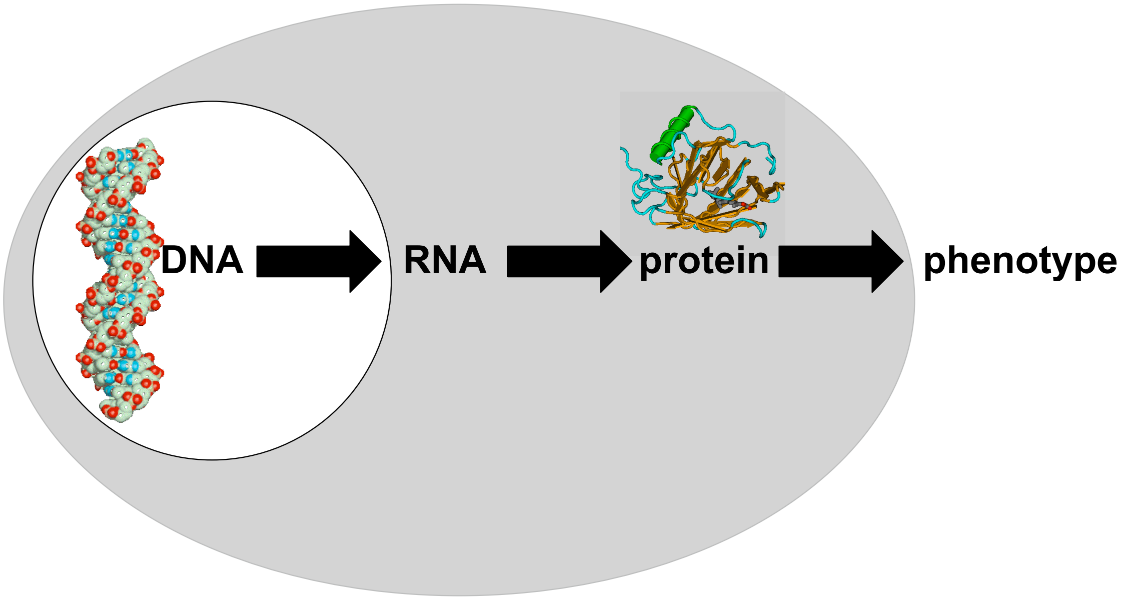
## Outline: today's topic

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



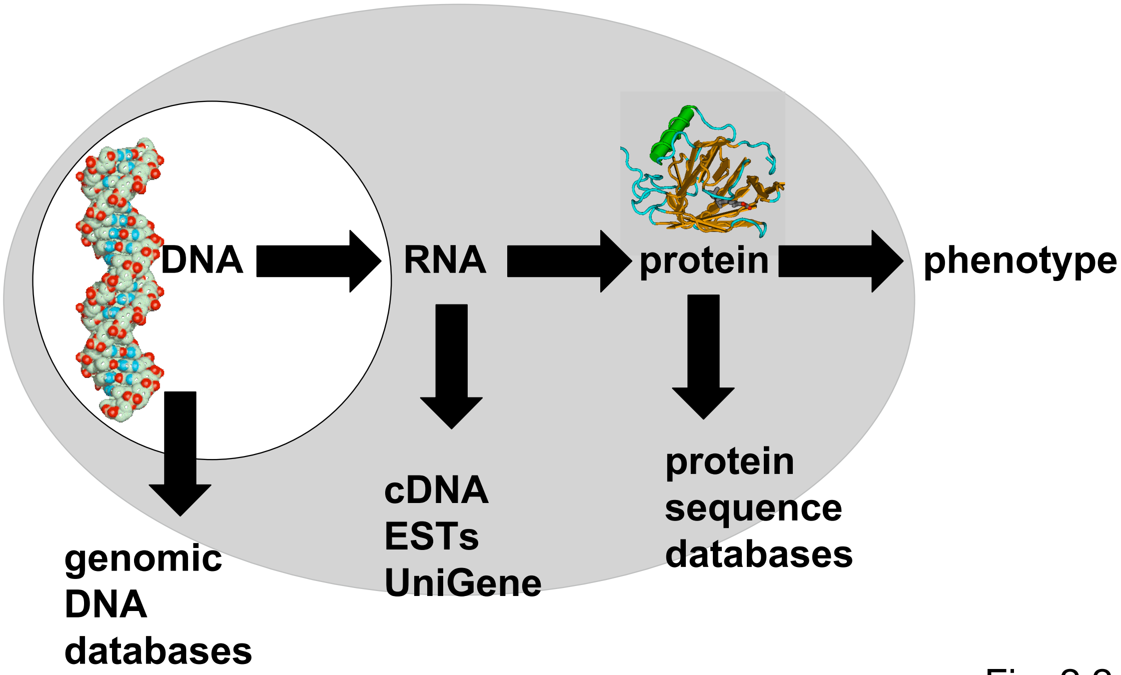


Fig. 2.2
Page 20

## DNA

| | | |
|---|---|---|
| GenBank | EBI | DDBJ |
| NCBI | EMBL<br>EUROPEAN MOLECULAR BIOLOGY LABORATORY | DDBJ |

## protein

| | | | | |
|---|---|---|---|---|
| | UniProt (www.uniprot.org) | | | |
| NCBI<br>Entrez | EBI<br>European Bioinformatics Institute | ExPASy<br>ExPASy | PIR<br>Protein Information Resource | PROTEIN DATA BANK PDB<br>Protein Data Bank |

**Growth of GenBank**

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

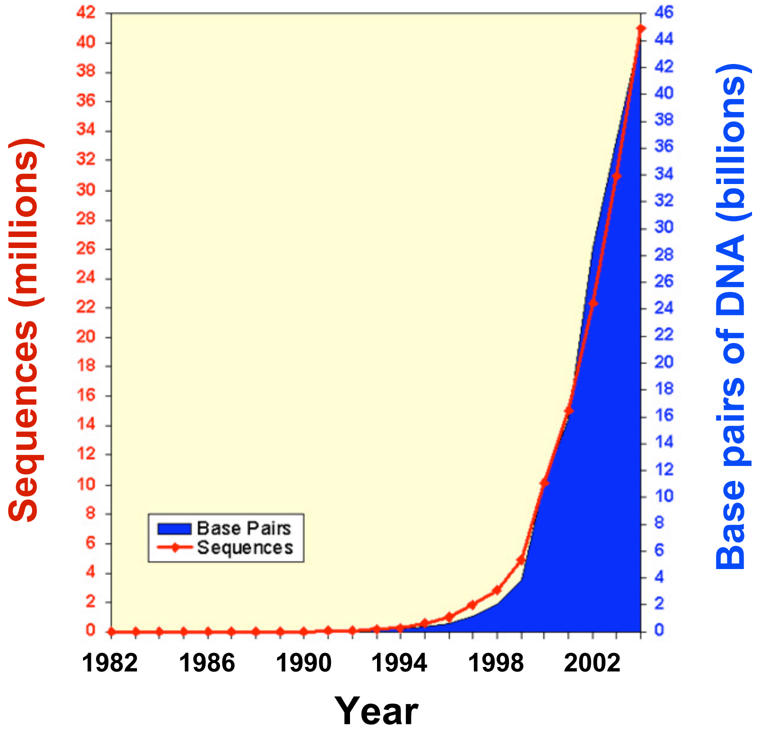
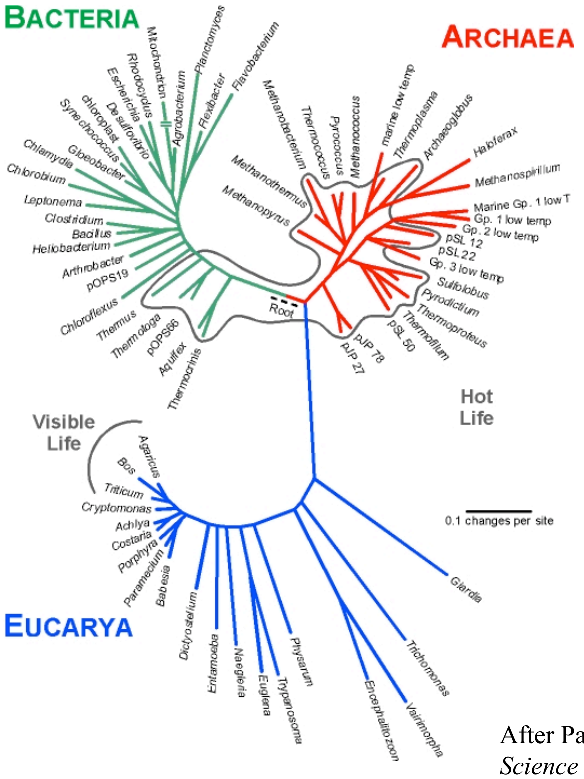
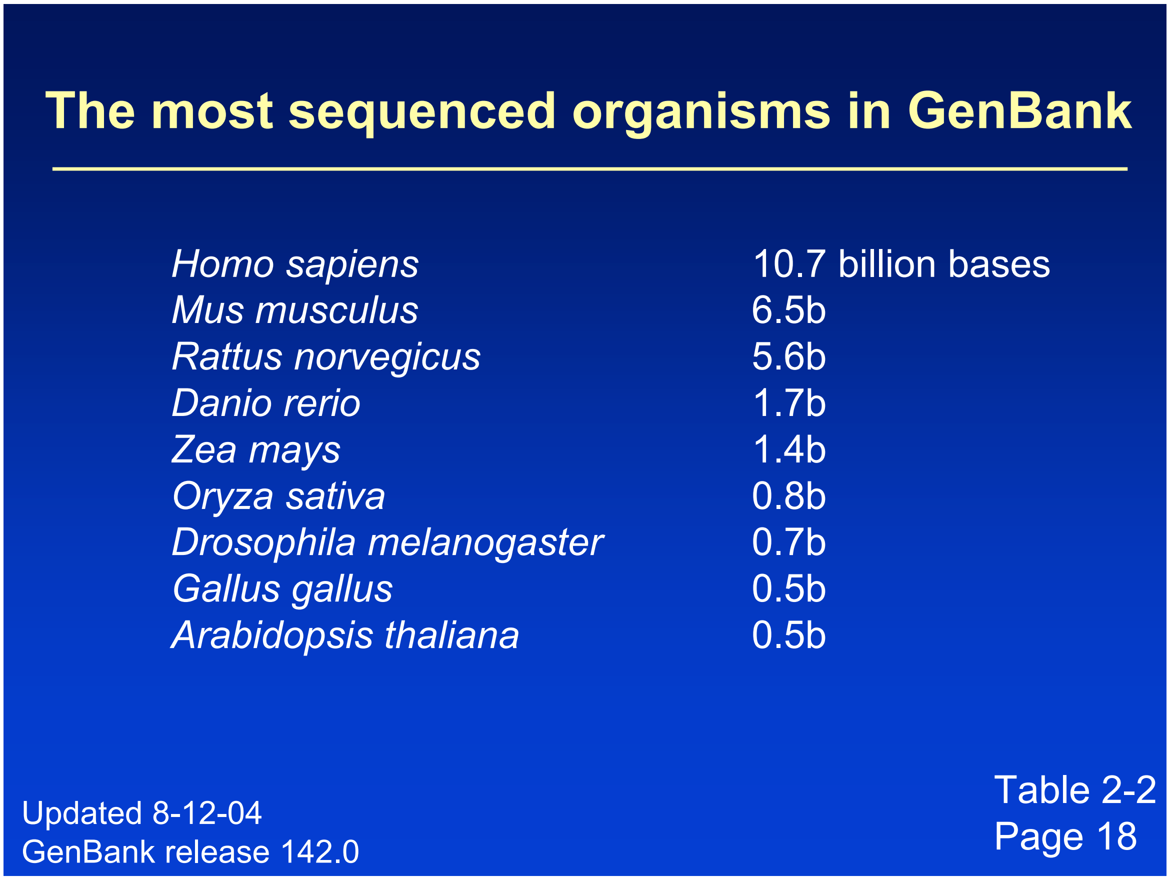


Fig. 2.1
Page 17



After Pace NR (1997)
*Science* 276:734

Page 6

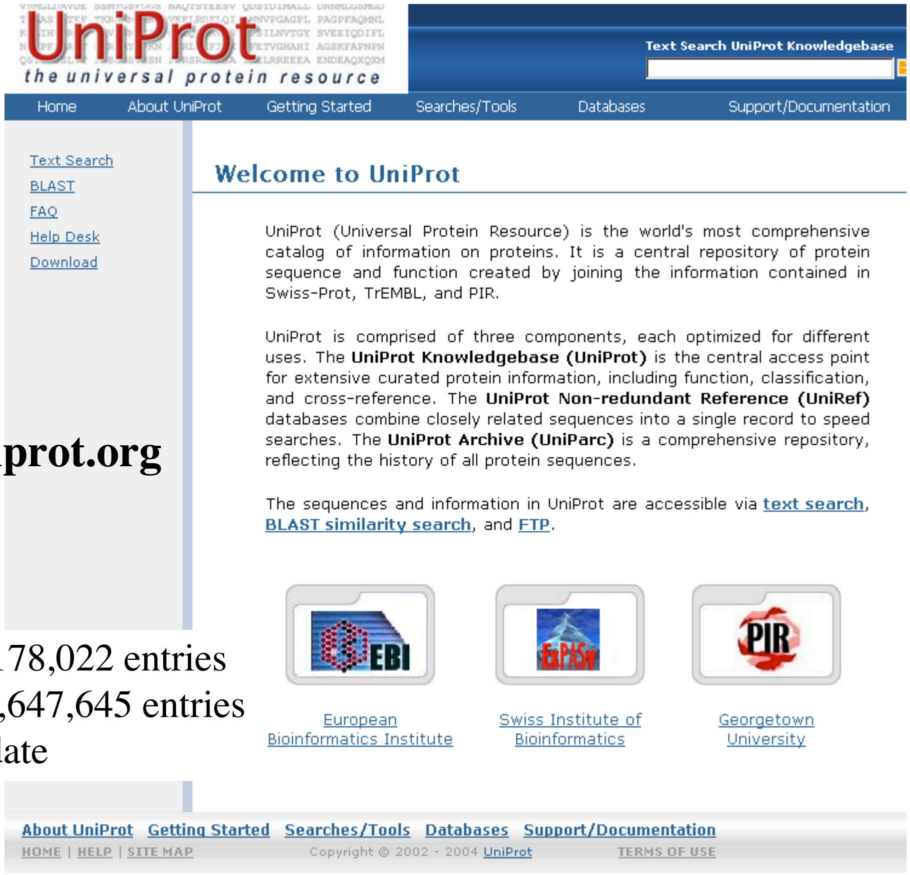


# The most sequenced organisms in GenBank

| | |
|---|---|
| *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

Table 2-2
Page 18



UniProt
the universal protein resource

Text Search UniProt Knowledgebase

Home | About UniProt | Getting Started | Searches/Tools | Databases | Support/Documentation

- Text Search
- BLAST
- FAQ
- Help Desk
- Download

## Welcome to UniProt

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 (UniRef)** 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.

