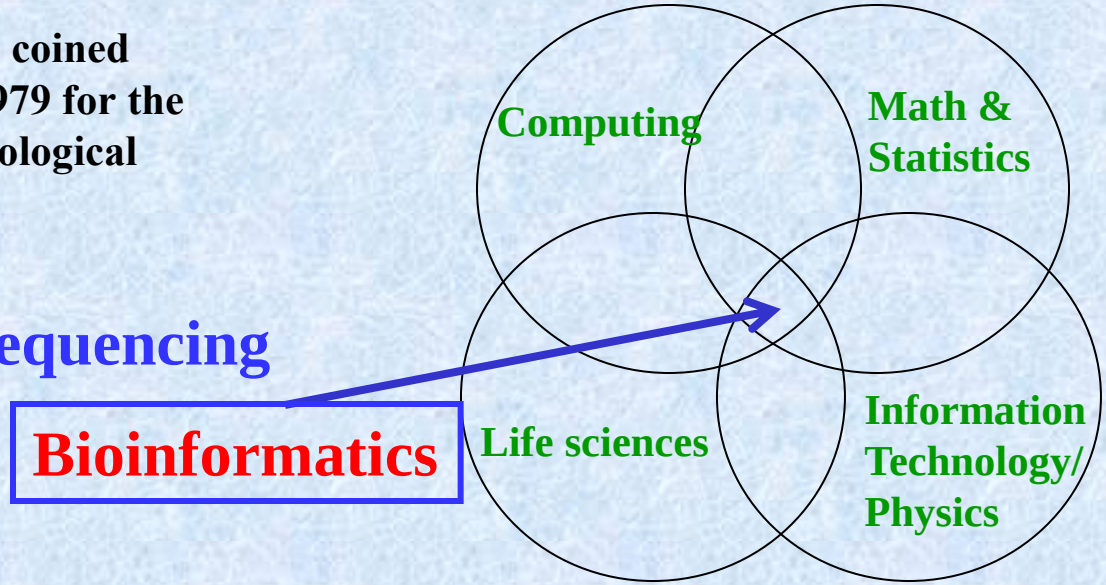
# Bioinformatics

**Bioinformatics is a field of science in which biology, computer science, and information technology merge into a single discipline to analyze biological information using computers and statistical techniques.**

**The term bioinformatics was coined by Paulien Hogeweg in 1979 for the study of informatics in biological systems.**

**Next Generation Sequencing**




**Biological experiments can generate large amounts of data (macromolecular sequences, structures, expression profiles, pathways etc).**

**Bioinformatics is about acquiring, managing, analyzing and understanding those data.**

# Bioinformatics: major aspects

1. Well organized databases

2. Computationally derived hypothesis

3. Web servers (tools/online applications)

4. Virtual screening of compounds for drug development

# Example 1

**Food pattern and weight control**

1. Rice 50%; wheat 25%; meat 10%; fruits 10%; vegetables: 5%: Not control
2. Rice 40%, wheat 5%; meat 15%; fruits 5%; vegetables: 35%: Not control
3. Rice 25%, wheat 10%, meat 10%, fruits 25%, vegetables 30%: Control
4. Rice 15%, wheat 15%, meat 40%, fruits 20%, vegetables 10%: Not control
5. Rice 15%, wheat 20%, meat 5%; fruits 15%; vegetables 45%: Control

**Known experimentally**

**Test case:**

Rice 20%; wheat 10%, meat 10%, fruits, 20%, vegetables 40%

?

Computer science (statistical methods)

Statistical approach

Rice + wheat ≤ 35%

AND

Meat < 15%

AND

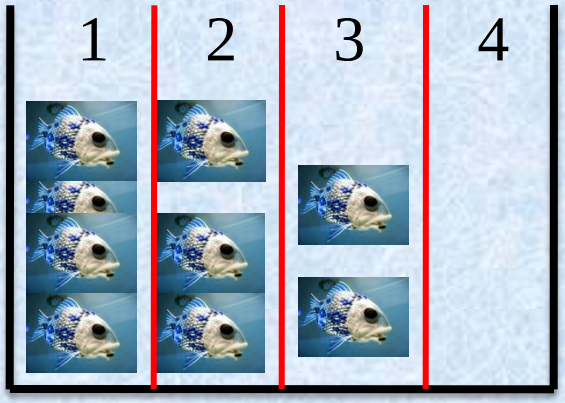
Vegetables ≥ 30%

## Answer: Control

Bioinformatics can handle large amount of data and provide possible solutions

# Example 2

## Finding fish in a pond




Experiments:

All compounds (trial and error)

35 million compounds, ZINC database

2.2 million: Enamine

35,000: Natural compounds

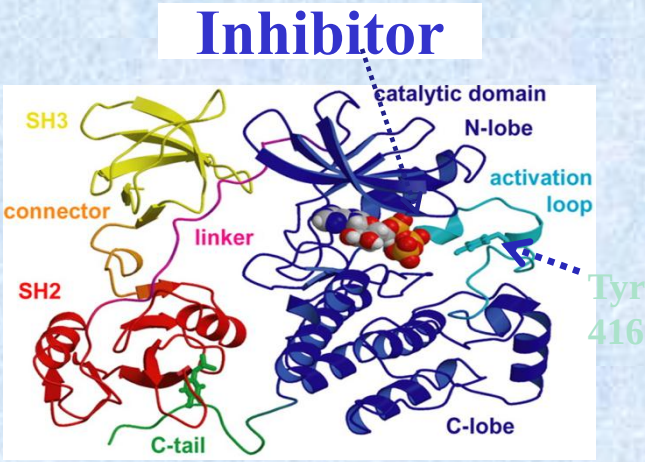
Finding suitable drug for a disease

| | Disease |
|---|---|
| Compound 1 | X |
| Compound 2 | X |
| Drug | O |

Solution: **BIOINFORMATICS**

**Computational techniques assist one in searching drug target and designing drug.**

# Initiative for Parallel Bioinformatics (IPAB)




**Organizer: Tokyo Institute of Technology**

**Target: cYes Kinase**

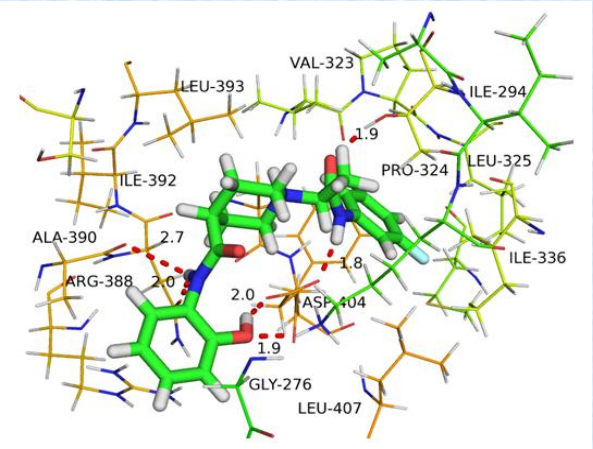
**Library: Enamine (2.2 million compounds)**

**Identification: 120 compounds**

**Test: 50 compounds**

**Inhibition: 4 Compounds**

**Probable hit: 1 compound**

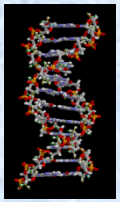



**2015: 200 compounds; inhibition: 5; hits: 2**

# Complexity of biological systems

**Nucleotide sequence** → **Nucleotide structure** → **Gene expression**

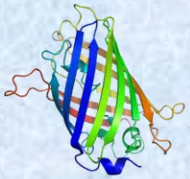
ATCGGGATAT

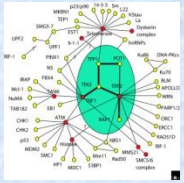


↓

**Protein sequence** ← **Protein structure** ← **Protein interactions**

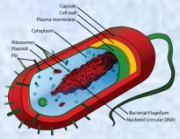
ATCHKLVYAT





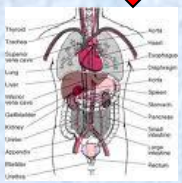
↓

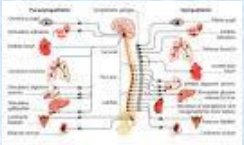
**Tissues** → **Cell signaling** → **Cell**



↓

**Organs** → **Physiology** → **Organism**





# The genetic material

DNA (deoxyribonucleic acid) is the genetic material.