European Bioinformatics Institute | Swiss Institute of Bioinformatics | Georgetown University

**About UniProt** **Getting Started** **Searches/Tools** **Databases** **Support/Documentation**

HOME | HELP | SITE MAP Copyright © 2002 - 2004 UniProt TERMS OF USE

**www.uniprot.org**

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

# PDB content growth (www.pdb.org)

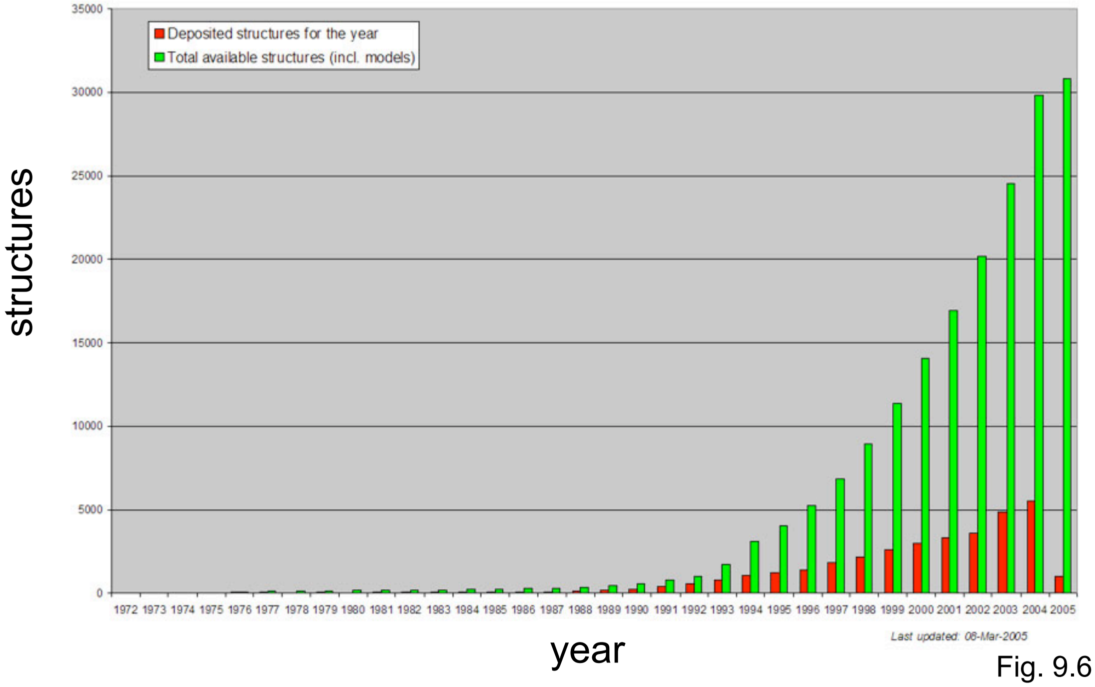


Fig. 9.6
Page 281

## 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

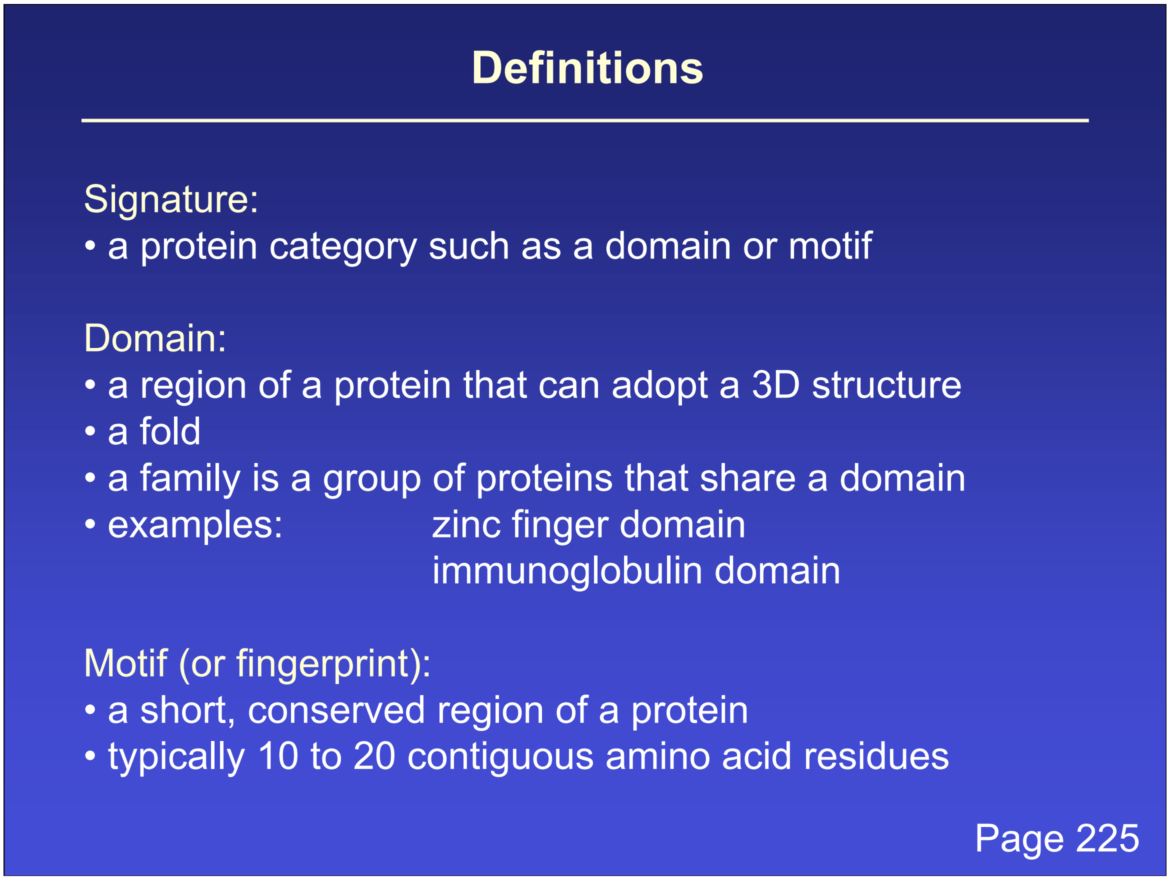
4. How to align the sequences of two proteins: pairwise alignment

## Definitions

Signature:

- a protein category such as a domain or motif

Page 225



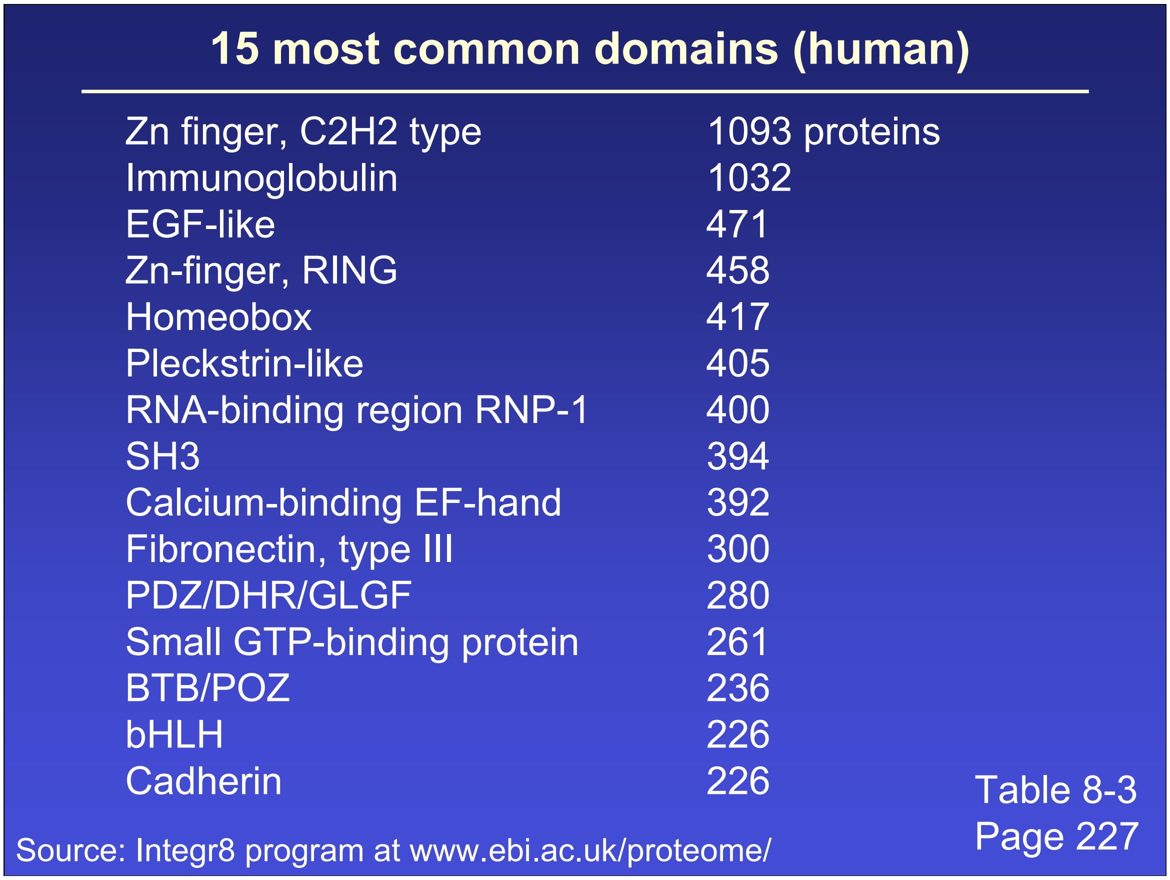
## Definitions

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 |

Table 8-3
Page 227

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

## Pairwise alignments in the 1950s

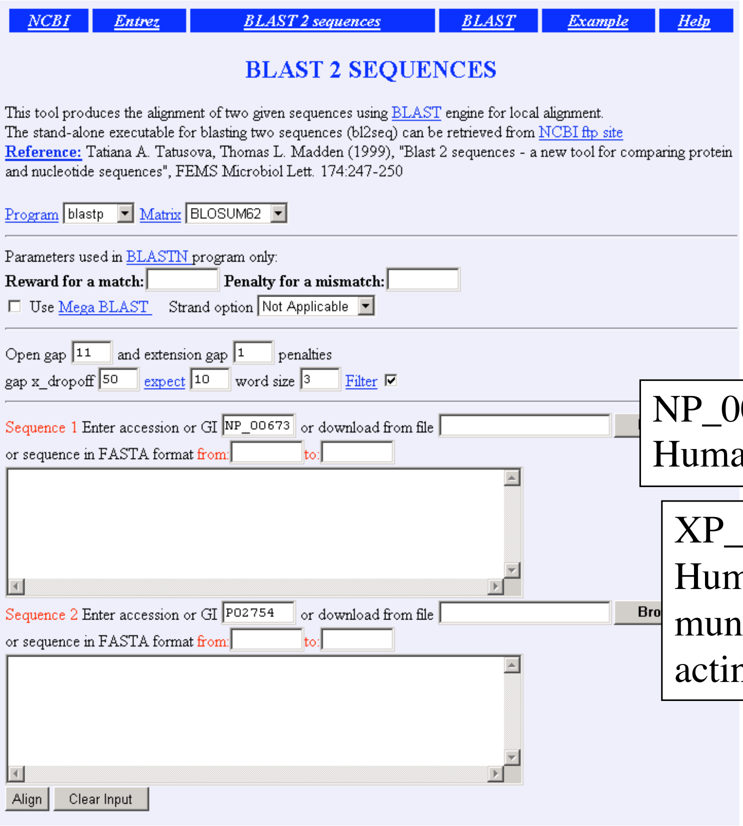
| | |
|---|---|
| **β-corticotropin (sheep)** | `ala gly glu asp asp glu` |
| **Corticotropin A (pig)** | `asp gly ala glu asp glu` |

| | |
|---|---|
| **Oxytocin** | `CYIQNCPLG` |
| **Vasopressin** | `CYFQNCPRG` |

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

## 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



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 (bl2seq) can be retrieved from NCBI ftp site
**Reference:** Tatiana A. Tatusova, Thomas L. Madden (1999), "Blast 2 sequences - a new tool for comparing protein and nucleotide sequences", FEMS Microbiol Lett. 174:247-250

Program blastp Matrix BLOSUM62

Parameters used in BLASTN program only:
**Reward for a match:** **Penalty for a mismatch:**
□ Use Mega BLAST Strand option Not Applicable

Open gap 11 and extension gap 1 penalties
gap x_dropoff 50 expect 10 word size 3 Filter ☑

Sequence 1 Enter accession or GI NP_00673 or download from file
or sequence in FASTA format from: to:

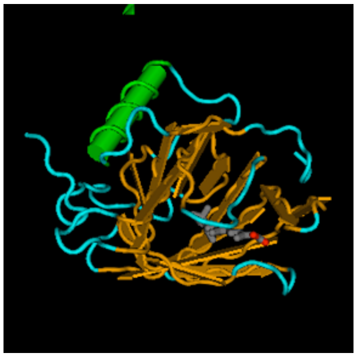
Sequence 2 Enter accession or GI P02754 or download from file Bro
or sequence in FASTA format from: to:

NP_005494
Human amyloid β

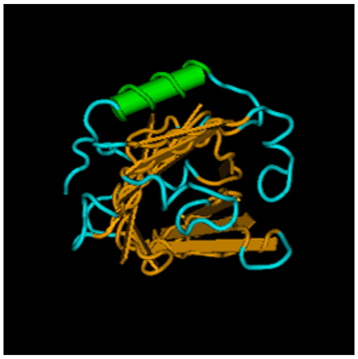
XP_372565
Human neuronal munc18-1-interacting protein 2

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
```

Page 44

# Definitions: two types of homology

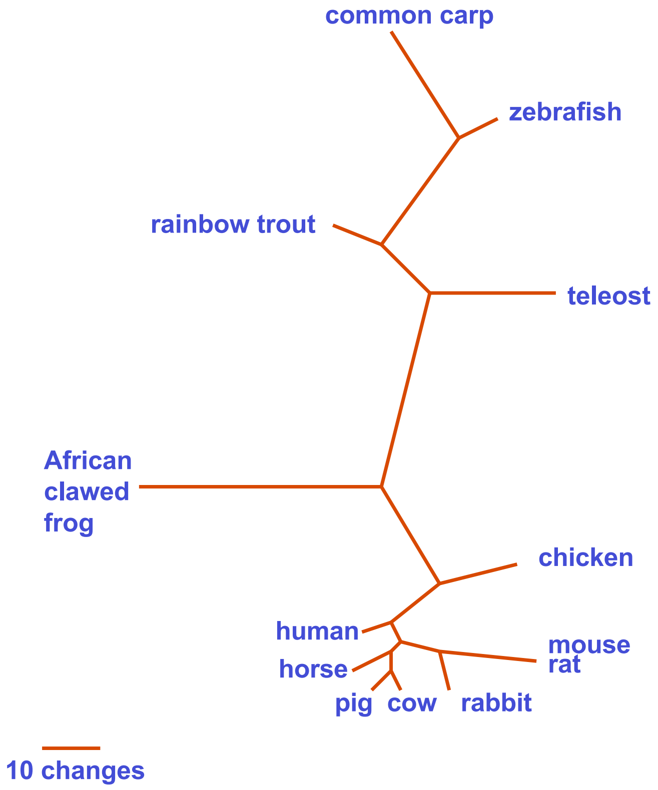
## Orthologs

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.

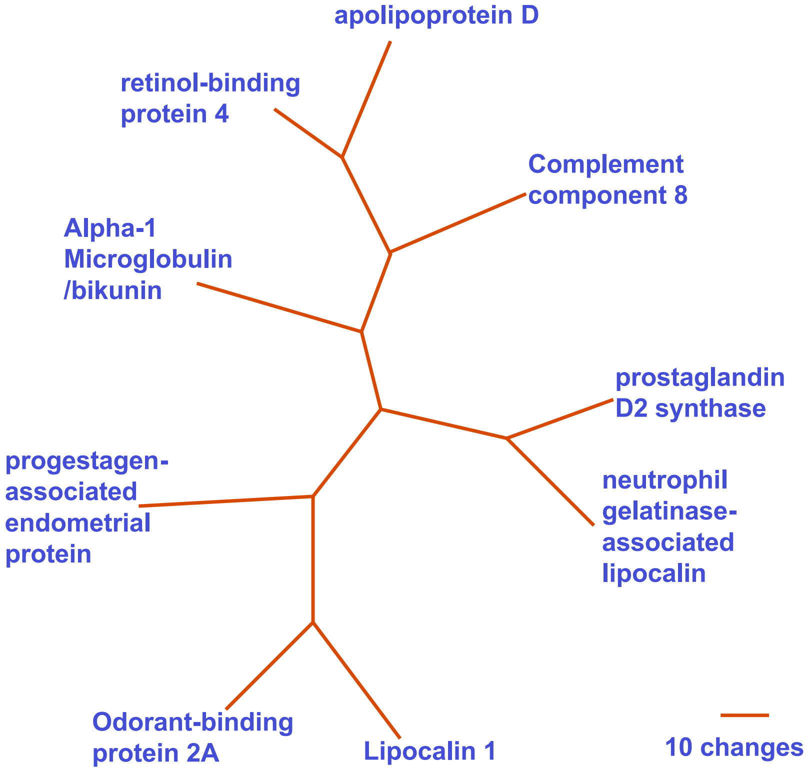
Page 43



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

This tree shows 13 RBP orthologs.

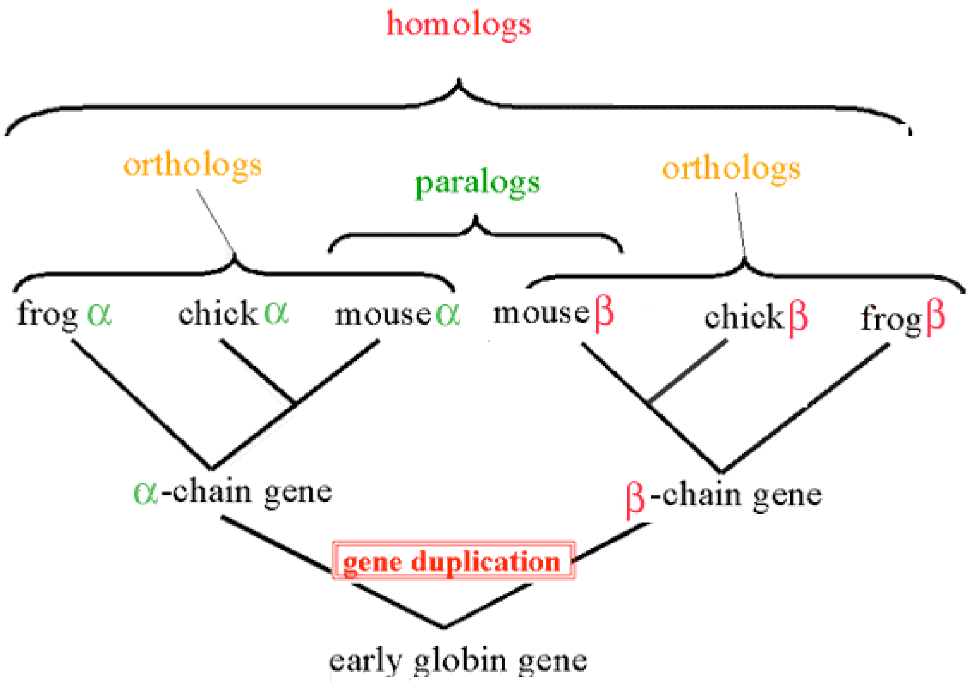
Page 43
Fig. 3.2


apolipoprotein D
retinol-binding protein 4
Complement component 8
Alpha-1 Microglobulin /bikunin
prostaglandin D2 synthase
neutrophil gelatinase-associated lipocalin
progestagen-associated endometrial protein
Odorant-binding protein 2A
Lipocalin 1
10 changes

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

This tree shows 9 human lipocalins.

Page 44
Fig. 3.3



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

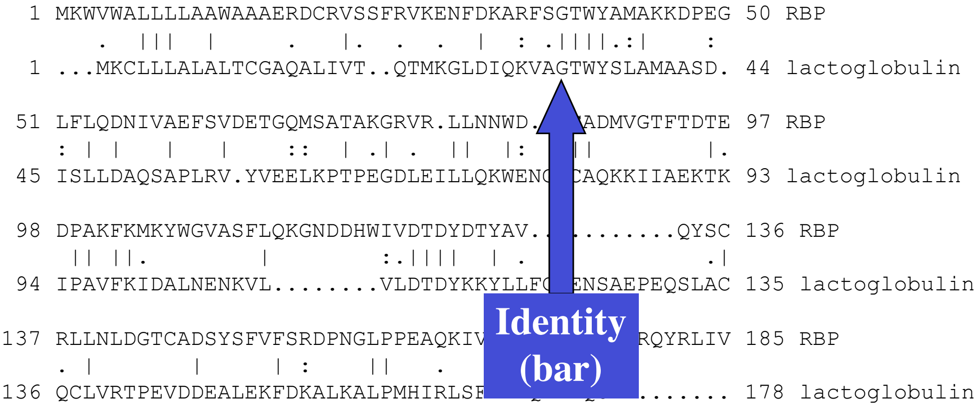
The extent to which two sequences are invariant.

## 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

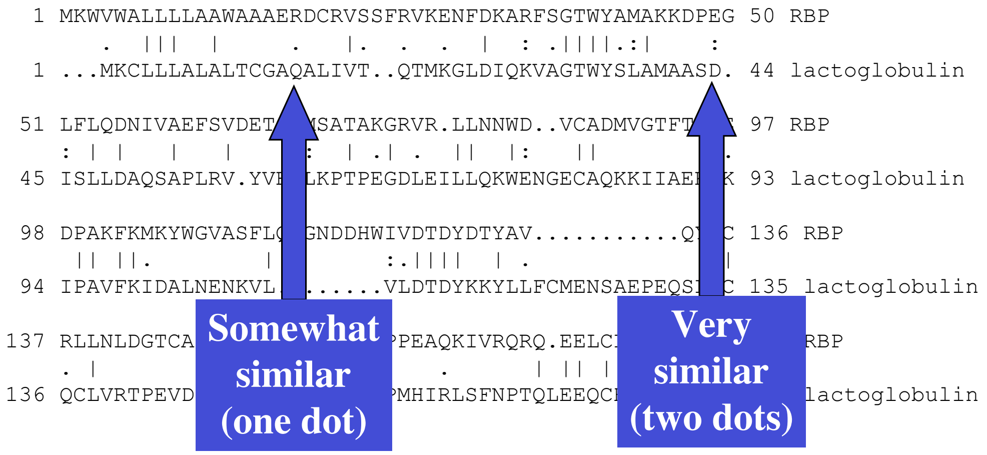