The information stored in DNA allows the organisms that are able to regulate their internal chemical composition, growth and reproduction.

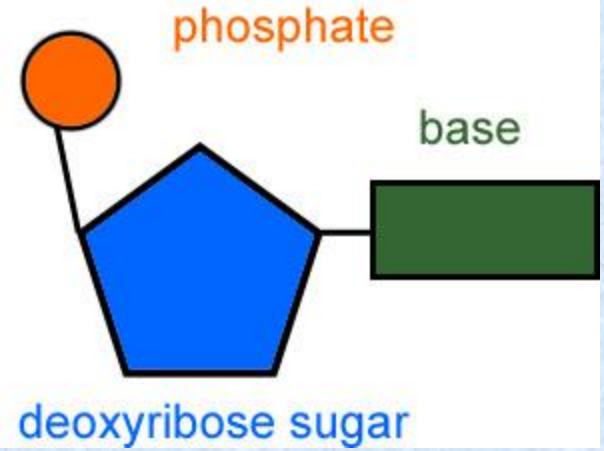
It allows to inherit our mother's curly hair, father's eyes etc.

The various units that govern those characteristics at the genetic level are called **genes**.

Genes contains specific sequence of nucleotides.

Nucleotide

(i) phosphate group,
(ii) central deoxyribose sugar and
(iii) base

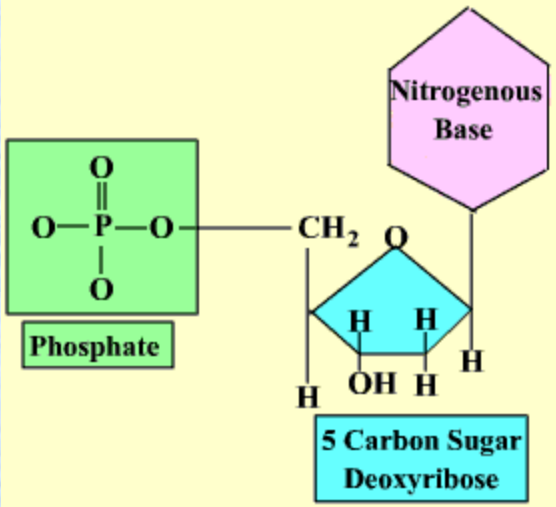


# Four different bases

DNA is made and resides in the nucleus of living cells.

DNA gets its name from the sugar molecule contained in its backbone (deoxyribose);

Four different nucleotide bases occur in DNA: **adenine (A), cytosine (C), guanine (G), and thymine (T).**

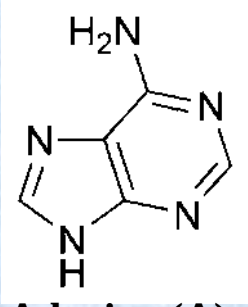



Deoxyribose Nucleotide

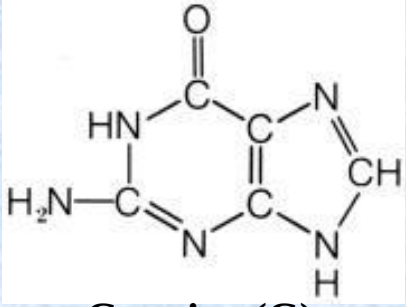
**1.Methylation protects the DNA.**
**2.Maintains the fidelity of DNA replication.**

## Purines

**The purines have a double-ring shape, with a pentagon attached to one side of a hexagon**



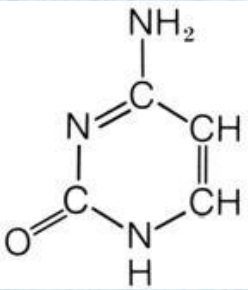
**Adenine (A)**



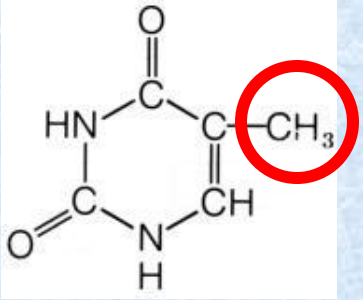
**Guanine (G)**

## Pyrmidines

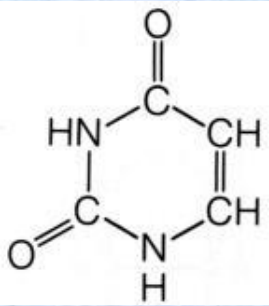
**The pyrimidines have a hexagon shape, generally made with four carbons and two nitrogen**



**Cytosine (C)**

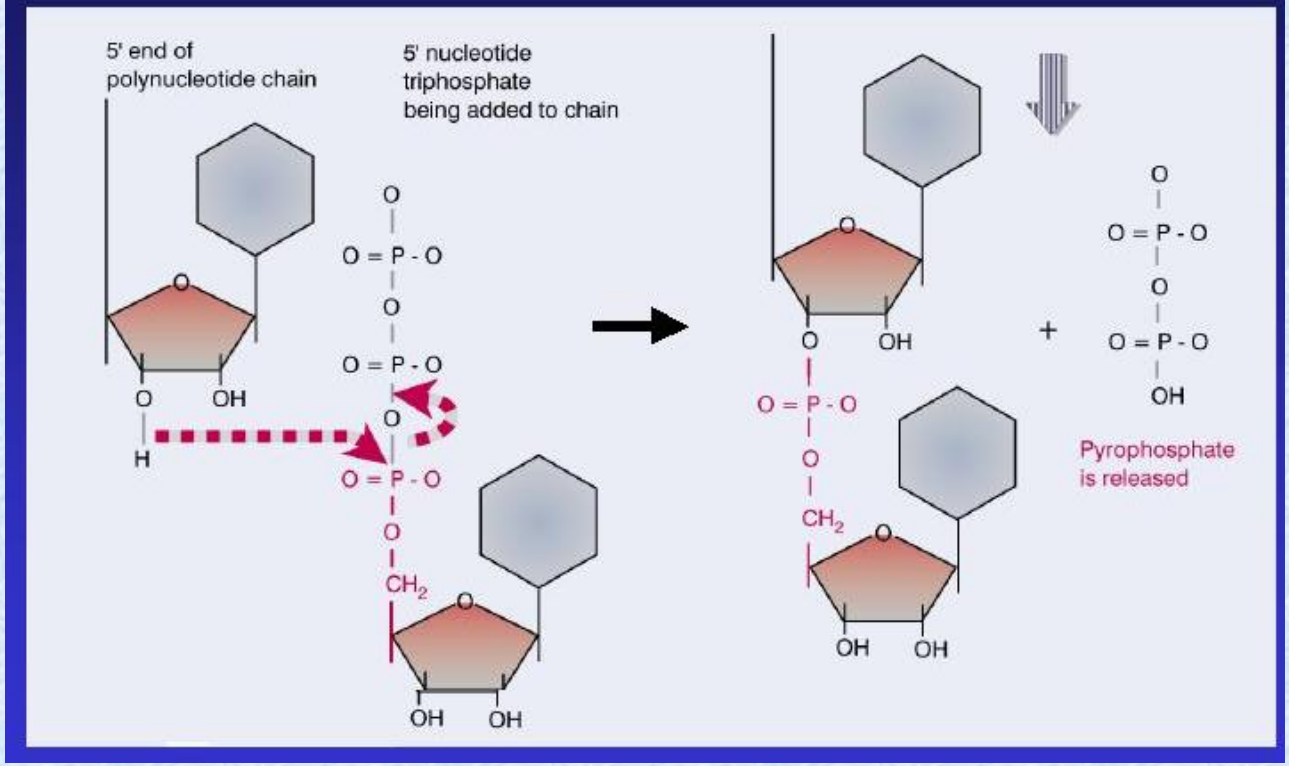


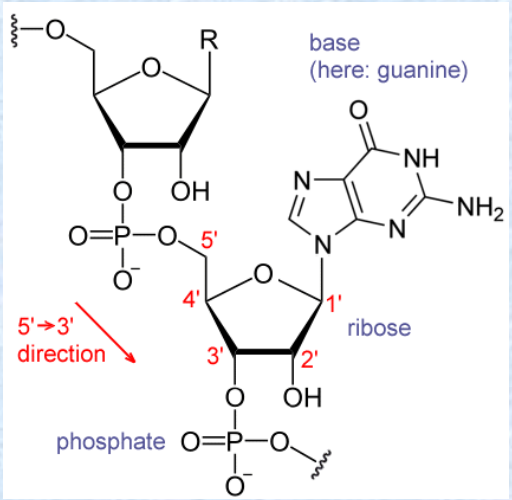
**Thymine (T)**

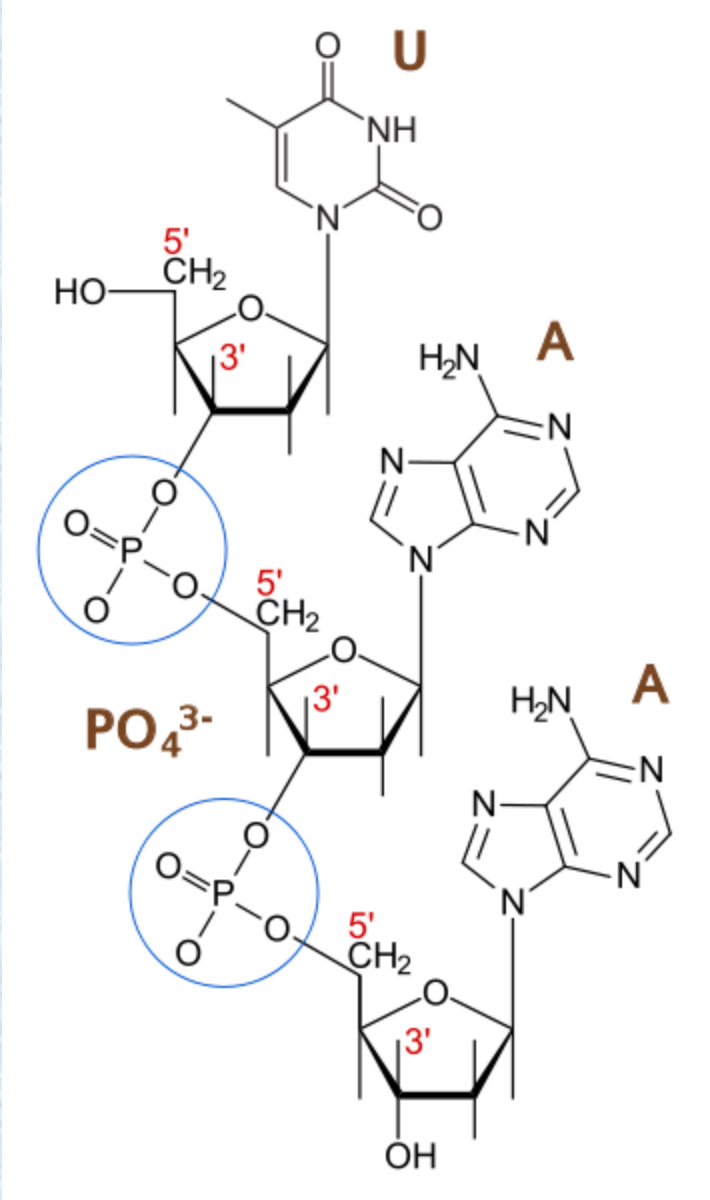


**Uracil (U)**

# Phosphodiester bond formation



# DNA

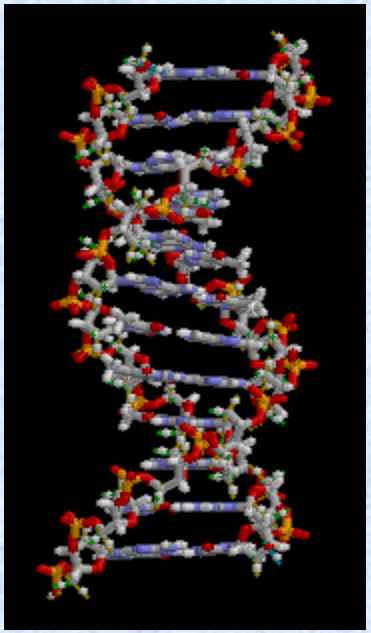
**DNA molecules within cells typically exist as double-stranded molecules.**

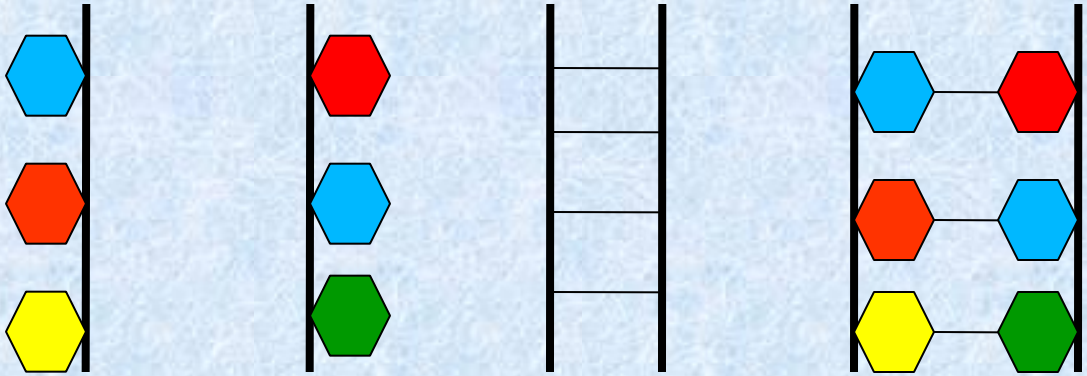
**The information content on one of those strands is essentially redundant with the information on the other (1953).**

**The redundant strand is not the same – it is complementary.**

**For every G on one strand, a C is found on its complementary and vice-versa**

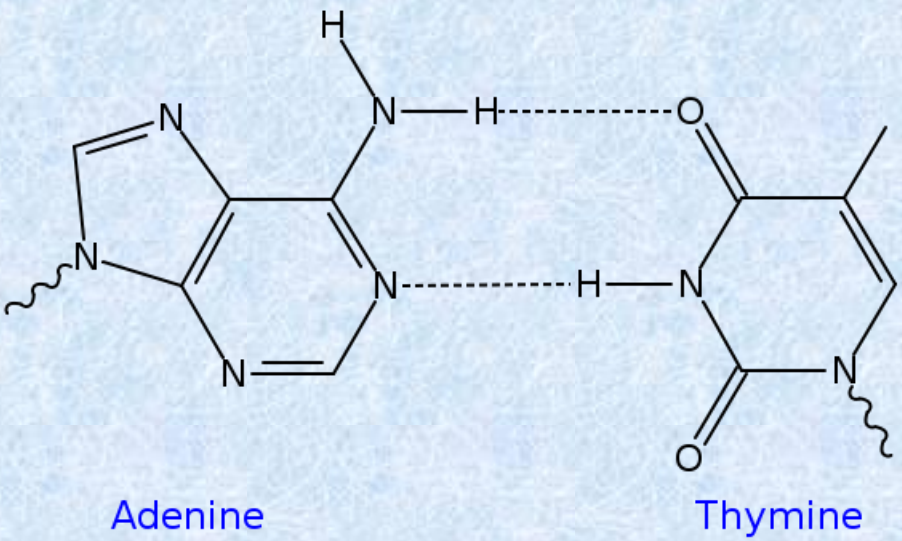
**For every A on one strand, a T is found on its complementary and vice-versa**