```
  1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP
       .  |||  |     .   |.  .  .  |  : .||||.:|    :
  1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

 51 LFLQDNIVAEFSVDETGQMSATAKGRVR.LLNNWD.   ADMVGTFTDTE 97 RBP
    : | |   |   |    ::  | .| .  ||  |:   ||        |.
 45 ISLLDAQSAPLRV.YVEELKPTPEGDLEILLQKWENG CAQKKIIAEKTK 93 lactoglobulin

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

137 RLLNLDGTCADSYSFVFSRDPNGLPPEAQKIV           RQYRLIV 185 RBP
    . |       |     | :    ||   .
136 QCLVRTPEVDDEALEKFDKALKALPMHIRLSF           ....... 178 lactoglobulin
```

Identity (bar)

Page 46
Fig. 3.5

## Pairwise alignment of retinol-binding protein and $\beta$-lactoglobulin



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

 51 LFLQDNIVAEFSVDET  MSATAKGRVR.LLNNWD..VCADMVGTFT    97 RBP
    : | |   |   |     :  | .| .  ||  |:   ||         .
 45 ISLLDAQSAPLRV.YV  LKPTPEGDLEILLQKWENGECAQKKIIAE  K 93 lactoglobulin

 98 DPAKFKMKYWGVASFL  GNDDHWIVDTDYDTYAV...........Q  C 136 RBP
     || ||.        |        :.||||  | .              |
 94 IPAVFKIDALNENKVL  ......VLDTDYKKYLLFCMENSAEPEQS  C 135 lactoglobulin

137 RLLNLDGTCA               PEAQKIVRQRQ.EELC              RBP
    . |                         .      | || |
136 QCLVRTPEVD               MHIRLSFNPTQLEEQC              lactoglobulin
```

Somewhat similar (one dot)

Very similar (two dots)

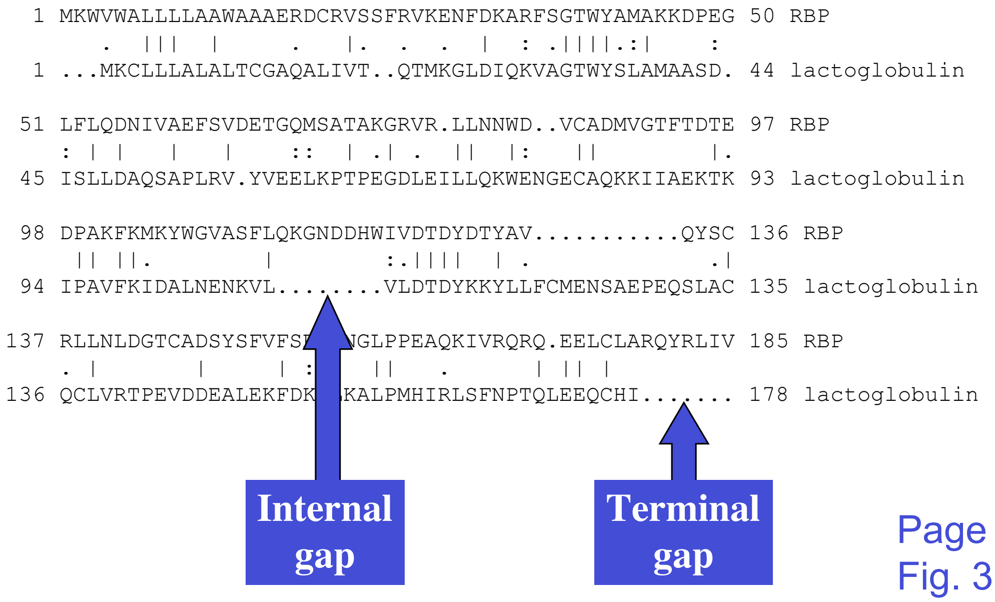
Page 46
Fig. 3.5

## 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 RLLNLDGTCADSYSFVFS[illegible]NGLPPEAQKIVRQRQ.EELCLARQYRLIV 185 RBP
    . |       |     |  :    ||   .      | || |
136 QCLVRTPEVDDEALEKFDK[illegible]KALPMHIRLSFNPTQLEEQCHI....... 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

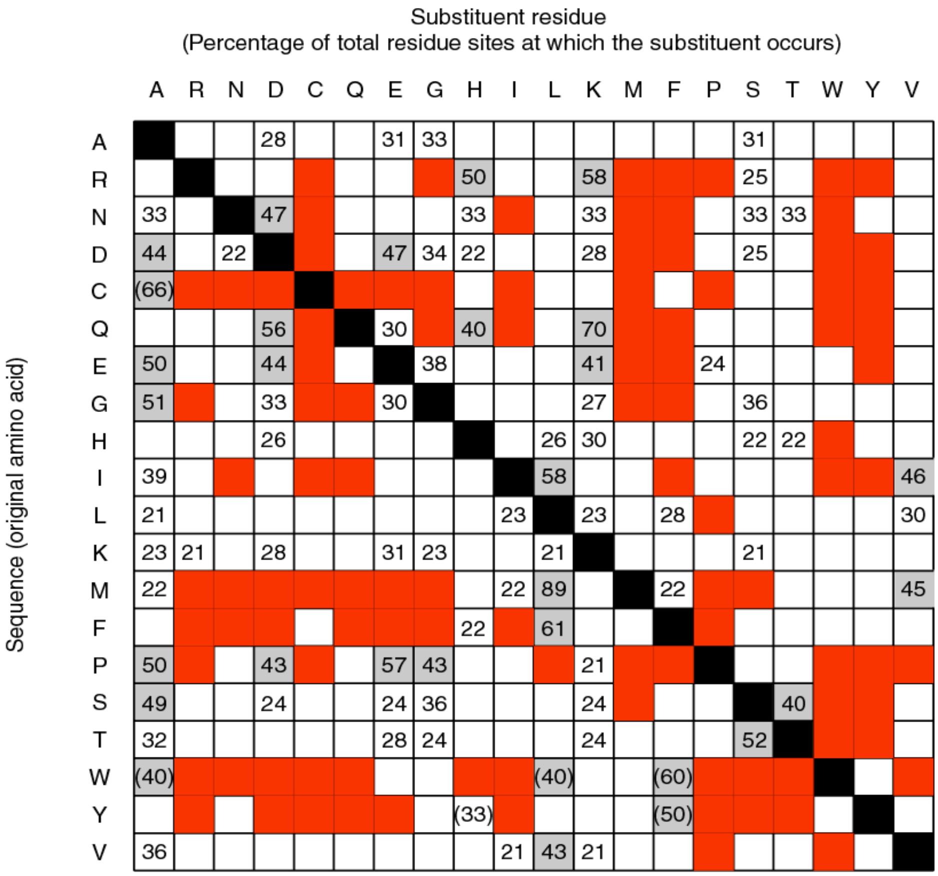
3. How to align the sequences of two proteins: Dayhoff’s evolutionary perspective

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.

Page 80



Substituent residue
(Percentage of total residue sites at which the substituent occurs)

Sequence (original amino acid)

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | ■ | | | 28 | | | 31 | 33 | | | | | | | | 31 | | | | |
| R | | ■ | | | 🟥 | | | 🟥 | 50 | | | 58 | 🟥 | 🟥 | 🟥 | 25 | | 🟥 | 🟥 | |
| N | 33 | | ■ | 47 | 🟥 | | | | 33 | 🟥 | | 33 | 🟥 | 🟥 | | 33 | 33 | 🟥 | | |
| D | 44 | | 22 | ■ | 🟥 | | 47 | 34 | 22 | | | 28 | 🟥 | 🟥 | | 25 | | 🟥 | 🟥 | |
| C | (66) | 🟥 | 🟥 | 🟥 | ■ | 🟥 | 🟥 | 🟥 | | 🟥 | | | 🟥 | | 🟥 | | | 🟥 | 🟥 | |
| Q | | | | 56 | 🟥 | ■ | 30 | 🟥 | 40 | 🟥 | | 70 | 🟥 | 🟥 | | | | 🟥 | 🟥 | |
| E | 50 | | | 44 | 🟥 | | ■ | 38 | | | | 41 | 🟥 | 🟥 | 24 | | | | 🟥 | |
| G | 51 | 🟥 | | 33 | 🟥 | 🟥 | 30 | ■ | | | | 27 | 🟥 | 🟥 | | 36 | | | | |
| H | | | | 26 | | | | | ■ | | 26 | 30 | | | | 22 | 22 | 🟥 | | |
| I | 39 | | 🟥 | | 🟥 | 🟥 | | | | ■ | 58 | | | 🟥 | | | | 🟥 | 🟥 | 46 |
| L | 21 | | | | | | | | | 23 | ■ | 23 | | 28 | 🟥 | | | | | 30 |
| K | 23 | 21 | | 28 | | | 31 | 23 | | | 21 | ■ | | | | 21 | | | | |
| M | 22 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | | 22 | 89 | | ■ | 22 | 🟥 | 🟥 | | | | 45 |
| F | | 🟥 | 🟥 | 🟥 | | 🟥 | 🟥 | 🟥 | 22 | 🟥 | 61 | | | ■ | 🟥 | | | | | |
| P | 50 | 🟥 | | 43 | 🟥 | | 57 | 43 | | | 🟥 | 21 | 🟥 | 🟥 | ■ | | | 🟥 | 🟥 | 🟥 |
| S | 49 | | | 24 | | | 24 | 36 | | | | 24 | 🟥 | | | ■ | 40 | 🟥 | 🟥 | |
| T | 32 | | | | | | 28 | 24 | | | | 24 | | | | 52 | ■ | 🟥 | 🟥 | |
| W | (40) | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | | | 🟥 | 🟥 | (40) | | | (60) | 🟥 | 🟥 | 🟥 | ■ | | 🟥 |
| Y | | 🟥 | | 🟥 | 🟥 | 🟥 | 🟥 | | (33) | 🟥 | | | | (50) | 🟥 | 🟥 | 🟥 | | ■ | |
| V | 36 | | | | | | | | | 21 | 43 | 21 | | | 🟥 | | | 🟥 | | ■ |

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.

Page 50

# Dayhoff’s 34 protein superfamilies

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

Page 50

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

| AAl | aRArg | NAsn | DAsp | CCys | QGln | EGlu | GGly | AR30N10917D1 |
|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