**Watson and Crick, 1962, Nobel prize**

# DNA base pairing



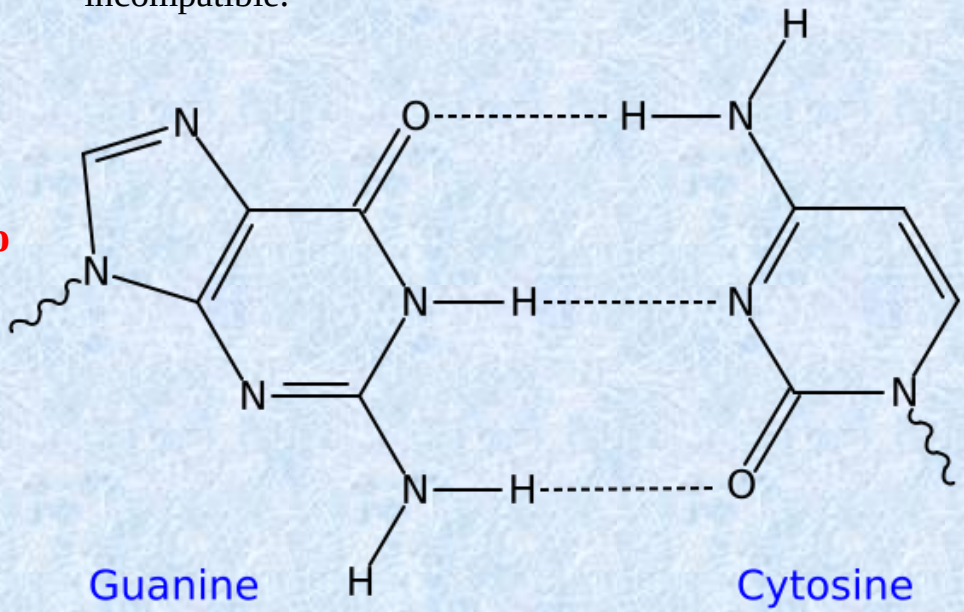

The nitrogenous base G with its two-ringed structure is simply too large to pair with a two-ringed A or another G in the space that usually exists between the two DNA strands.

Same way, it is too small between T and T/C.

Space is not a barrier to interaction between G's and T's or A's and C's but their chemical nature are incompatible.

**This bonding is specific, and adenine always bonds to thymine (and vice versa) and guanine always bonds to cytosine (and vice versa).**

**Right spacing and interaction between their chemical groups to form stable base pairs.**



# Refresh

**Bioinformatics: introduction**

**Features of bioinformatics**

**Applications of bioinformatics**

**Bioinformatics in different complexities of biological systems**

# Complementary strand

**Although the two strands of a DNA molecule are complementary they are not in the same 5'/3' orientation.**

**Instead the two strands are said to be antiparallel.**

**5' ACGTTACG 3'**

**3' TGCAATGC 5'**

**5' CGTAACGT 3' (most cellular process involving DNA occur in the 5' to the 3' direction).**

**The two strands of double stranded DNA molecule are reverse complements of each other.**

**Example:**

**5' AGCCGTTAAGCTAATTCTGCTAGC 3'**

**Complementary strand is: ?**

**5'**

# Public domain program

## EMBOSS

**EMBOSS**

degapseq
descseq

EMBOSS (European Molecular Biology Open Software Suite) is a suite of free software tools for sequence analysis. There are a wide variety of programs that make up the suite, ranging in application from database searching to presentation of sequence data.

snowseq
sixpack
EDIT
ligncopy
biosed
codcopy
cutseq
degapseq
descseq
entret
extractalign
extractfeat
extractseq
featcopy
featreport
listor
makenucseq
makeprotseq
maskambignuc
maskambigprot
maskfeat
maskseq
newseq
nohtml
noreturn
nospace
notab

**REVSEQ**

*(Reverse and complement a nucleotide sequence)*

Help ?

*Fields with a coloured background are optional and can safely be ignored...*

[ Hide optional fields ]

**1. SET THE PARAMETERS FOR THE RUN (OR ACCEPT THE DEFAULTS...)**

input section

Select a set of sequences.

Use one of the following three fields: *(file must contain DNA sequences)*

1. To access a sequence from a database, enter the USA path here: *(dbname:entry)*
2. Or, upload a sequence file from your local computer here:

   Choose File No file chosen

3. Or enter the sequence data manually here:

ACTGACC

makeprotseq
maskambignuc
maskambigprot
maskfeat
maskseq
newseq
nohtml
noreturn
nospace
notab
notseq
nthseq
pasteseq
revseq
seqret
seqretsplit
sizeseq
skipredundant
skipseq
splitsource
splitter
trimest

3. Or enter the sequence data manually here:

ACTGACC

advanced section

Reverse sequence? Yes

Complement sequence? Yes

output section

Output file format: Pearson FASTA

**2. SUBMIT TO REVSEQ...**

run revseq

**REVSEQ: OUTPUT**

**OUTPUT FILE: outseq** [ RIGHT CLICK TO SAVE ]

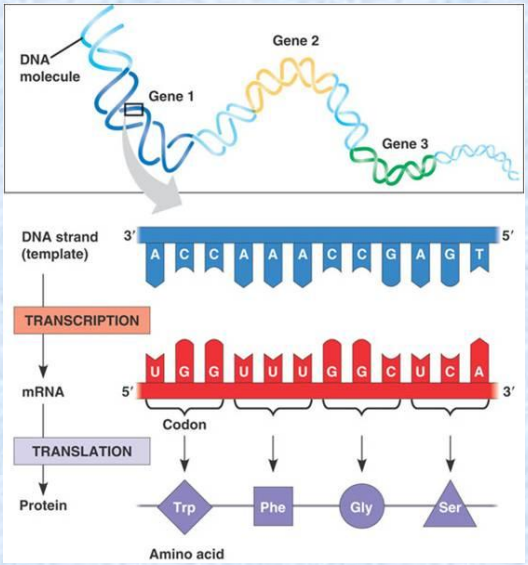
```
>EMBOSS_001 Reversed:
GGTCAGT
```

# Protein synthesis: Central dogma in molecular biology

Information stored in DNA is used to make a more transient, single-stranded polynucleotide called RNA (ribonucleic acid) this is in turn used to make proteins. The process of making an RNA copy of a gene is called **transcription** and is accomplished through an enzymatic activity of an RNA polymerase.

There is one-to-one correspondence between the nucleotides to make RNA (G, A, U, uracil and C).

The process of converting that information from nucleotide sequences in RNA to amino acid sequences that make a protein is called **translation**: and it is performed by a complex of proteins and RNA called **ribosomes**.




DNA $\xrightarrow[\text{RNA polymerase}]{\text{Transcription}}$ RNA $\xrightarrow[\text{Ribosomes}]{\text{Translation}}$ Protein

# Genetic code

**Only 4 different nucleotides are used to make DNA/RNA molecules**

**20 different amino acids are used in protein synthesis .**

**There cannot be one-to-one correspondence between nucleotide and amino acid**

**Combination of 2 gives $4^2$=16, which is less than 20.**

**Four nucleotides can be arranged in a total of 64 different combinations of three.**

**18 of the 20 amino acids are coded by more than one codon and this feature is called degeneracy.**

**It is possible to make the same amino acid sequence with changes in nucleotide.**

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

# DNA/RNA sequence to protein

**What sequence of amino acids would the following RNA sequence code for?**

**ACGUGCGCAUGCAACCGAAUGA**

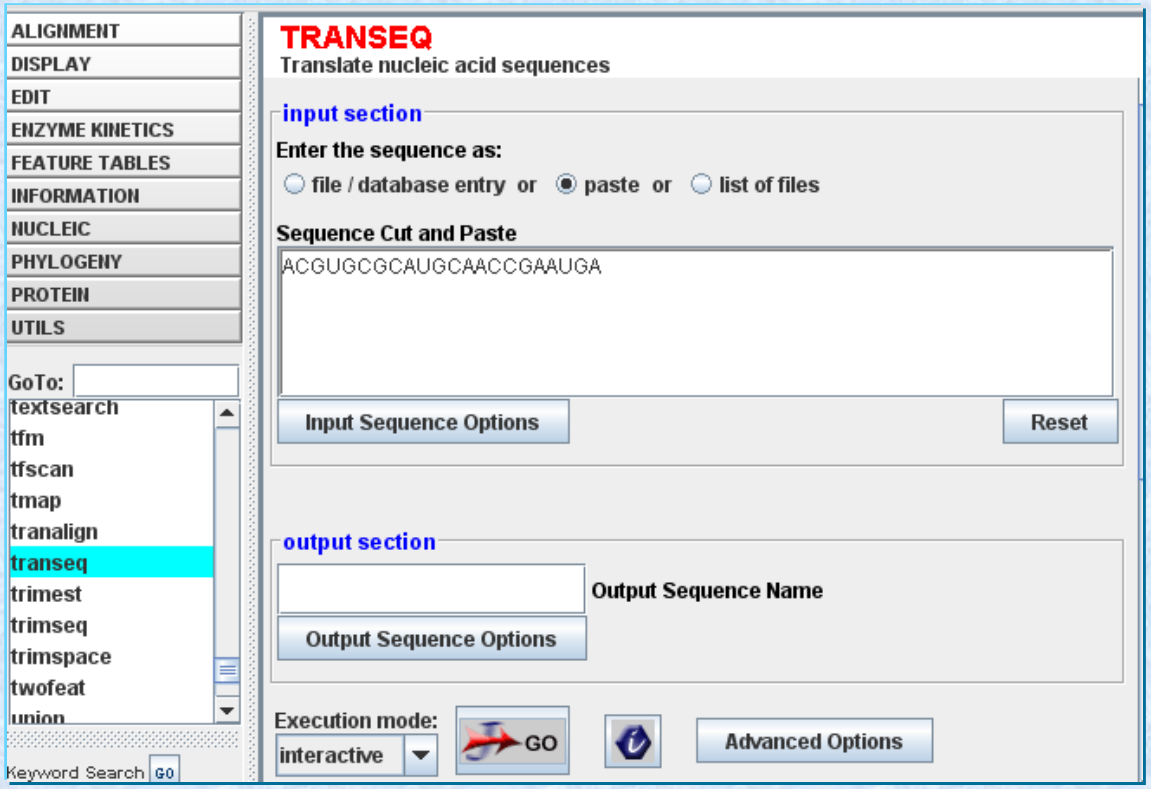
**TCACNRMX**

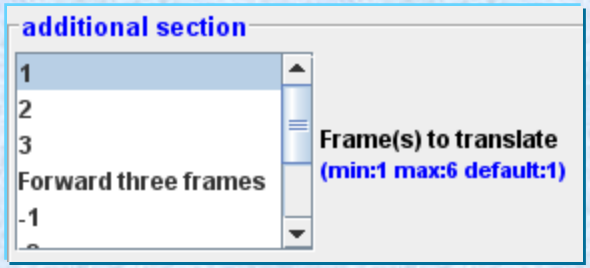
**What will happen if the first nucleotide "A" is deleted?**

**RAHATE***

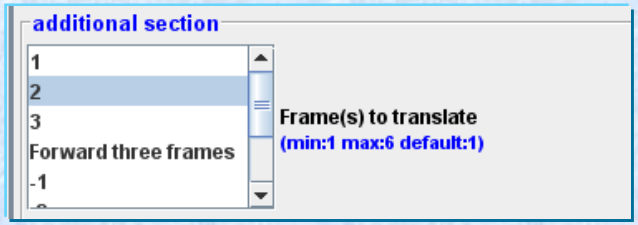
| First letter | Second letter: U | Second letter: C | Second letter: A | Second letter: G | Third letter |
|---|---|---|---|---|---|
| U | UUU, UUC Phe<br>UUA, UUG Leu | UCU, UCC, UCA, UCG Ser | UAU, UAC Tyr<br>UAA Stop<br>UAG Stop | UGU, UGC Cys<br>UGA Stop<br>UGG Trp | U<br>C<br>A<br>G |
| C | CUU,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

# Resources




**TCACNRMX**




**RAHATE***

# Reading frames

```
e.g. 5' CAATGGCTAGGTACTATGTATGAGATCATGATCTTTACAAATCCGAG  3'  DNA

Forward Frames

       CAA TGG CTA GGT ACT ATG TAT GAG ATC ATG ATC TTT ACA AAT CCG AG DNA
        Q   W   L   G   T   M   Y   E   I   M   I   F   T   N   P     Amino Acids

     C AAT GGC TAG GTA CTA TGT ATG AGA TCA TGA TCT TTA CAA ATC CGA G  DNA
        N   G   *   V   L   C   M   R   S   *   S   L   Q   I   R     Amino Acids

    CA ATG GCT AGG TAC TAT GTA TGA GAT CAT GAT CTT TAC AAA TCC GAG    DNA
        M   A   R   Y   Y   V   *   D   H   D   L   Y   K   S   E     Amino Acids
```

# Reverse frames

**Reverse Frames**

Here we take the reverse/complimentry (bottom) strand and reverse it so it starts with the 5' end.

```
5' CAATGGCTAGGTACTATGTATGAGATCATGATCTTTACAAATCCGAG  3'  DNA Top Strand
   ||||||||||||||||||||||||||||||||||||||||||||||
3' GTTACCGATCCATGATACATACTCTAGTACTAGAAATGTTTAGGCTC  5'  DNA Bottom (Complimentry) Strand

5' CTCGGATTTGTAAAGATCATGATCTCATACATAGTACCTAGCCATTG  3'  DNA Bottom (Complimentry) Strand Reversed


    CTC GGA TTT GTA AAG ATC ATG ATC TCA TAC ATA GTA CCT AGC CAT TG DNA
     L   G   F   V   K   I   M   I   S   Y   I   V   P   S   H   X Amino Acids

  C TCG GAT TTG TAA AGA TCA TGA TCT CAT ACA TAG TAC CTA GCC ATT G  DNA
     S   D   L   *   R   S   *   S   H   T   *   Y   L   A   I  X  Amino Acids

 CT CGG ATT TGT AAA GAT CAT GAT CTC ATA CAT AGT ACC TAG CCA TTG    DNA
     R   I   C   K   D   H   D   L   I   H   S   T   *   P   L     Amino Acids
```

ALIGNMENT
DISPLAY
EDIT
ENZYME KINETICS
FEATURE TABLES
INFORMATION
NUCLEIC
PHYLOGENY
PROTEIN
UTILS

GoTo:

textsearch
tfm
tfscan
tmap
tranalign
transeq
trimest
trimseq
trimspace
twofeat
union
vectorstrip
water
whichdb
wobble
wordcount
wordfinder
wordmatch
wossname
yank

# TRANSEQ

Translate nucleic acid sequences

## input section

Enter the sequence as:

○ file / database entry or ◉ paste or ○ list of files

Sequence Cut and Paste

CAATGGCTAGGTACTATGTATGAGATCATGATCTTTACAAATCCGAG

Input Sequence Options | Reset

## output section

Output Sequence Name

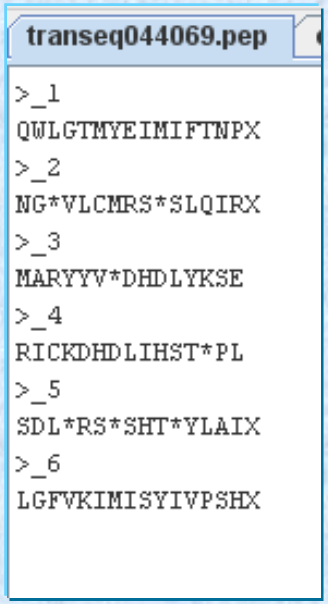
Output Sequence Options

Execution mode: interactive | GO | Advanced Options

## additional section

-1
-2
-3
Reverse three frames
All six frames

Frame(s) to translate
(min:1 max:6 default:1)




**>_1**
**QWLGTMYEIMIFTNPX**
**>_2**
**NG*VLCMRS*SLQIRX**
**>_3**
**MARYYV*DHDLYKSE**
**>_4**
**RICKDHDLIHST*PL**
**>_5**
**SDL*RS*SHT*YLAIX**
**>_6**
**LGFVKIMISYIVPSHX**

# Programming biological problems

1. Find the complementary strand of DNA

2. Convert DNA sequence into protein sequence

3. Number of nucleotides in a DNA sequence.

4. Pair preference

# Computer programming

**Computer hardware requires an operating system for its function.**

**General operating systems are Windows, MacOS and Unix**

**Linux is an open source version of Unix**

**Many libraries and tools in bioinformatics and computational biology are available in linux platforms.**

**The commonly used programming languages are FORTRAN, C, C++, JAVA, PERL, Python etc.**

**The unix environment is very convenient to write computer programs**

**(i) Set up unix operating system**

**(ii) Use unix operating system for programming.**

**When you log on Unix, it prompts with $ for any command**

# Unix commands

## Files

- **ls** --- lists your files
  **ls -l** --- lists your files in 'long format', which contains lots of useful information, e.g. the exact size of the file, who owns the file and who has the right to look at it, and when it was last modified.
  **ls -a** --- lists all files, including the ones whose filenames begin in a dot, which you do not always want to see.
  There are many more options, for example to list files by size, by date, recursively etc.