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
```

Page 48
Fig. 3.7

## 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 |

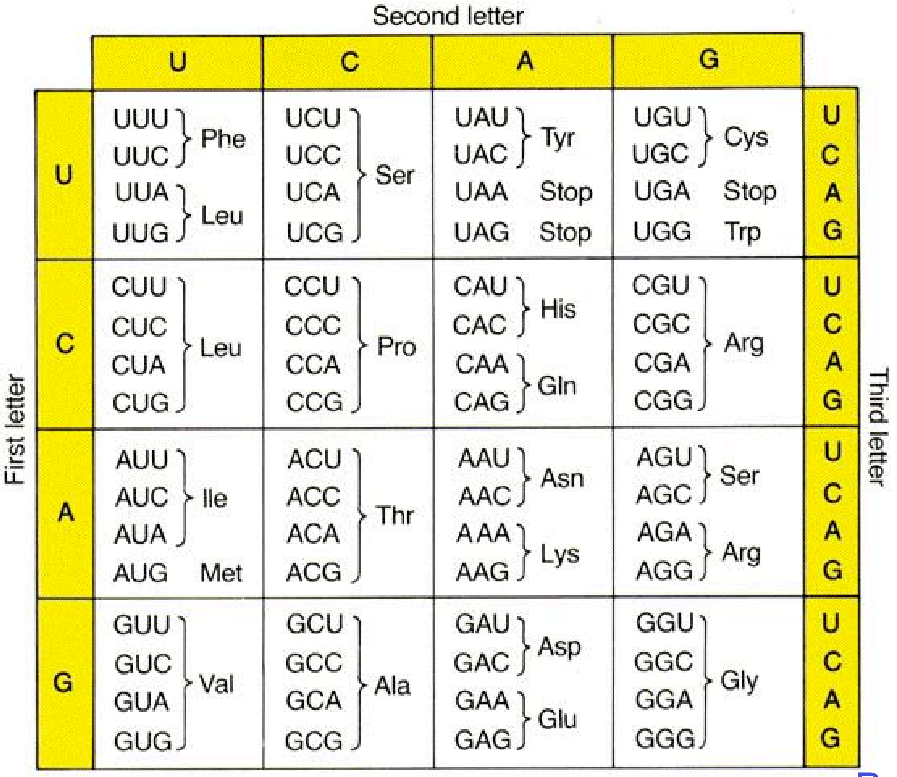
Table 3.1
Page 53

## 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



Second letter

| First letter | U | C | A | G | Third letter |
|---|---|---|---|---|---|
| U | UUU } Phe<br>UUC }<br>UUA } Leu<br>UUG } | UCU } Ser<br>UCC<br>UCA<br>UCG } | UAU } Tyr<br>UAC }<br>UAA Stop<br>UAG Stop | UGU } Cys<br>UGC }<br>UGA Stop<br>UGG Trp | U<br>C<br>A<br>G |
| C | CUU } Leu<br>CUC<br>CUA<br>CUG } | CCU } Pro<br>CCC<br>CCA<br>CCG } | CAU } His<br>CAC }<br>CAA } Gln<br>CAG } | CGU } Arg<br>CGC<br>CGA<br>CGG } | U<br>C<br>A<br>G |
| A | AUU } Ile<br>AUC<br>AUA }<br>AUG Met | ACU } Thr<br>ACC<br>ACA<br>ACG } | AAU } Asn<br>AAC }<br>AAA } Lys<br>AAG } | AGU } Ser<br>AGC }<br>AGA } Arg<br>AGG } | U<br>C<br>A<br>G |
| G | GUU } Val<br>GUC<br>GUA<br>GUG } | GCU } Ala<br>GCC<br>GCA<br>GCG } | GAU } Asp<br>GAC }<br>GAA } Glu<br>GAG } | GGU } Gly<br>GGC<br>GGA<br>GGG } | U<br>C<br>A<br>G |

Page 54

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

AAlaRArgNAsnDAspCCysQGlnEGluGGlyAR30N10917C

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |

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**

| AAl | aRArg | NAsn | DAspC | CysQG | lnEGl | uGGly | HHisI | IleA98 | 6729103 |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |

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.

# 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

| AAl | aRArg | NAsnD | AspC | CysQG | lnEGl | uGGly | HHisI | IleA 98 | 6729103 |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |

Page 55

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

| PAM0AAlaRArgNAsnDAspCCysQGlnEGluGGlyA100%0%0%0%0%0%0%0%R | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |

Top: original amino acid
Side: replacement amino acid

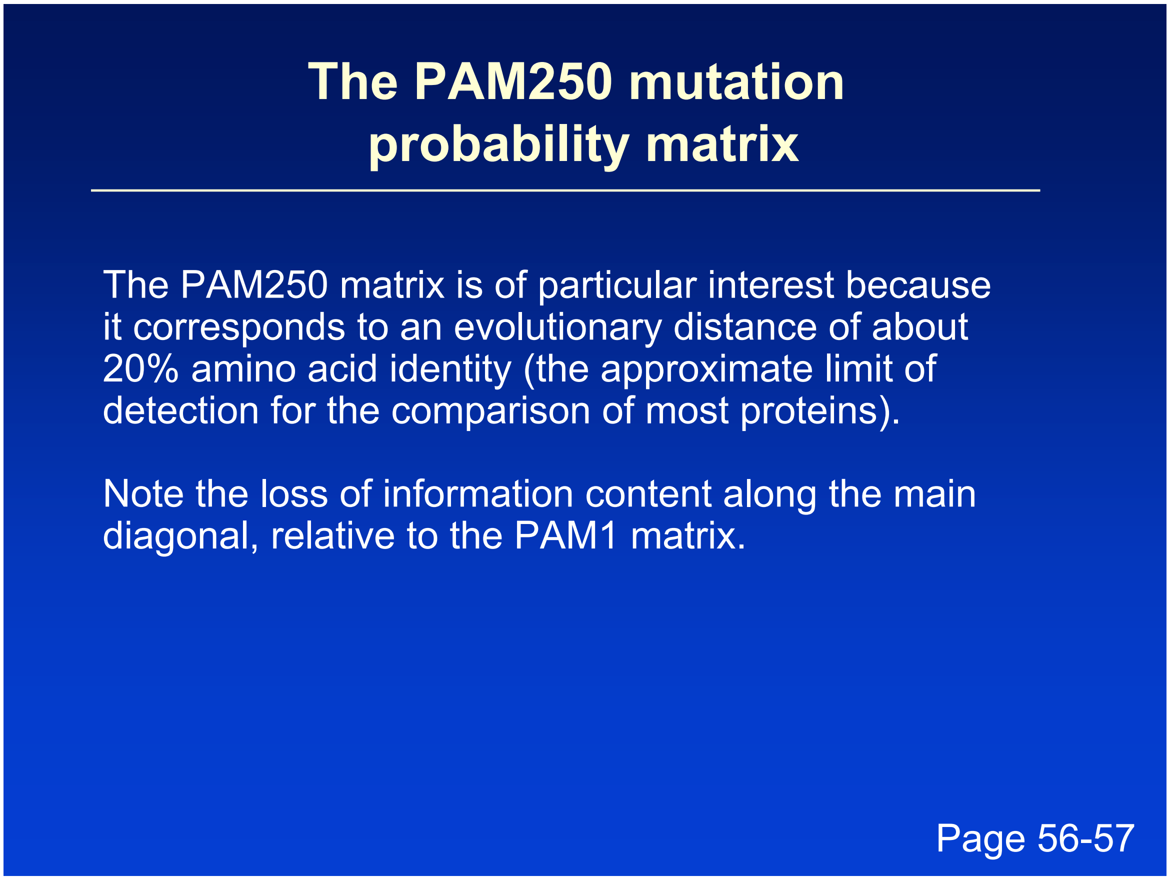
Fig. 3.12
Page 56

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

| PAM∞ | A<br>Ala | R<br>Arg | N<br>Asn | D<br>Asp | C<br>Cys | Q<br>Gln | E<br>Glu | G<br>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 |

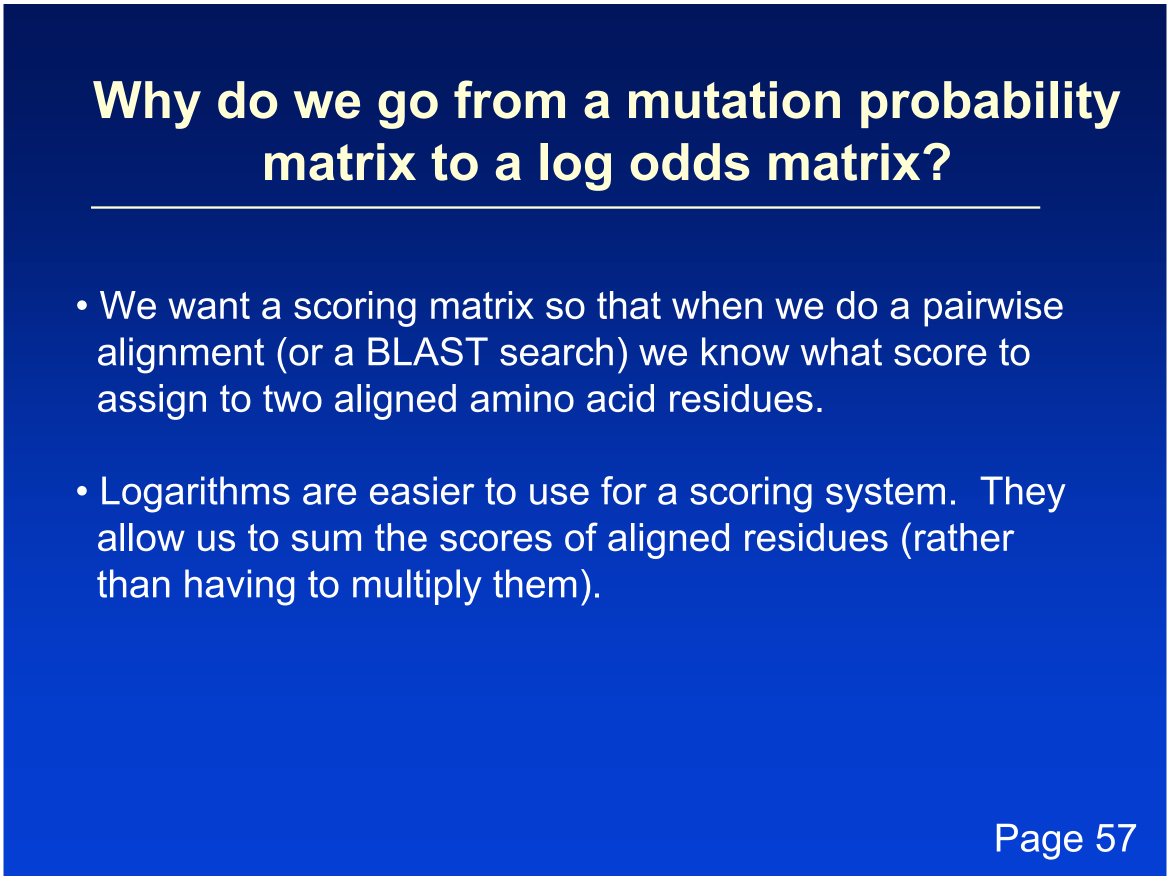
Top: original amino acid
Side: replacement amino acid

Fig. 3.13
Page 57

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **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** |

**PAM250 log odds scoring matrix**

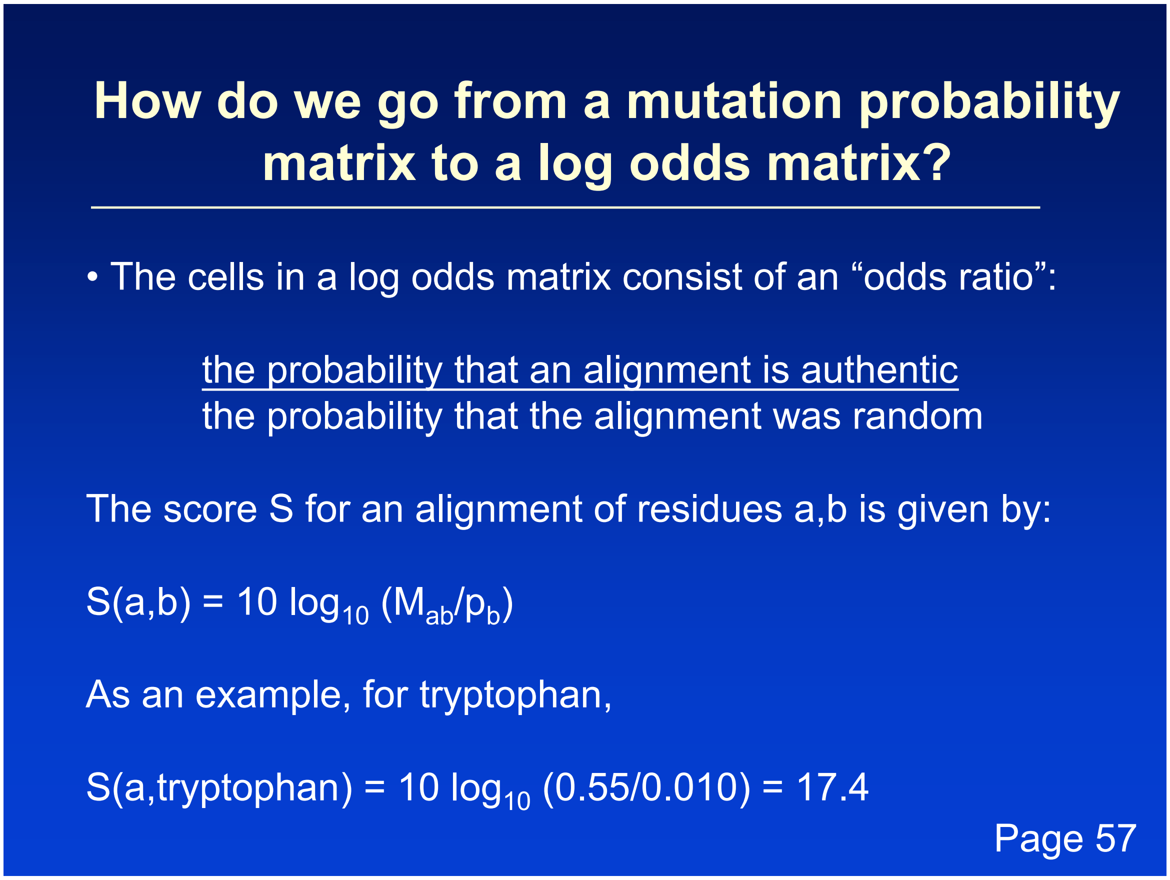
Fig. 3.14
Page 58



## 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).



## 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”:

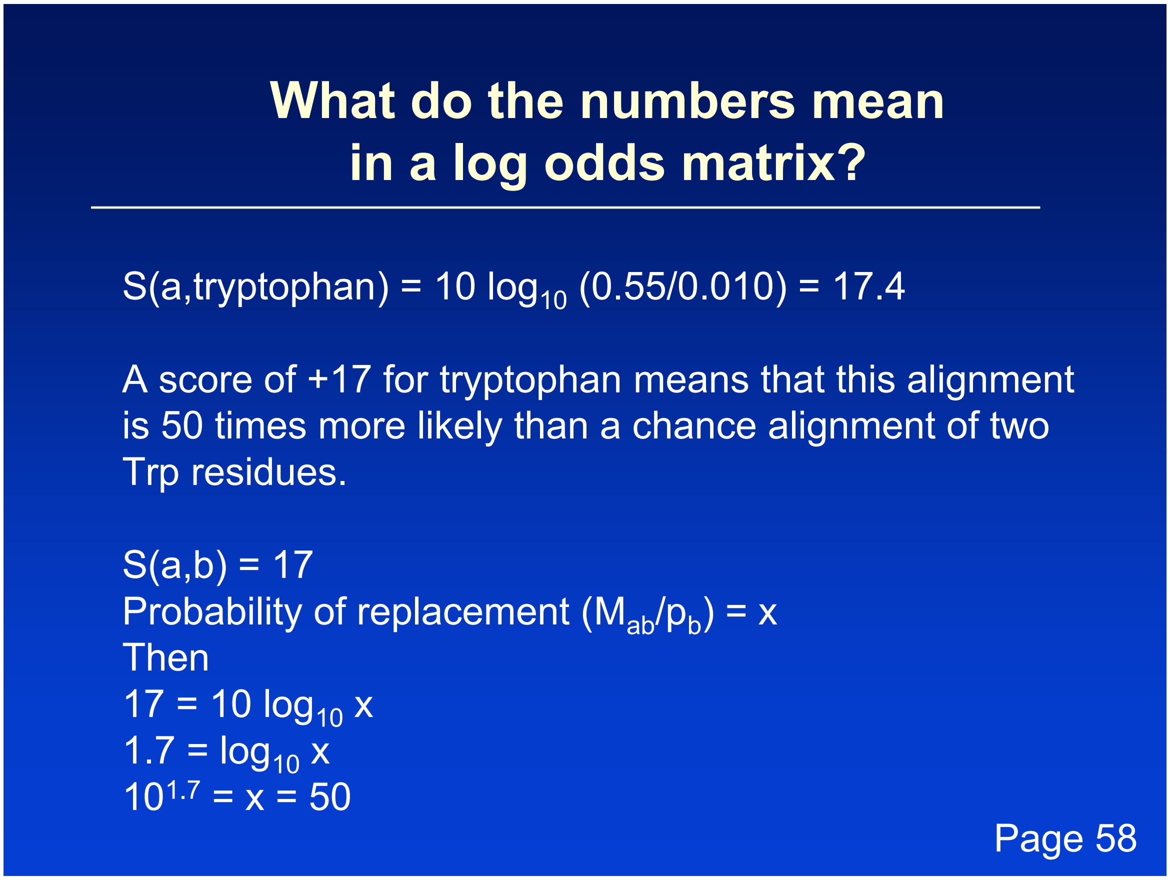
$$\frac{\text{the probability that an alignment is authentic}}{\text{the probability that the alignment was random}}$$

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,

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

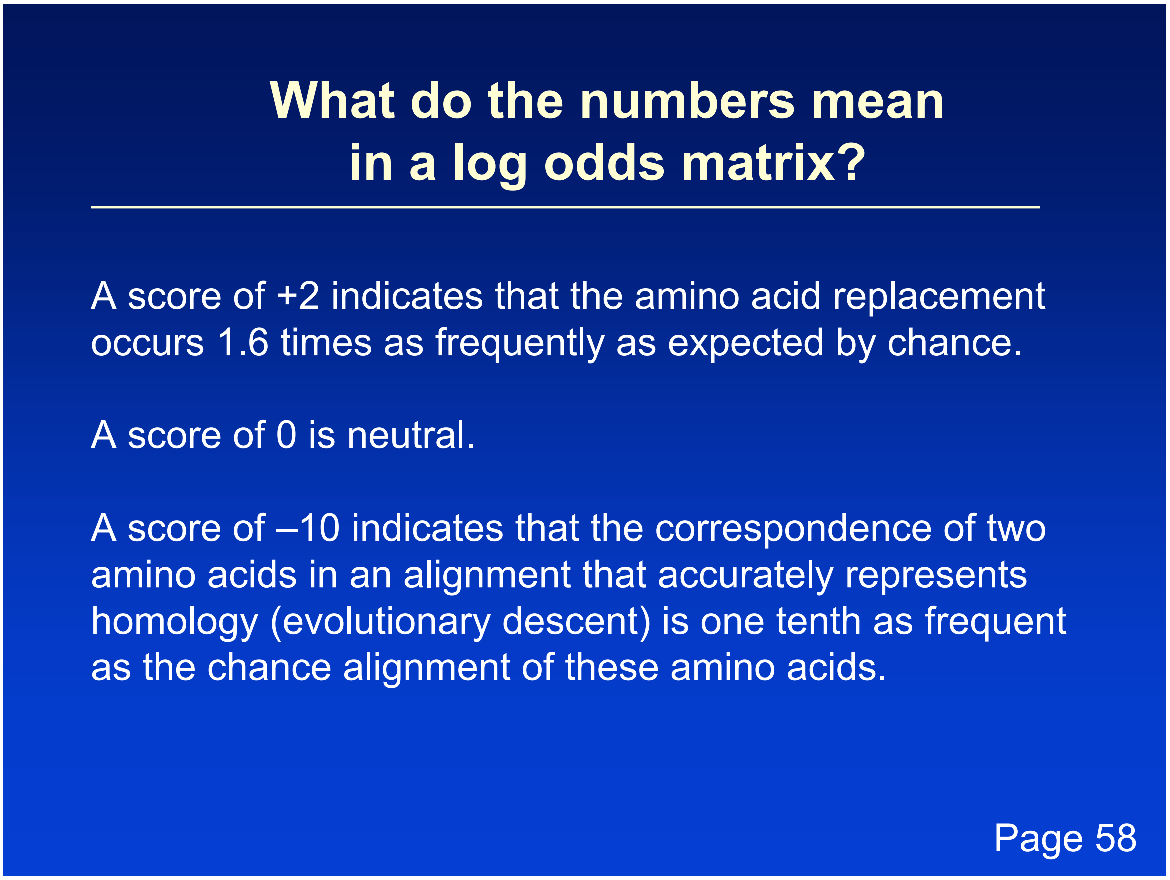


## What do the numbers mean in a log odds matrix?

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

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

$S(a,b) = 17$<br>
Probability of replacement $(M_{ab}/p_b) = x$<br>
Then<br>
$17 = 10 \log_{10} x$<br>
$1.7 = \log_{10} x$<br>
$10^{1.7} = x = 50$



## 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.

Page 58

**PAM250 log odds scoring matrix**

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **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

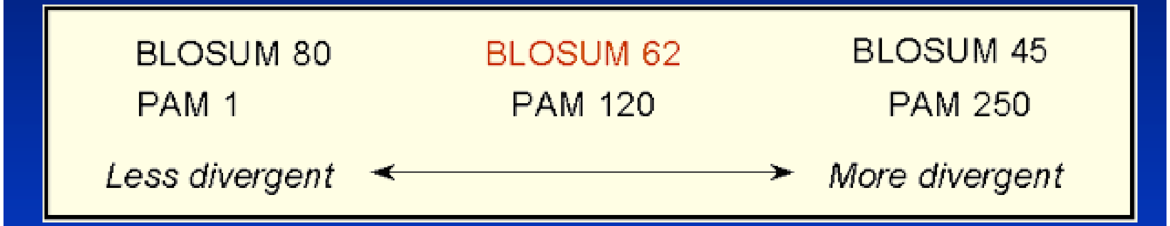
## PAM10 log odds scoring matrix

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

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

Fig. 3.15
Page 59

| | | | |
|---|---|---|---|
| BLOSUM90 | BLOSUM80 | BLOSUM62 | BLOSUM45 |
| PAM30 | PAM120 | PAM180 | PAM240 |



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
             **        *  ** **
```

## 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

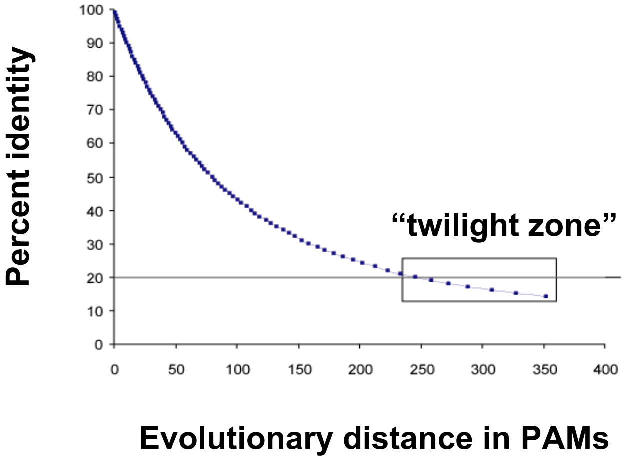


Fig. 3.19
Page 62

## 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

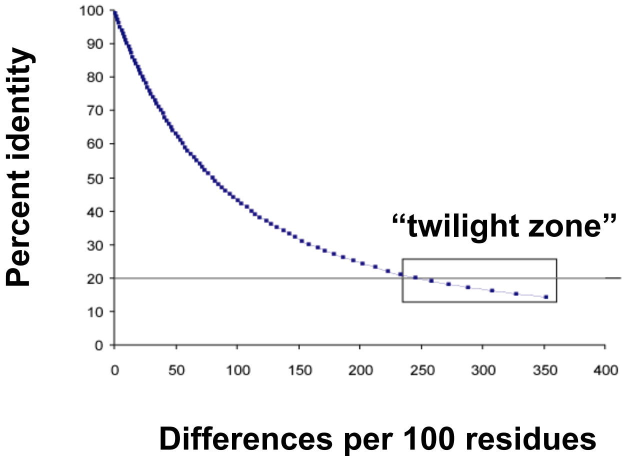
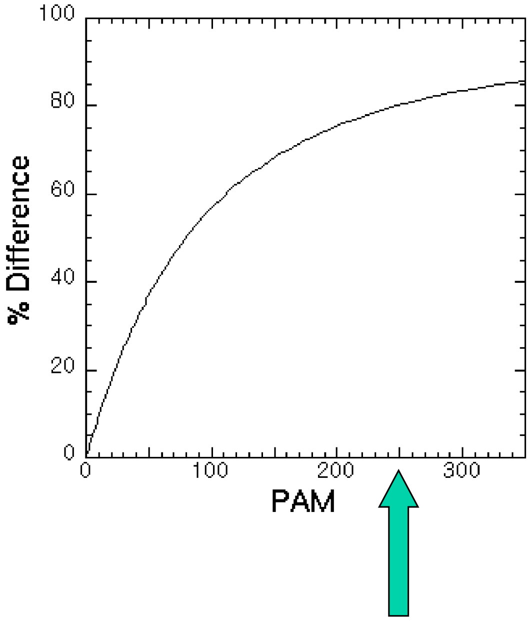


Fig. 3.19
Page 62

## PAM matrices reflect different degrees of divergence



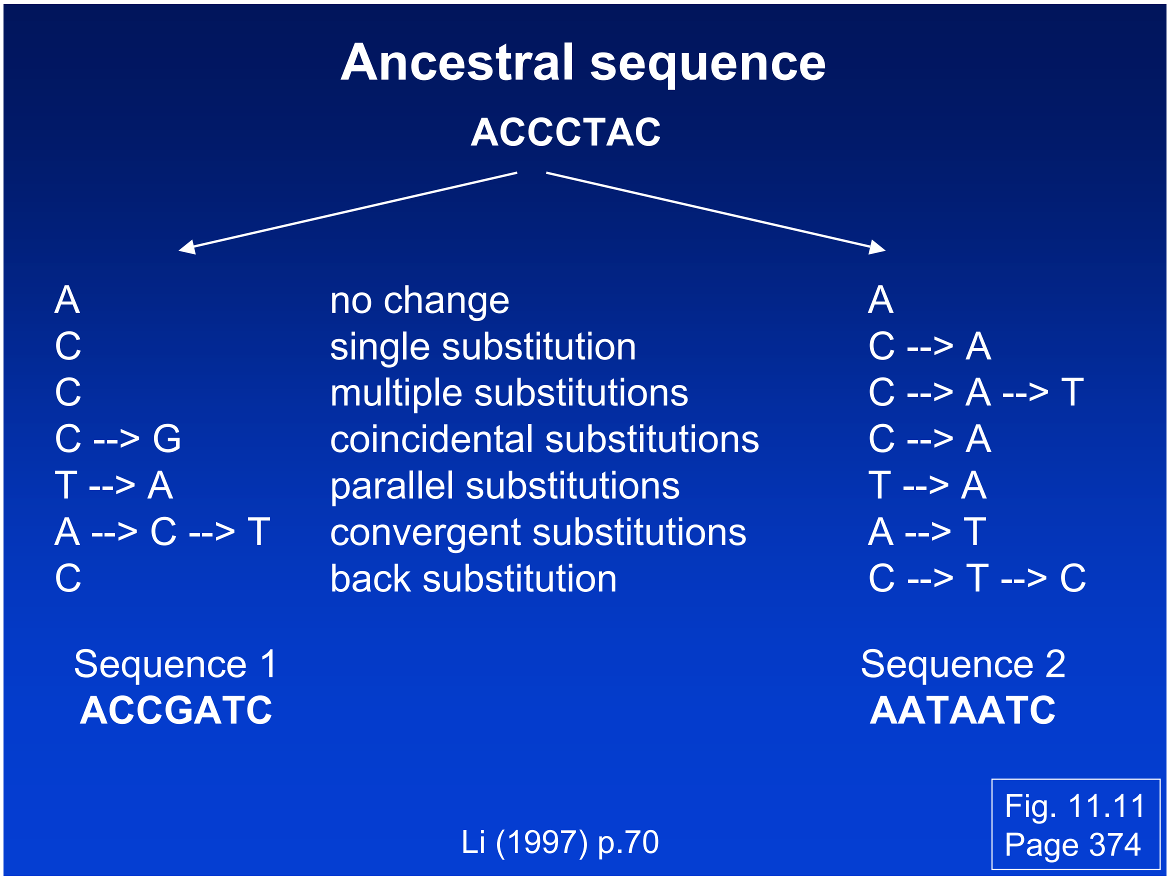
**PAM250**

## 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)



## 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

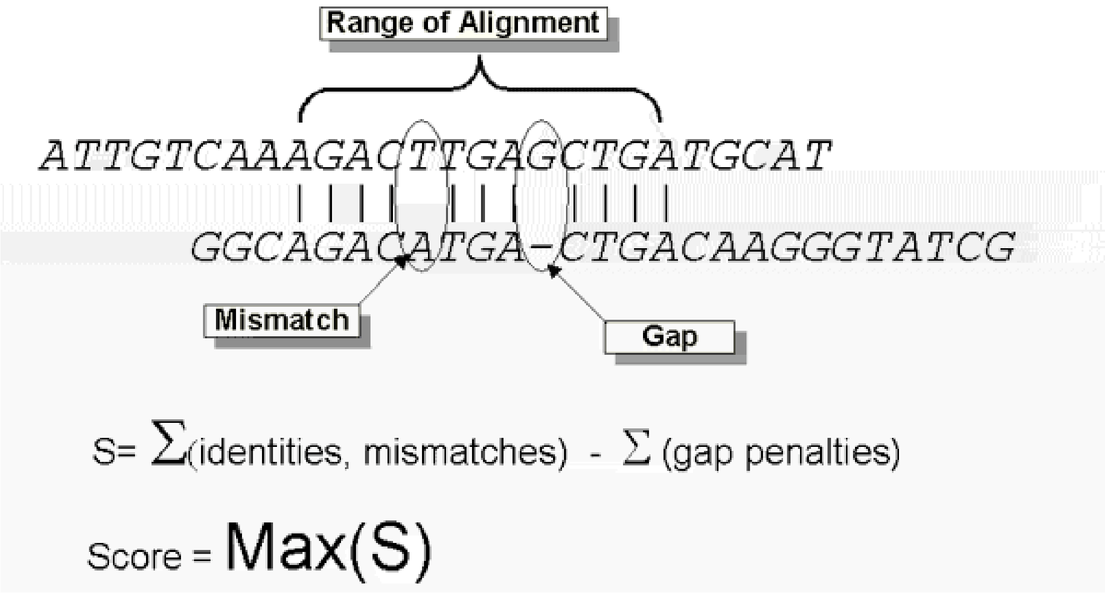
Page 50

## 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

Page 62

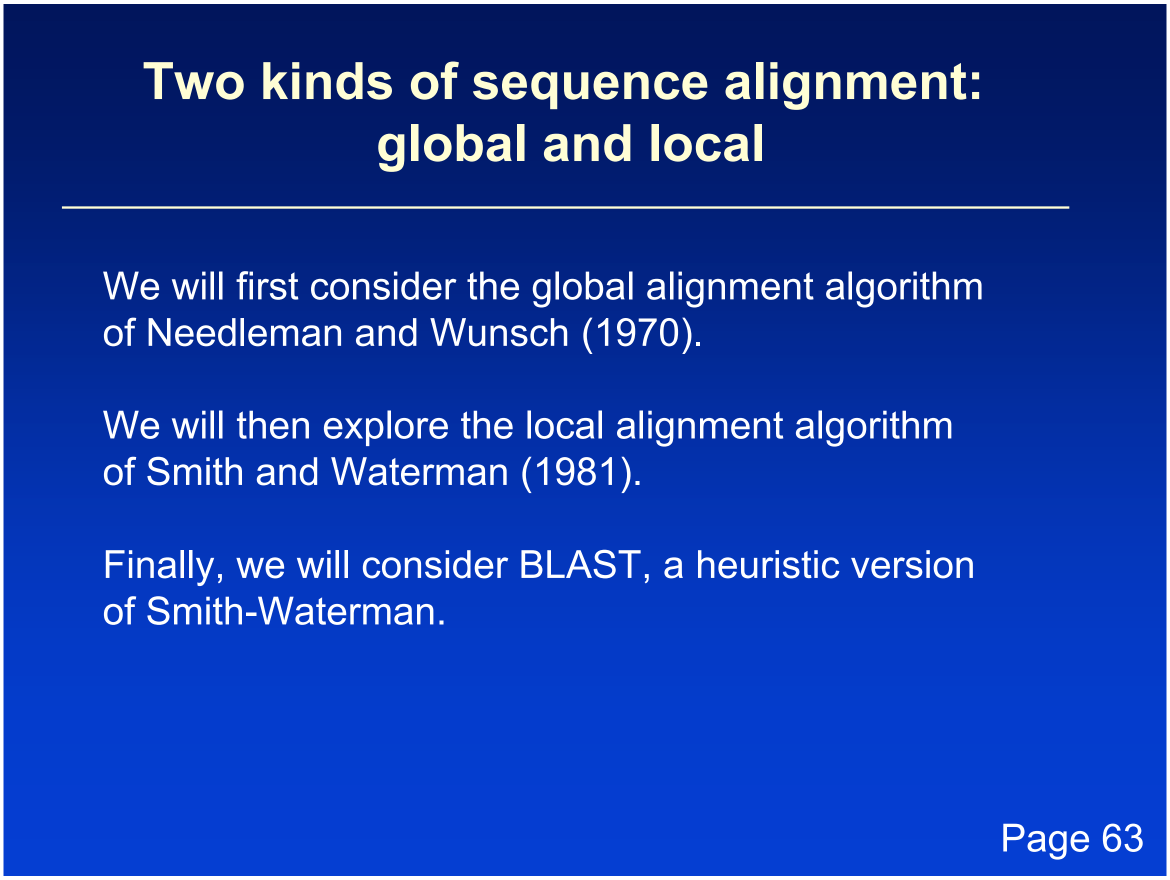
## Calculation of an alignment score



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

$\text{Score} = \text{Max}(S)$

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

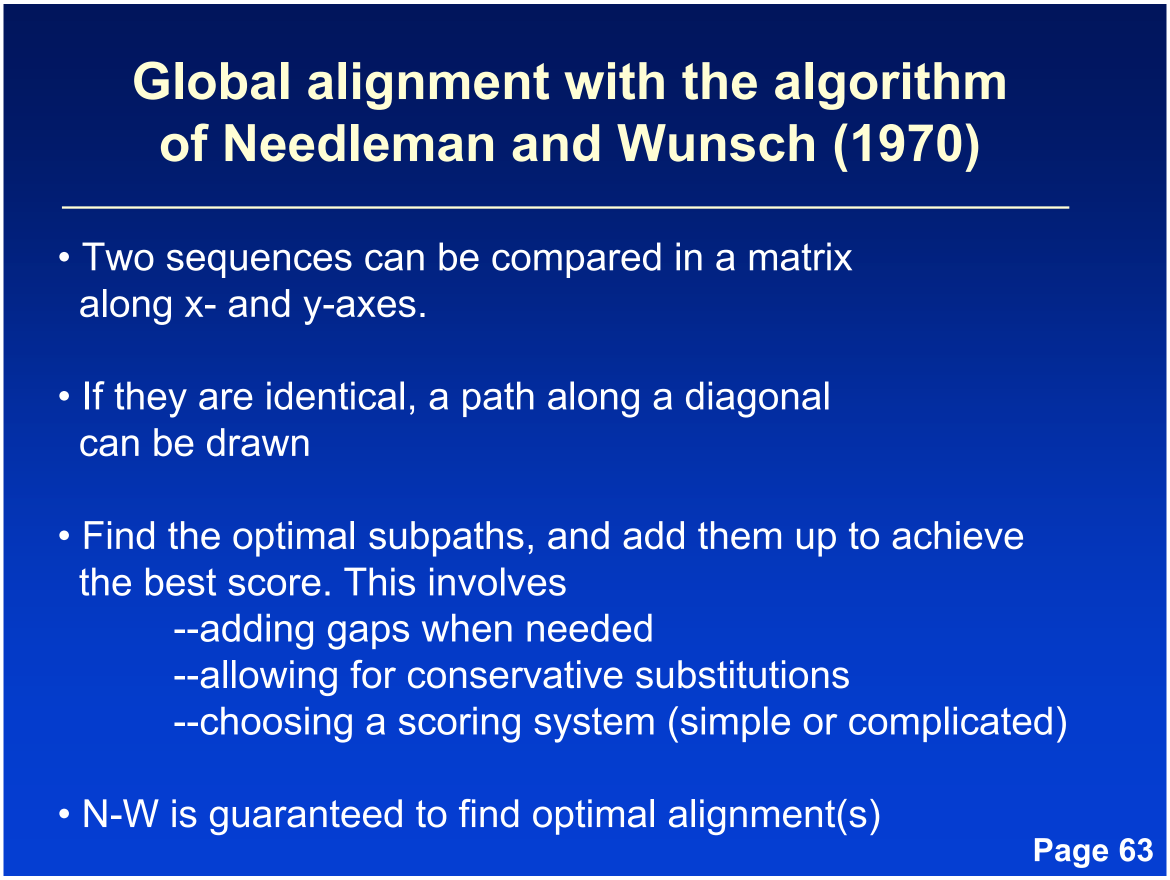


# Two kinds of sequence alignment: global and local

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

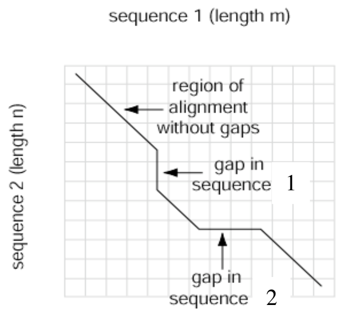
[1] set up a matrix

[2] score the matrix

[3] identify the optimal alignment(s)