- **vi *filename*** --- is an editor that lets you create and edit a file.

- **mv *filename1 filename2*** --- moves a file

- **cp *filename1 filename2*** --- copies a file

- **rm *filename*** --- removes a file. It is wise to use the option rm -i, which will ask you for confirmation before actually deleting anything.

- **diff *filename1 filename2*** --- compares files, and shows where they differ

- **wc *filename*** --- tells you how many lines, words, and characters there are in a file

- **chmod *options filename*** --- lets you change the read, write, and execute permissions on your files.

## Finding things

**grep *string filename(s)*** --- looks for the string in the files.

# Unix commands

- **File Compression**

  **gzip** ***filename*** --- compresses files, so that they take up much less space.

  **gunzip** ***filename*** --- uncompresses files compressed by gzip.

  **tar** –cvf home.tar home/

  **tar** –xvf home

  **lpr** ***filename*** --- print.

  **lpq** --- check out the printer queue

  Passwd --- change password

**Directories**

- **mkdir** ***dirname*** --- make a new directory
- **cd** ***dirname*** --- change directory.
- **pwd** --- tells you where you currently are.

# Communicating with other computers

**1. Web (Netscape, Firefox, Explorer, Chrome etc.)**

**The computers recognize each other by their internet protocol (IP) addresses. IP addresses consists of four numbers separated by dots (e.g., 10.93.219.140). These are interpreted as directions to the host by network software.**

**Computers also have hostnames biotech.iitm.ac.in**

**2. Telnet**

**Usage: telnet hostname**

**It opens a shell on a remote Unix machine and one has to give username and password for the unix machine. Then one can use the computer from remote.**

**3. ftp: file transfer protocol**

**It is a method for transferring files from one computer to another**

**E.g. ftp ftp.wwpdb.org (for downloading Protein Structure Data)**

# ftp commands

***bin:*** ***to set the mode of file transfer to binary mode***

**bye**: ***to exit the FTP environment***

**cd**: ***to change directory on the remote machine***

**get**: ***to copy one file from the remote machine to the local machine***

**mget**: ***to copy multiple files from the remote machine to the local machine;***

**mput**: ***to copy multiple files from the local machine to the remote machine;***

**put**: ***to copy one file from the local machine to the remote machine***

**lcd**: ***to change directory on your local machine (same as UNIX cd)***

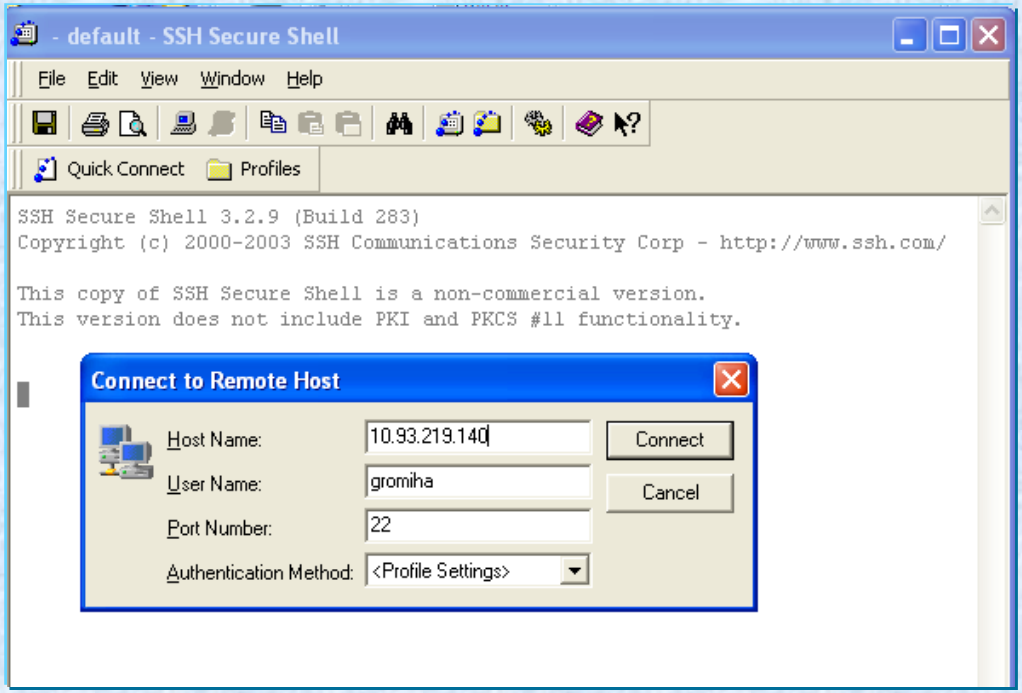
**ls**: ***to list the names of the files in the current remote directory***

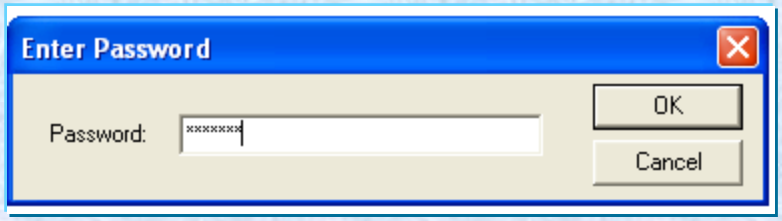
**help**: ***to request a list of all available FTP commands***

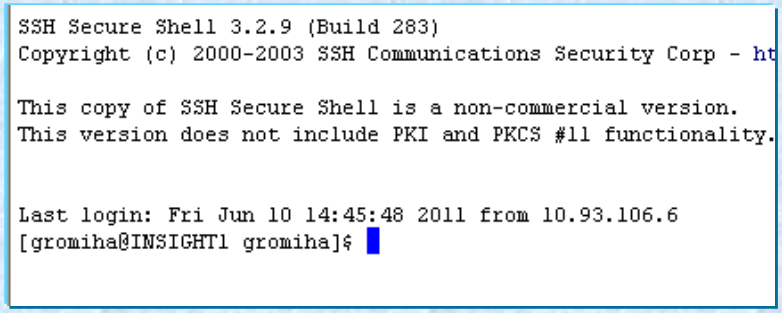
# Connecting with remote computer

SSH is secure shell protocol for file transfer or connecting to remote computer.

It is available for Windows, Unix and MacOS.