Page 63

# 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

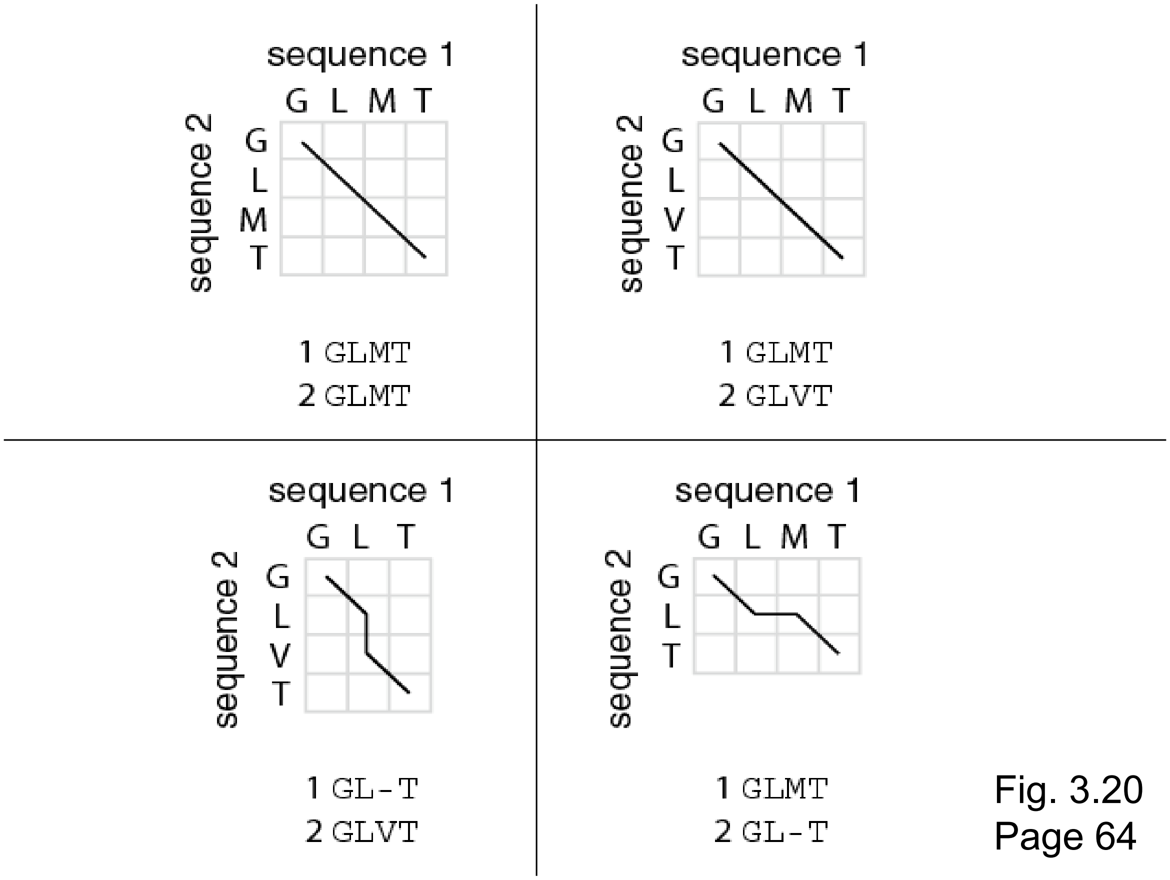


Fig. 3.20
Page 64

# Start Needleman-Wunsch with an identity matrix

| | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | | | | | | | | | | | | |
| J | | | | | 1 | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| J | | | | | 1 | | | | | | | | |
| N | | | | 1 | | | | | | | | | |
| R | | | | | | 1 | | | | | 1 | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| K | | | | | | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| R | | | | | | 1 | | | | | 1 | | |
| B | | 1 | | | | | | | | | | | |
| P | | | | | | | | | | | | 1 | |

Fig. 3.21
Page 65

# Start Needleman-Wunsch with an identity matrix

| | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | | | | | | | | | | | | |
| J | | | | | 1 | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| J | | | | | 1 | | | | | | | | |
| N | | | | 1 | | | | | | | | | |
| R | | | | | | 1 | | | | | 1 | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| K | | | | | | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| R | | | | | | 1 | | | | | 1 | | |
| B | | 1 | | | | | | | | | | | |
| P | | | | | | | | | | | | 1 | |

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

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

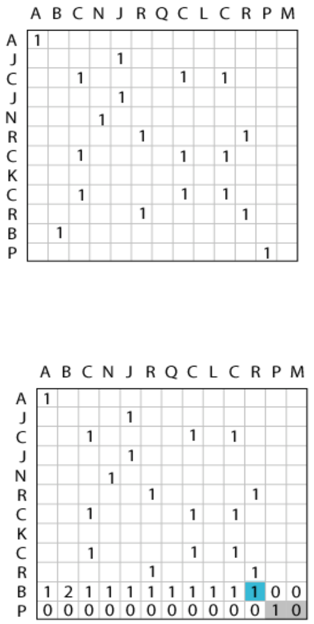
Fig. 3.21
Page 65

# Fill in the matrix starting from the bottom right

| | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | | | | | | | | | | | | |
| J | | | | | 1 | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| J | | | | | 1 | | | | | | | | |
| N | | | | 1 | | | | | | | | | |
| R | | | | | | 1 | | | | | 1 | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| K | | | | | | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| R | | | | | | 1 | | | | | 1 | | |
| B | | 1 | | | | | | | | | | | |
| P | | | | | | | | | | | | 1 | |

| | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | | | | | | | | | | | | |
| J | | | | | 1 | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| J | | | | | 1 | | | | | | | | |
| N | | | | 1 | | | | | | | | | |
| R | | | | | | 1 | | | | | 1 | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| K | | | | | | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| R | | | | | | 1 | | | | | 1 | | |
| B | | 1 | | | | | | | | | | | |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

Fig. 3.21
Page 65

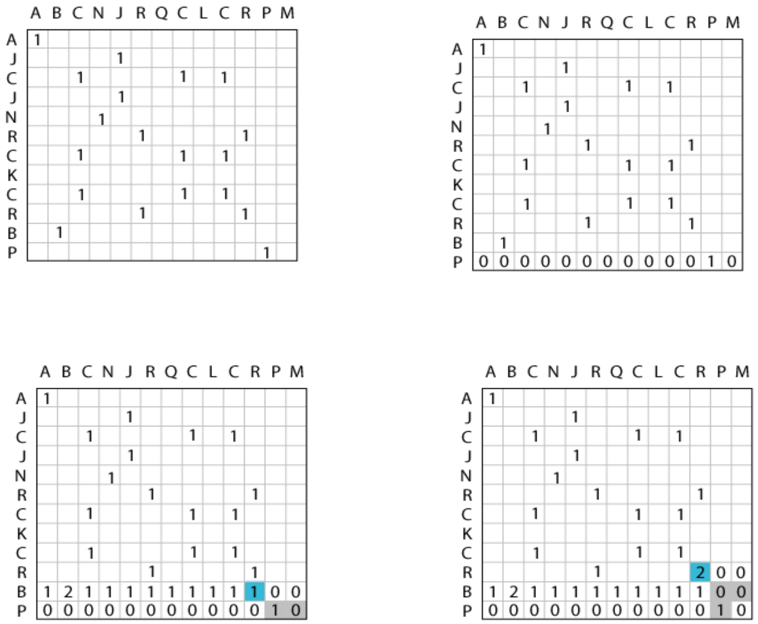


| | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | | | | | | | | | | | | |
| J | | | | | 1 | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| J | | | | | 1 | | | | | | | | |
| N | | | | 1 | | | | | | | | | |
| R | | | | | | 1 | | | | | 1 | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| K | | | | | | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| R | | | | | | 1 | | | | | 1 | | |
| B | | 1 | | | | | | | | | | | |
| P | | | | | | | | | | | | 1 | |

| | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | | | | | | | | | | | | |
| J | | | | | 1 | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| J | | | | | 1 | | | | | | | | |
| N | | | | 1 | | | | | | | | | |
| R | | | | | | 1 | | | | | 1 | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| K | | | | | | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| R | | | | | | 1 | | | | | 1 | | |
| B | | 1 | | | | | | | | | | | |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

| | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | | | | | | | | | | | | |
| J | | | | | 1 | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| J | | | | | 1 | | | | | | | | |
| N | | | | 1 | | | | | | | | | |
| R | | | | | | 1 | | | | | 1 | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| K | | | | | | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| R | | | | | | 1 | | | | | 1 | | |
| B | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

Fig. 3.21
Page 65



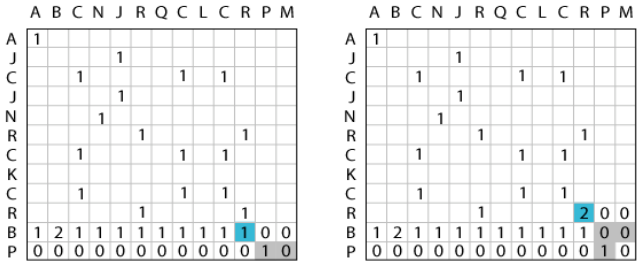
| | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | | | | | | | | | | | | |
| J | | | | | 1 | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| J | | | | | 1 | | | | | | | | |
| N | | | | 1 | | | | | | | | | |
| R | | | | | | 1 | | | | | 1 | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| K | | | | | | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| R | | | | | | 1 | | | | | 1 | | |
| B | | 1 | | | | | | | | | | | |
| P | | | | | | | | | | | | 1 | |

| | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | | | | | | | | | | | | |
| J | | | | | 1 | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| J | | | | | 1 | | | | | | | | |
| N | | | | 1 | | | | | | | | | |
| R | | | | | | 1 | | | | | 1 | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| K | | | | | | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| R | | | | | | 1 | | | | | 1 | | |
| B | | 1 | | | | | | | | | | | |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

| | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | | | | | | | | | | | | |
| J | | | | | 1 | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| J | | | | | 1 | | | | | | | | |
| N | | | | 1 | | | | | | | | | |
| R | | | | | | 1 | | | | | 1 | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| K | | | | | | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| R | | | | | | 1 | | | | | 1 | | |
| B | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

| | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | | | | | | | | | | | | |
| J | | | | | 1 | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| J | | | | | 1 | | | | | | | | |
| N | | | | 1 | | | | | | | | | |
| R | | | | | | 1 | | | | | 1 | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| K | | | | | | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| R | | | | | | 1 | | | | | 2 | 0 | 0 |
| B | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

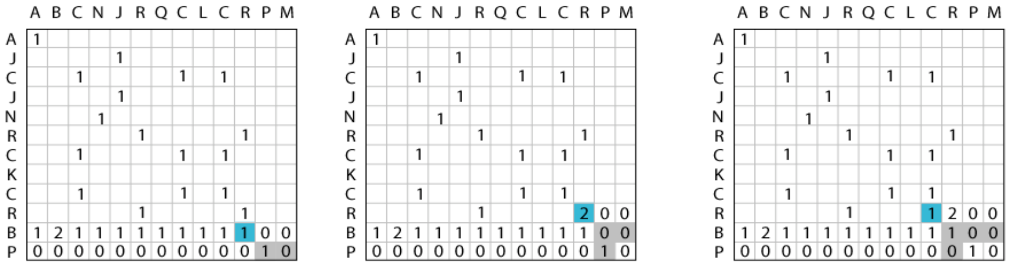
Fig. 3.21
Page 65



|   | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 |  |  |  |  |  |  |  |  |  |  |  |  |
| J |  |  |  |  | 1 |  |  |  |  |  |  |  |  |
| C |  |  | 1 |  |  |  |  | 1 |  | 1 |  |  |  |
| J |  |  |  |  | 1 |  |  |  |  |  |  |  |  |
| N |  |  |  | 1 |  |  |  |  |  |  |  |  |  |
| R |  |  |  |  |  | 1 |  |  |  |  | 1 |  |  |
| C |  |  | 1 |  |  |  |  | 1 |  | 1 |  |  |  |
| K |  |  |  |  |  |  |  |  |  |  |  |  |  |
| C |  |  | 1 |  |  |  |  | 1 |  | 1 |  |  |  |
| R |  |  |  |  |  | 1 |  |  |  |  | 1 |  |  |
| B | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

|   | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 |  |  |  |  |  |  |  |  |  |  |  |  |
| J |  |  |  |  | 1 |  |  |  |  |  |  |  |  |
| C |  |  | 1 |  |  |  |  | 1 |  | 1 |  |  |  |
| J |  |  |  |  | 1 |  |  |  |  |  |  |  |  |
| N |  |  |  | 1 |  |  |  |  |  |  |  |  |  |
| R |  |  |  |  |  | 1 |  |  |  |  | 1 |  |  |
| C |  |  | 1 |  |  |  |  | 1 |  | 1 |  |  |  |
| K |  |  |  |  |  |  |  |  |  |  |  |  |  |
| C |  |  | 1 |  |  |  |  | 1 |  | 1 |  |  |  |
| R |  |  |  |  |  | 1 |  |  |  |  | 2 | 0 | 0 |
| B | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

Fig. 3.21
Page 65

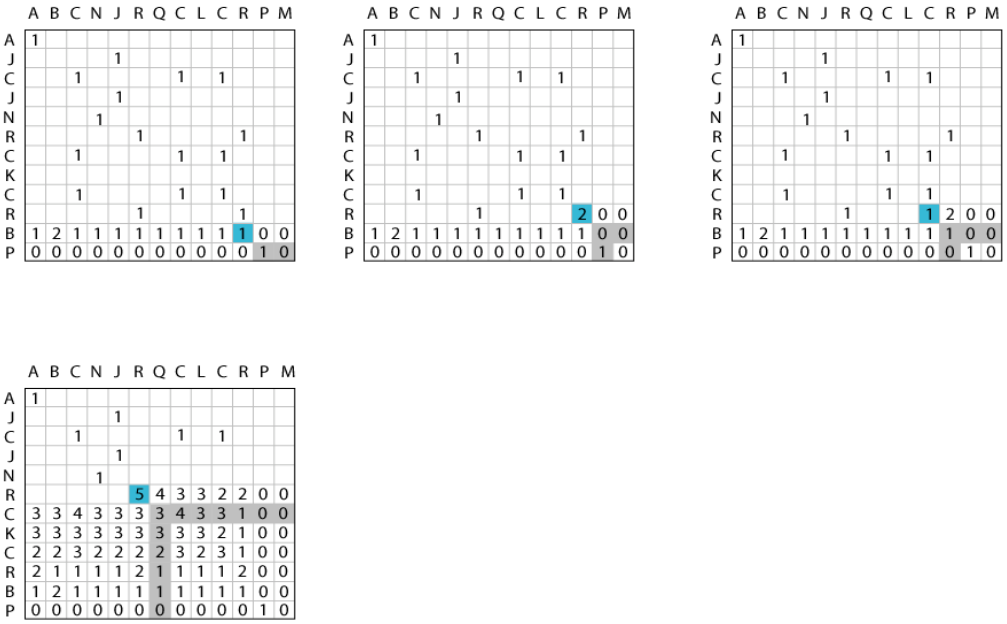


|   | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 |  |  |  |  |  |  |  |  |  |  |  |  |
| J |  |  |  |  | 1 |  |  |  |  |  |  |  |  |
| C |  |  | 1 |  |  |  |  | 1 |  | 1 |  |  |  |
| J |  |  |  |  | 1 |  |  |  |  |  |  |  |  |
| N |  |  |  | 1 |  |  |  |  |  |  |  |  |  |
| R |  |  |  |  |  | 1 |  |  |  |  | 1 |  |  |
| C |  |  | 1 |  |  |  |  | 1 |  | 1 |  |  |  |
| K |  |  |  |  |  |  |  |  |  |  |  |  |  |
| C |  |  | 1 |  |  |  |  | 1 |  | 1 |  |  |  |
| R |  |  |  |  |  | 1 |  |  |  |  | 1 |  |  |
| B | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

|   | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 |  |  |  |  |  |  |  |  |  |  |  |  |
| J |  |  |  |  | 1 |  |  |  |  |  |  |  |  |
| C |  |  | 1 |  |  |  |  | 1 |  | 1 |  |  |  |
| J |  |  |  |  | 1 |  |  |  |  |  |  |  |  |
| N |  |  |  | 1 |  |  |  |  |  |  |  |  |  |
| R |  |  |  |  |  | 1 |  |  |  |  | 1 |  |  |
| C |  |  | 1 |  |  |  |  | 1 |  | 1 |  |  |  |
| K |  |  |  |  |  |  |  |  |  |  |  |  |  |
| C |  |  | 1 |  |  |  |  | 1 |  | 1 |  |  |  |
| R |  |  |  |  |  | 1 |  |  |  |  | 2 | 0 | 0 |
| B | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

|   | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 |  |  |  |  |  |  |  |  |  |  |  |  |
| J |  |  |  |  | 1 |  |  |  |  |  |  |  |  |
| C |  |  | 1 |  |  |  |  | 1 |  | 1 |  |  |  |
| J |  |  |  |  | 1 |  |  |  |  |  |  |  |  |
| N |  |  |  | 1 |  |  |  |  |  |  |  |  |  |
| R |  |  |  |  |  | 1 |  |  |  |  | 1 |  |  |
| C |  |  | 1 |  |  |  |  | 1 |  | 1 |  |  |  |
| K |  |  |  |  |  |  |  |  |  |  |  |  |  |
| C |  |  | 1 |  |  |  |  | 1 |  | 1 |  |  |  |
| R |  |  |  |  |  | 1 |  |  |  | 1 | 2 | 0 | 0 |
| B | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