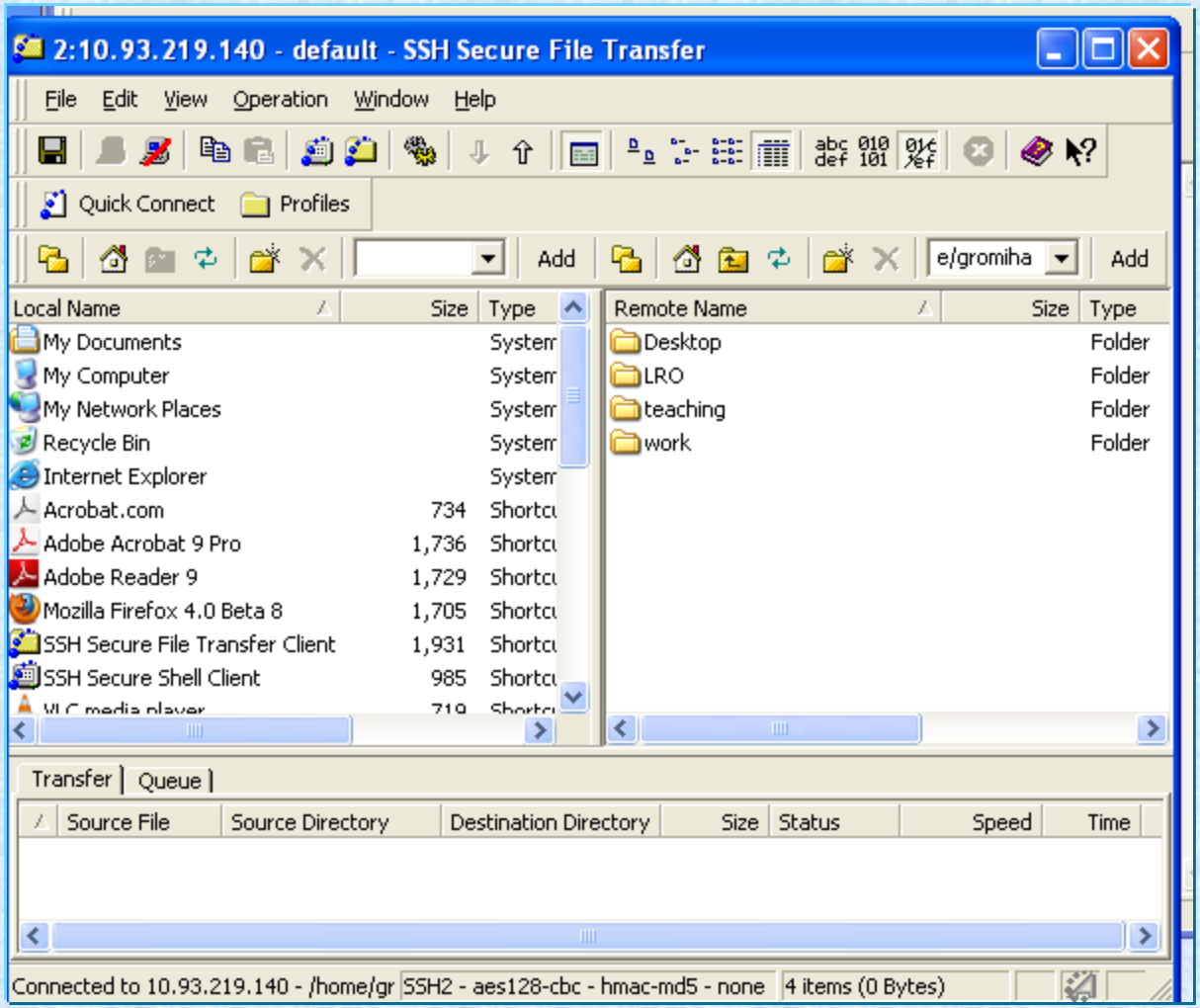




10.21.50.194

bioinfo
btclass123

# File transfer using SSH



# Basic vi commands (Edit)

**Start vi**

**vi filename**: *edit filename starting at line 1*

**vi -r filename**: *recover filename that was being edited when system crashed*

***Exit vi***

**:wq<Return>**: *quit vi, writing out modified file to file named in original invocation*

**:q<Return>**: *quit (or exit) vi*

**:q!<Return>:** *quit vi even though latest changes have not been saved for this vi call*

***Moving cursor***

**j [*or* down-arrow]**: *move cursor down one line;* **k [*or* up-arrow]**: *move cursor up one line*

**h [*or* left-arrow]**: *move cursor left one character;* **l [*or* right-arrow]**: *move cursor right one character*

**0 (zero):** *move cursor to start of current line (the one with the cursor)*

**$:** *move cursor to end of current line*; **W:** *move cursor to beginning of next word*

**B:** *move cursor back to beginning of preceding word;* **:0<Return>:** *move cursor to first line in file*

**:n<Return>** *move cursor to line n;* **:$<Return>:** *move cursor to last line in file*

# Basic vi commands (Edit)

**Insert text**

i: *insert text before cursor, until <Esc> hit*

a: *append text after cursor, until <Esc> hit*

**Deleting text**

x: *delete single character under cursor*

Dd: *delete entire current line*

Ndd *: delete N lines, beginning with the current line; e.g., 5dd deletes 5 lines*

**Cutting and pasting**

Nyy: *copy (yank, cut) the next N lines, including the current line, into the buffer*

p: *put (paste) the line(s) in the buffer into the text after the current line*

**Searching text**

/string: *search forward for occurrence of string in text*

?string: *search backward for occurrence of string in text*

n: *move to next occurrence of search string*

N: *move to next occurrence of search string in opposite direction*

# Details about unix and vi commands

**http://unlser1.unl.csi.cuny.edu/tutorials/viunix/unixman.html**

**http://unix.t-a-y-l-o-r.com/index.html**

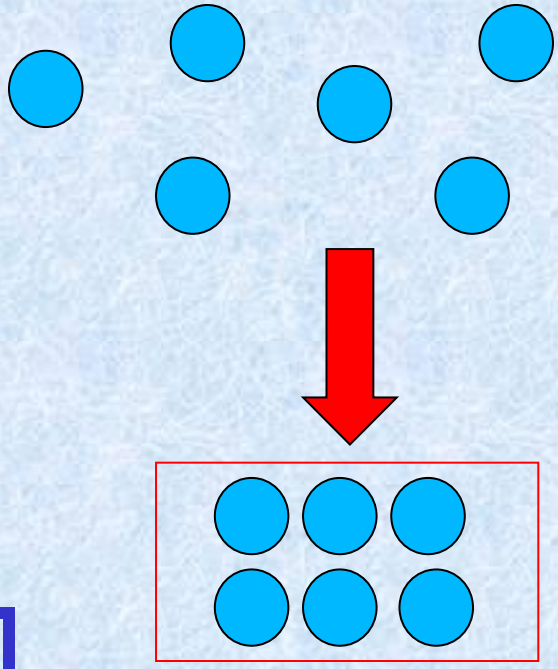
**PDF and ps files**

# Databases

Biological experiments **(macromolecular sequences, structures, expression profiles, pathways etc)** provide wealth of data.

The data are available randomly in the literature

It is necessary to collect the scattered data and put in proper order in the form of a database



**Database is an organized collection of information, in computer-readable form.**

# Databases: Characteristics

1. The contents
2. The ontology: the list of valid terms and their definitions
3. The logical structure, or the expression of the inter-relationships among the data, called the schema.
4. The format of the data
5. The routes for selective retrieval of data, and presentation of results, or pasting them on to a program for analysis
6. Links to other information resources: other databases, references to original publications of data, tutorial background etc.

# Characteristics

| | |
|---|---|
| NO. | 193 |
| ***** Sequence and structural information***** | |
| PROTEIN NAME | Lysozyme |
| SOURCE | Bacteriophage T4 |
| LENGTH | 164 |
| MOLECULAR WEIGHT | 18603.94 |
| PIR_ID | LZBPT4 |
| SWISSPROT_ID | LYCV_BPT4 (P00720) |
| E.C.NUMBER | EC3.2.1.17 Go to BRENDA |
| PMD NO | A921426 |
| PDB_wild | 2LZM Homologous PDB Entries |
| PDB_mutant | 1L56 |
| MUTATION | K 60 P |
| NO. OF MOLECULE | 1 |
| SECONDARY STRUCTURE | Helix |
| ACCESSIBLE SURFACE AREA | 107.7 A**2 |

1

```
NO.
PROTEIN          Staphylococcal nuclease
SOURCE           Staphylococcus aureus
MUTATION          G 20 E
MUTATED_CHAIN
PDB_mutant
SEC.STR.
ASA
NO_MOLECULE
STATE
dG_H2O        6.3
ddG_H2O       -3.2
dG
ddG
T             20
Tm
dTm
dHvH
dHcal
m              4.7
Cm             1.3
dCp
pH            3.9
BUFFER_NAME   HEPES
BUFFER_CONC   10 mM
ION_NAME_1    NaCl
ION_CONC_1    100 mM
ADDITIVES
PROTEIN_CONC  50 microg/mL
MEASURE       pI
METHOD        GdnHCl
REVERSIBILITY unknown
ACTIVITY
ACTIVITY_Km
ACTIVITY_Kcat
ACTIVITY_Kd
KEY_WORDS     dielectric effect; electrostatics; hydration; pKa; bioenergetics
REFERENCE     Proc Natl Acad Sci U S A. 2010 Sep 14;107(37):16096-100.
AUTHOR        Isom DG, Castañeda CA, Cannon BR, Velu PD, García-Moreno E B.
REMARKS
//
```