Fig. 3.22
Page 66



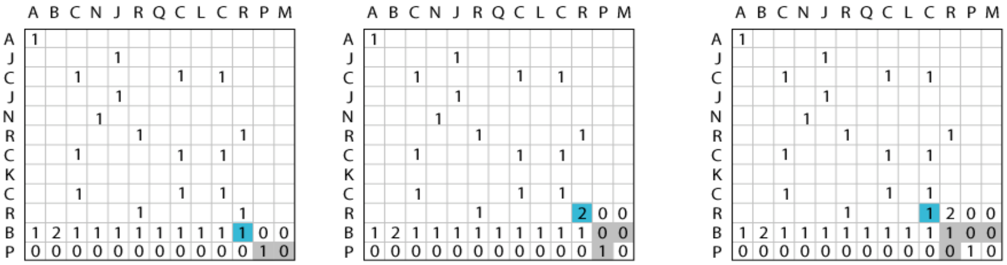
| | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | | | | | | | | | | | | |
| J | | | | | 1 | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| J | | | | | 1 | | | | | | | | |
| N | | | | 1 | | | | | | | | | |
| R | | | | | | 1 | | | | | 1 | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| K | | | | | | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| R | | | | | | 1 | | | | | 1 | | |
| B | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

| | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | | | | | | | | | | | | |
| J | | | | | 1 | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| J | | | | | 1 | | | | | | | | |
| N | | | | 1 | | | | | | | | | |
| R | | | | | | 1 | | | | | 1 | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| K | | | | | | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| R | | | | | | 1 | | | | | 2 | 0 | 0 |
| B | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

| | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | | | | | | | | | | | | |
| J | | | | | 1 | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| J | | | | | 1 | | | | | | | | |
| N | | | | 1 | | | | | | | | | |
| R | | | | | | 1 | | | | | 1 | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| K | | | | | | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| R | | | | | | 1 | | | | 1 | 2 | 0 | 0 |
| B | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

| | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | | | | | | | | | | | | |
| J | | | | | 1 | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| J | | | | | 1 | | | | | | | | |
| N | | | | 1 | | | | | | | | | |
| R | | | | | | 5 | 4 | 3 | 3 | 2 | 2 | 0 | 0 |
| C | 3 | 3 | 4 | 3 | 3 | 3 | 3 | 4 | 3 | 3 | 1 | 0 | 0 |
| K | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 1 | 0 | 0 |
| C | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 3 | 2 | 3 | 1 | 0 | 0 |
| R | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 0 | 0 |
| B | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

Fig. 3.22
Page 66

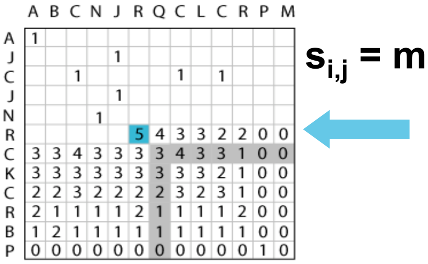


| | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | | | | | | | | | | | | |
| J | | | | | 1 | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| J | | | | | 1 | | | | | | | | |
| N | | | | 1 | | | | | | | | | |
| R | | | | | | 1 | | | | | 1 | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| K | | | | | | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| R | | | | | | 1 | | | | | 1 | | |
| B | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

| | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | | | | | | | | | | | | |
| J | | | | | 1 | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| J | | | | | 1 | | | | | | | | |
| N | | | | 1 | | | | | | | | | |
| R | | | | | | 1 | | | | | 1 | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| K | | | | | | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| R | | | | | | 1 | | | | | 2 | 0 | 0 |
| B | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

| | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | | | | | | | | | | | | |
| J | | | | | 1 | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| J | | | | | 1 | | | | | | | | |
| N | | | | 1 | | | | | | | | | |
| R | | | | | | 1 | | | | | 1 | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| K | | | | | | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| R | | | | | | 1 | | | | 1 | 2 | 0 | 0 |
| B | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

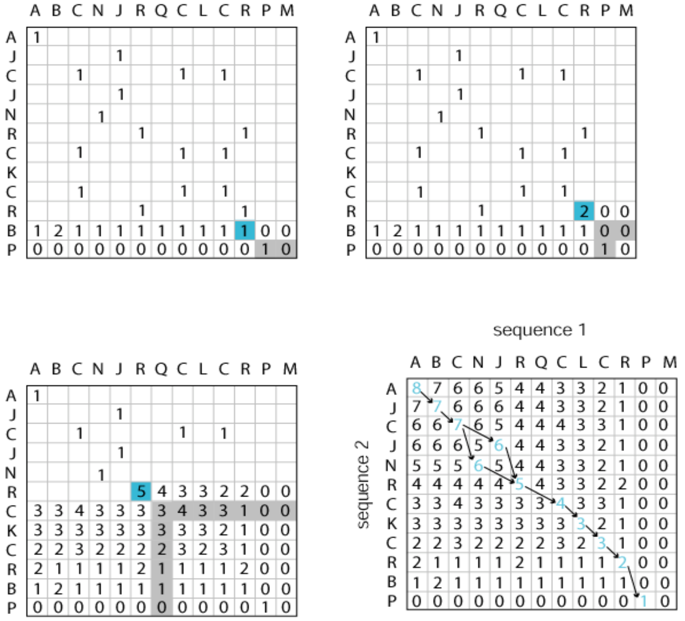
**Rule for assigning score in position i, j:**



| | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | | | | | | | | | | | | |
| J | | | | | 1 | | | | | | | | |
| C | | | 1 | | | | | 1 | | 1 | | | |
| J | | | | | 1 | | | | | | | | |
| N | | | | 1 | | | | | | | | | |
| R | | | | | | 5 | 4 | 3 | 3 | 2 | 2 | 0 | 0 |
| C | 3 | 3 | 4 | 3 | 3 | 3 | 3 | 4 | 3 | 3 | 1 | 0 | 0 |
| K | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 1 | 0 | 0 |
| C | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 3 | 2 | 3 | 1 | 0 | 0 |
| R | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 0 | 0 |
| B | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

$$s_{i,j} = \max \left[ \begin{array}{l} s_{i-1,j-1} + s(a_i b_j) \\ s_{i-x,j} \text{ (i.e. add a gap of length x)} \\ s_{i,j-x} \text{ (i.e. add a gap of length x)} \end{array} \right.$$

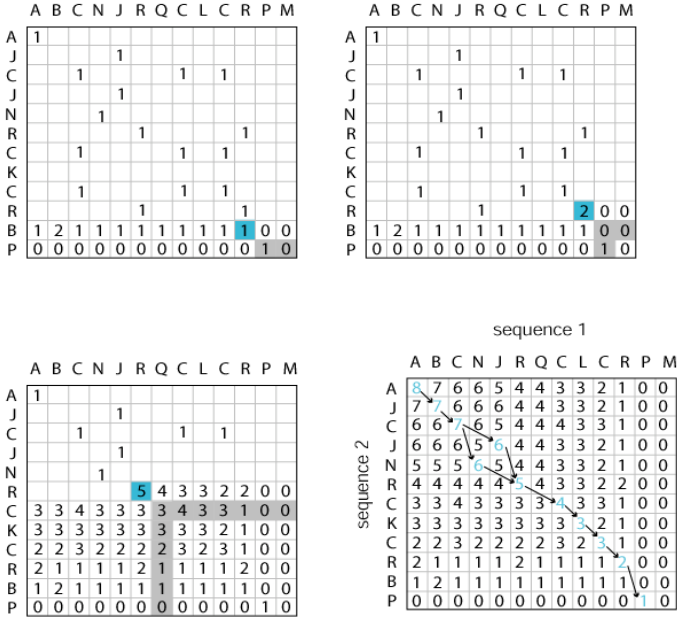
Fig. 3.22
Page 66



|  | A | B | C | N | J | R | Q | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 8 | 7 | 6 | 6 | 5 | 4 | 4 | 3 | 3 | 2 | 1 | 0 | 0 |
| J | 7 | 7 | 6 | 6 | 6 | 4 | 4 | 3 | 3 | 2 | 1 | 0 | 0 |
| C | 6 | 6 | 7 | 6 | 5 | 4 | 4 | 4 | 3 | 3 | 1 | 0 | 0 |
| J | 6 | 6 | 6 | 5 | 6 | 4 | 4 | 3 | 3 | 2 | 1 | 0 | 0 |
| N | 5 | 5 | 5 | 6 | 5 | 4 | 4 | 3 | 3 | 2 | 1 | 0 | 0 |
| R | 4 | 4 | 4 | 4 | 4 | 5 | 4 | 3 | 3 | 2 | 2 | 0 | 0 |
| C | 3 | 3 | 4 | 3 | 3 | 3 | 3 | 4 | 3 | 3 | 1 | 0 | 0 |
| K | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 1 | 0 | 0 |
| C | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 3 | 2 | 3 | 1 | 0 | 0 |
| R | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 0 | 0 |
| B | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

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

Page 66



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

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

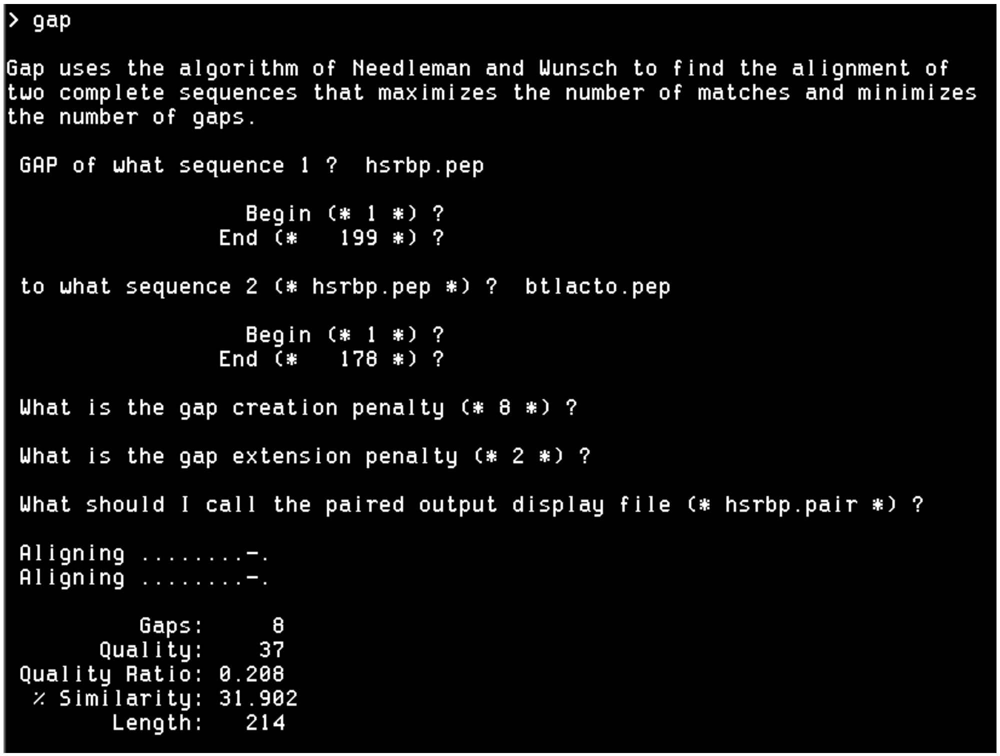
Fig. 3.22
Page 66

# Needleman-Wunsch: dynamic programming

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

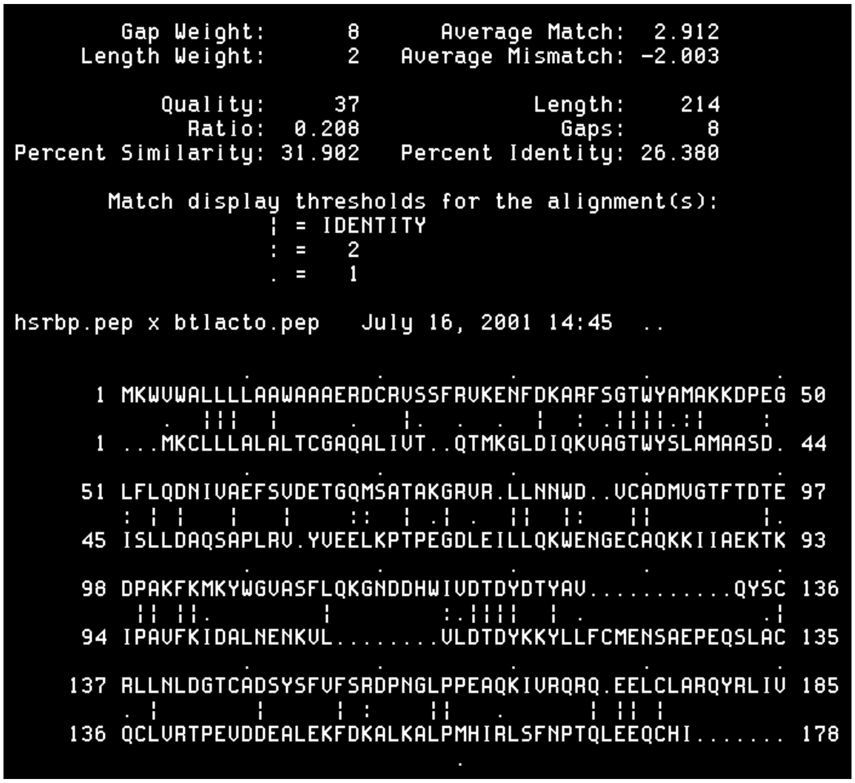
It is an example of a dynamic programming algorithm: an optimal path (alignment) is identified by incrementally extending optimal subpaths.
Thus, a series of decisions is made at each step of the alignment to find the pair of residues with the best score.

Page 67



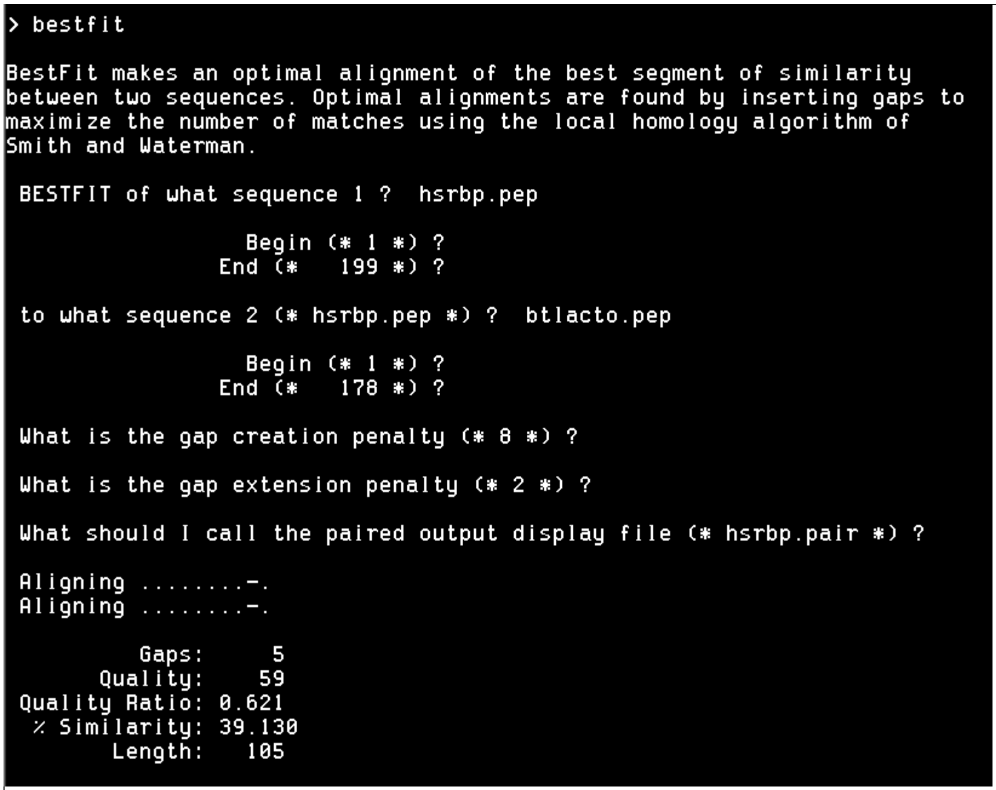
```
> gap

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
Page 68



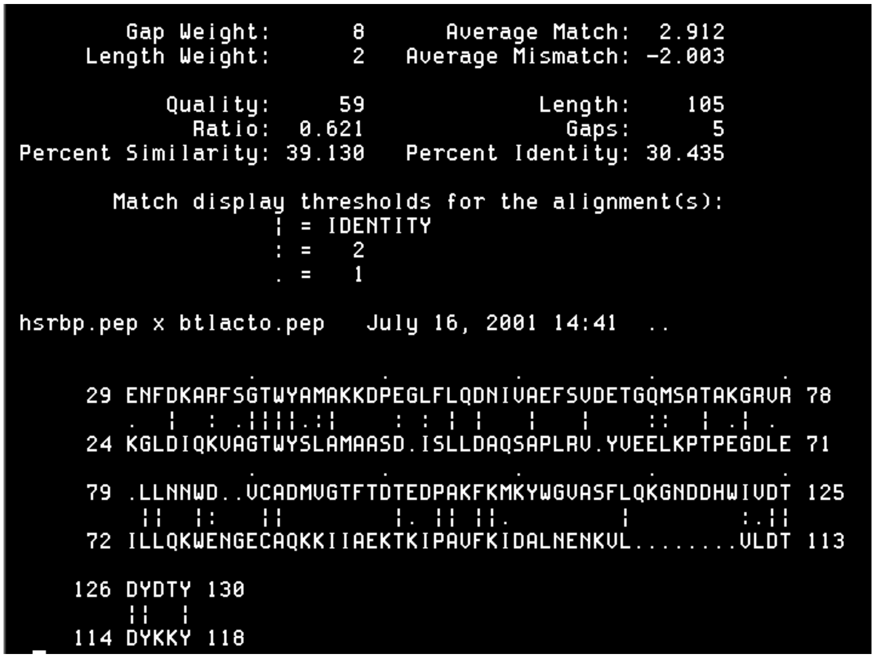
```
        Gap Weight:      8      Average Match:  2.912
     Length Weight:      2   Average Mismatch: -2.003

           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
Page 69



```
> 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
Page 71



```
        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)

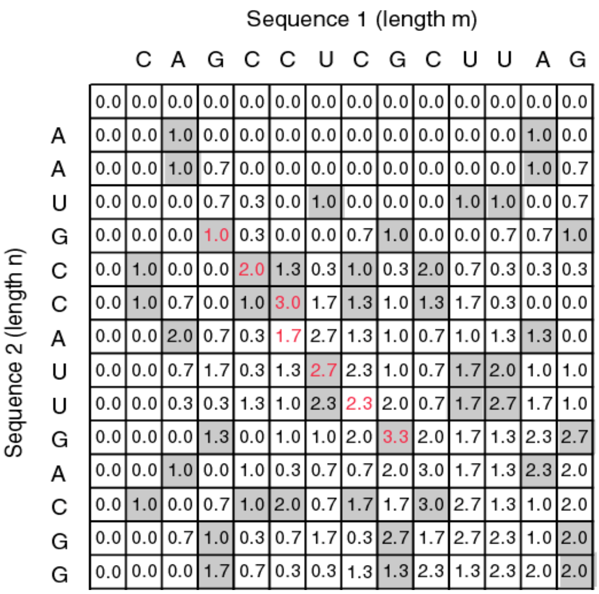
No values in the scoring matrix can be negative! S $\geq$ 0

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



Sequence 1 (length m)

Sequence 2 (length n)

| | | C | A | G | C | C | U | C | G | C | U | U | A | G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| A | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| A | 0.0 | 0.0 | 1.0 | 0.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.7 |
| U | 0.0 | 0.0 | 0.0 | 0.7 | 0.3 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.7 |
| G | 0.0 | 0.0 | 0.0 | 1.0 | 0.3 | 0.0 | 0.0 | 0.7 | 1.0 | 0.0 | 0.0 | 0.7 | 0.7 | 1.0 |
| C | 0.0 | 1.0 | 0.0 | 0.0 | 2.0 | 1.3 | 0.3 | 1.0 | 0.3 | 2.0 | 0.7 | 0.3 | 0.3 | 0.3 |
| C | 0.0 | 1.0 | 0.7 | 0.0 | 1.0 | 3.0 | 1.7 | 1.3 | 1.0 | 1.3 | 1.7 | 0.3 | 0.0 | 0.0 |
| A | 0.0 | 0.0 | 2.0 | 0.7 | 0.3 | 1.7 | 2.7 | 1.3 | 1.0 | 0.7 | 1.0 | 1.3 | 1.3 | 0.0 |
| U | 0.0 | 0.0 | 0.7 | 1.7 | 0.3 | 1.3 | 2.7 | 2.3 | 1.0 | 0.7 | 1.7 | 2.0 | 1.0 | 1.0 |
| U | 0.0 | 0.0 | 0.3 | 0.3 | 1.3 | 1.0 | 2.3 | 2.3 | 2.0 | 0.7 | 1.7 | 2.7 | 1.7 | 1.0 |
| G | 0.0 | 0.0 | 0.0 | 1.3 | 0.0 | 1.0 | 1.0 | 2.0 | 3.3 | 2.0 | 1.7 | 1.3 | 2.3 | 2.7 |
| A | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.3 | 0.7 | 0.7 | 2.0 | 3.0 | 1.7 | 1.3 | 2.3 | 2.0 |
| C | 0.0 | 1.0 | 0.0 | 0.7 | 1.0 | 2.0 | 0.7 | 1.7 | 1.7 | 3.0 | 2.7 | 1.3 | 1.0 | 2.0 |
| G | 0.0 | 0.0 | 0.7 | 1.0 | 0.3 | 0.7 | 1.7 | 0.3 | 2.7 | 1.7 | 2.7 | 2.3 | 1.0 | 2.0 |
| G | 0.0 | 0.0 | 0.0 | 1.7 | 0.7 | 0.3 | 0.3 | 1.3 | 1.3 | 2.3 | 1.3 | 2.3 | 2.0 | 2.0 |

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

# 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

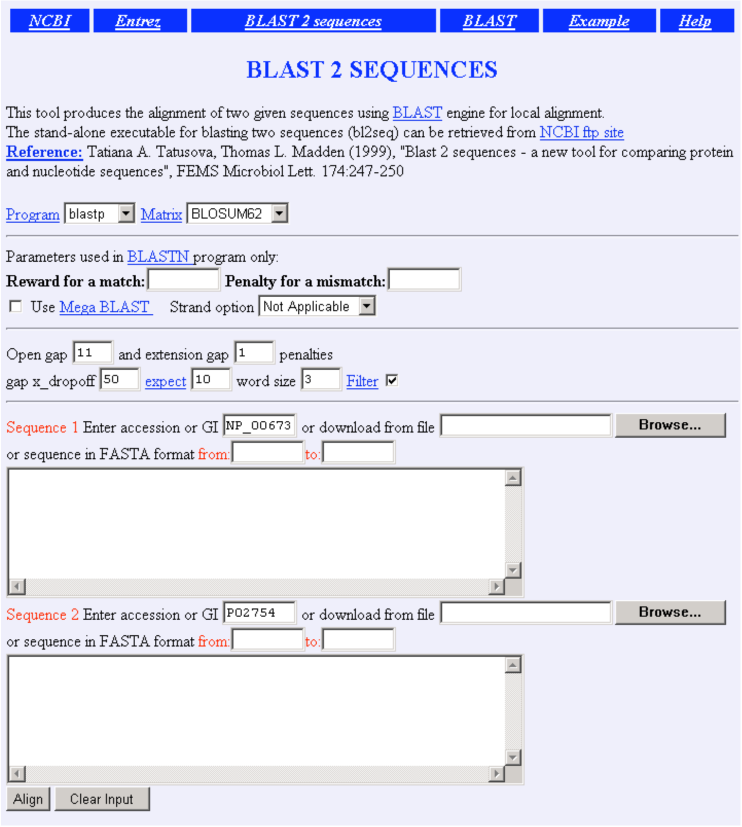
  --3 BLOSUM and 3 PAM matrices

  --gap creation and extension penalties

  --filtering

  --word size

  [4] click “align”



## 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 (bl2seq) can be retrieved from NCBI ftp site
**Reference:** Tatiana A. Tatusova, Thomas L. Madden (1999), "Blast 2 sequences - a new tool for comparing protein and nucleotide sequences", FEMS Microbiol Lett. 174:247-250

Program blastp Matrix BLOSUM62

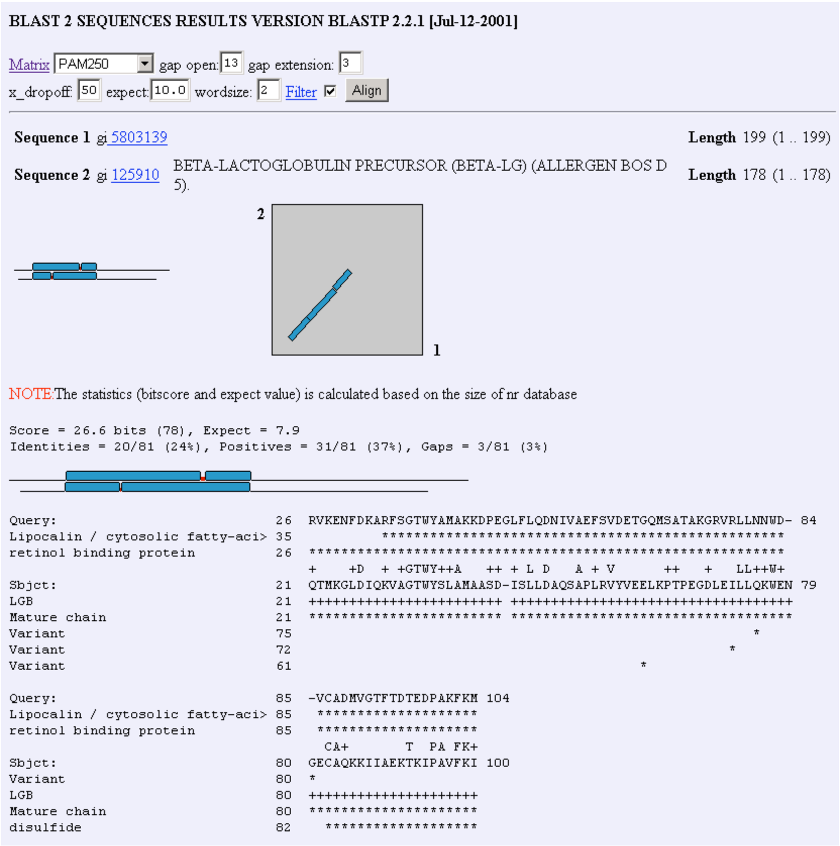
Parameters used in BLASTN program only:
**Reward for a match:** **Penalty for a mismatch:**
□ Use Mega BLAST Strand option Not Applicable

Open gap 11 and extension gap 1 penalties
gap x_dropoff 50 expect 10 word size 3 Filter ☑

Sequence 1 Enter accession or GI NP_00673 or download from file
or sequence in FASTA format from: to:

Sequence 2 Enter accession or GI P02754 or download from file
or sequence in FASTA format from: to:

Fig. 3.27
Page 73



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

Matrix PAM250 gap open: 13 gap extension: 3
x_dropoff: 50 expect: 10.0 wordsize: 2 Filter ☑

**Sequence 1** gi 5803139 **Length** 199 (1 .. 199)

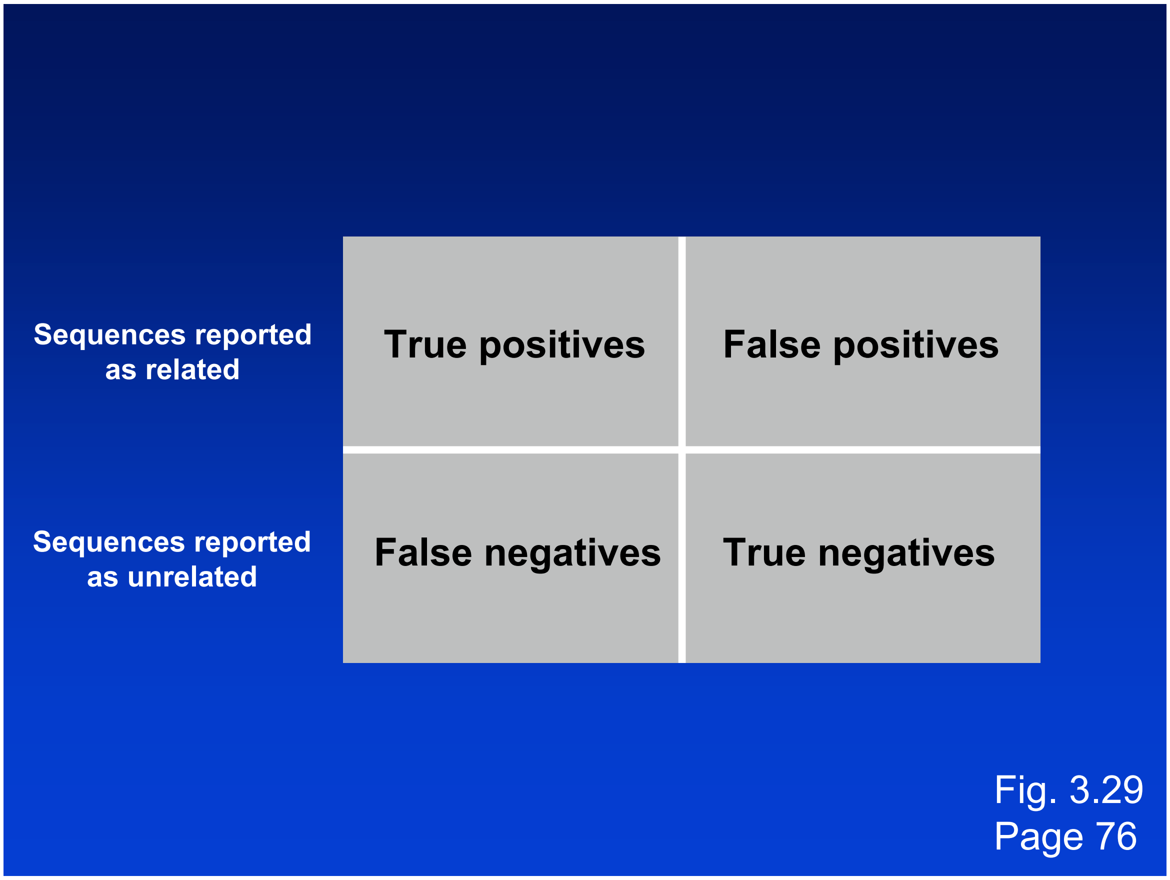
**Sequence 2** gi 125910 BETA-LACTOGLOBULIN PRECURSOR (BETA-LG) (ALLERGEN BOS D 5). **Length** 178 (1 .. 178)

NOTE:The statistics (bitscore and expect value) is calculated based on the size of nr database

```
Score = 26.6 bits (78), Expect = 7.9
Identities = 20/81 (24%), Positives = 31/81 (37%), Gaps = 3/81 (3%)
```

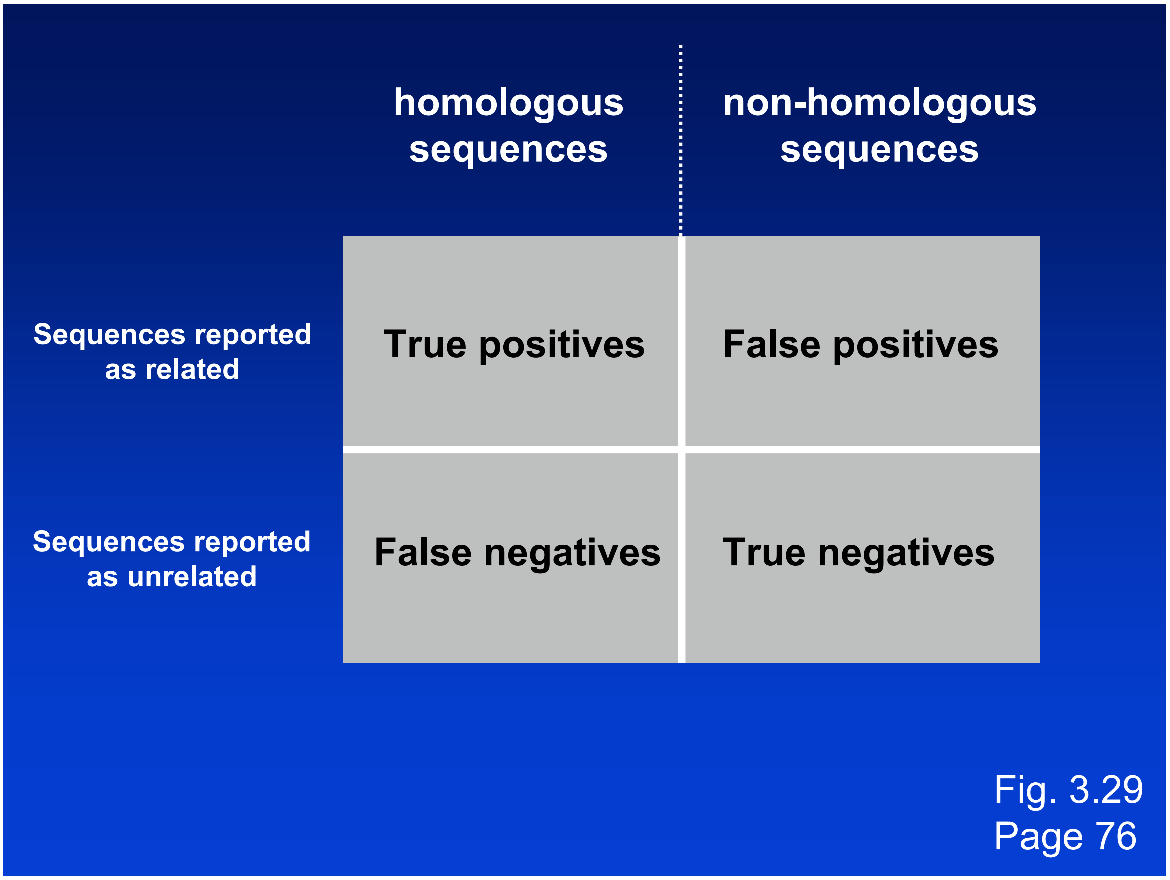
```
Query:                           26  RVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWD- 84
Lipocalin / cytosolic fatty-aci> 35           **************************************************
retinol binding protein          26  ***********************************************************
                                     +    +D  + +GTWY++A   ++ + L D   A + V      ++   +   LL++W+
Sbjct:                           21  QTMKGLDIQKVAGTWYSLAMAASD-ISLLDAQSAPLRVYVEELKPTPEGDLEILLQKWEN 79
LGB                              21  ++++++++++++++++++++++++ +++++++++++++++++++++++++++++++++++
Mature chain                     21  ************************ ***********************************
Variant                          75                                                         *
Variant                          72                                                      *
Variant                          61                                           *

Query:                           85  -VCADMVGTFTDTEDPAKFKM 104
Lipocalin / cytosolic fatty-aci> 85   ********************
retinol binding protein          85   ********************
                                       CA+       T  PA FK+
Sbjct:                           80  GECAQKKIIAEKTKIPAVFKI 100
Variant                          80  *
LGB                              80  +++++++++++++++++++++
Mature chain                     80  *********************
disulfide                        82    *******************
```

Fig. 3.28
Page 74



| | | |
|---|---|---|
| Sequences reported as related | True positives | False positives |
| Sequences reported as unrelated | False negatives | True negatives |

Fig. 3.29
Page 76



| | homologous sequences | non-homologous sequences |
|---|---|---|
| Sequences reported as related | True positives | False positives |
| Sequences reported as unrelated | False negatives | True negatives |

Fig. 3.29
Page 76

|  | homologous sequences | non-homologous sequences |
|---|---|---|
| Sequences reported as related | True positives | False positives |
| Sequences reported as unrelated | False negatives | True negatives |
|  | Sensitivity:<br>ability to find true positives | Specificity:<br>ability to minimize false positives |