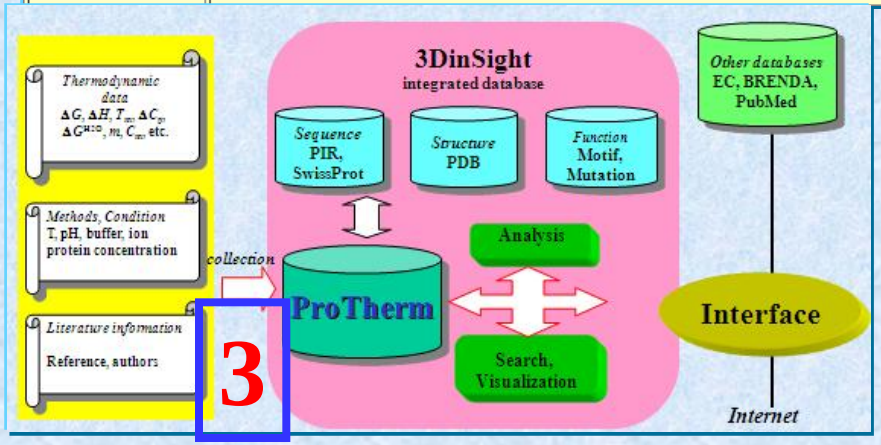
4

Entry [ ] - [ ] PDB Code [ ] Start Clear
Protein [Lysozyme] Source [Bacteriophage]
Mol-weight [ ] To [ ]
Mutation [ ] To [ ] ☑ Single ☐ Double ☐ Multiple ☐ Wild Type
Sec.Structure ☑ Helix ☑ Sheet ☐ Turn ☐ Coil
Accessibility ○ Any ○ Buried ○ Partially Buried ○ Exposed ◉ ASA [0] To [20] [%]
Measure ☐ Absorbance ☐ CD ☐ DSC ☐ Fluorescence ☐ NMR ☐ Others
Method ☑ Thermal ☐ Denaturants ☐ Others
pH [5] To [9]
dTm/Tm/T [dTm] [0] To [50] [C]
dH/dCp/dG/dG_H2O [dH] [ ] To [ ] energy unit: [kcal]
ddG/ddG_H2O [ddG] [ ] To [ ]
State ☑ 2 ☐ 3 ☐ >3
Reversibility [Yes]
Keyword [ ] [OR] [ ]
Author [ ] [OR] [ ]
Year Since [ ] Until [ ]

5

| Terms | Explanations |
|---|---|
| No | Entry number. This option can be used for getting data from a particular entry (Eg. 3012) or a range of entries (Eg. 10107-10365) with/without other search conditions |
| Protein Name | Name of the protein. Multiple words can be entered with spaces. Wild card can also be used: * for a string of characters and ? for a character. If * and ? were to be used as real characters, place "\" (backslash) before them. |
| E.C.No. | Enzyme Commission number |
| PMD No. | Protein Mutant Database accession number |
| PDB_wild | Protein Data Bank code for the native protein |

2




3

| Entry | Protein | PDB_wild | PDB_mutant | Mutation | Sec.Str. | ASA (%) | Tm | dTm | pH | REFERENCE |
|---|---|---|---|---|---|---|---|---|---|---|
| 1459 | LYSOZYME | 2LZM | 1QT6 | E 11 H | H | 18.89 | 65.20 | 0.10 | 5.40 | PROC NATL ACAD SCI U S A 92, 452-456 (1995) |
| 1522 | LYSOZYME | 2LZM | 1L87 | F 153 L | H | 0.05 | 65.68 | 0.80 | 5.70 | J MOL BIOL 229, 747-769 (1993) |
| 199 | LYSOZYME | 2LZM | 1L23 | G 77 A | H | 4.83 | 65.60 | 0.90 | 6.50 | BIOPOLYMERS 32, 1431-1441 (1992) |

PubMed.gov
U.S. National Library of Medicine
National Institutes of Health
Search: PubMed Limits Advanced search Help

Display Settings: Abstract

Nucleic Acids Res. 1999 Jan 1;27(1):286-8.

**ProTherm: Thermodynamic Database for Proteins and Mutants.**

Gromiha MM, An J, Kono H, Oobatake M, Uedaira H, Sarai A.

Tsukuba Life Science Center, The Institute of Physical and Chemical Research (RIKEN), 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan

6

# Organization: Relational database

**In 1970, E.F. Codd from IBM described the relational database.**

**The basic unit of a relational database is a set of correspondence between different features of the database contents, called tables.**

**Relational database is the one in which data are organized as tables, each table comprising a group of records with the same fields (known as attributes). This allows related data to be linked (reassembled) as required without reorganizing the original tables.**

**The set theoretic operations (union, intersection, difference, Cartesian product) on tables facilitate processing of logically complex queries.**

# Example

Table 1

Two tables from a relational database of properties of amino acids

| Amino acid | 3-letter code | 1-letter code | Volume (Å$^3$) | Surface area (Å$^2$) | Distal group |
|---|---|---|---|---|---|
| Alanine | Ala | A | 88.6 | 115 | Methyl |
| Arginine | Arg | R | 173.4 | 225 | Guanidinium |
| Asparagine | Asn | N | 111.1 | 150 | Amide |
| Aspartic acid | Asp | D | 114.1 | 160 | Carboxyl |
| Cysteine | Cys | C | 108.5 | 135 | Sulphydryl |
| Glutamic acid | Glu | E | 138.4 | 190 | Carboxyl |
| Glutamine | Gln | Q | 143.8 | 180 | Amide |
| Glycine | Gly | G | 60.1 | 75 | Hydrogen |
| Histidine | His | H | 153.2 | 195 | Imidazole |
| Isoleucine | Ile | I | 166.7 | 175 | Methyl |
| Leucine | Leu | L | 166.7 | 170 | Methyl |
| Lysine | Lys | K | 168.6 | 200 | Amino |
| Methionine | Met | M | 162.9 | 185 | Methyl |
| Phenylalanine | Phe | F | 189.9 | 210 | Phenyl |
| Proline | Pro | P | 112.7 | 145 | Pyrrolidine |
| Serine | Ser | S | 89.0 | 115 | Hydroxyl |
| Threonine | Thr | T | 116.1 | 140 | Hydroxyl |
| Tryptophan | Trp | W | 227.8 | 255 | Indole |
| Tyrosine | Tyr | Y | 193.6 | 230 | Phenol |
| Valine | Val | V | 140.0 | 155 | Methyl |

Two tables from a relational database of properties of amino acids *(continued)*

| Distal group | H-bond donor | H-bond acceptor |
|---|---|---|
| Amide | yes | yes |
| Amino | yes | no |
| Carboxyl | no | yes |
| Guanidinium | yes | yes |
| Hydrogen | no | no |
| Hydroxyl | yes | yes |
| Indole | yes | yes |
| Methyl | no | no |
| Phenol | yes | yes |
| Phenyl | no | no |
| Pyrrolidine | yes | no |
| Sulphydryl | yes | no |

**Simple: What are the three letter codes of the amino acids, which have distal carboxyl group?**

**View**

**Compound: What are the three letter codes of the amino acids with volume more than 125 A$^3$ and have distal carboxyl group?**

**What are the three letter codes of the amino acids, which can serve as hydrogen bond donors?**

**Join**

Lesk, 2008

# Example

**General form of joining is the Cartesian product of the two tables. If the set contains n and m elements the product will contain nm elements. Here, 20 amino acids and 12 distal groups and the total will be 240 rows.**

| From Table 1 | | | | | | From Table 2 | | | |
|---|---|---|---|---|---|---|---|---|---|
| AA | 3 | 1 | V | A | Group | Group | Donar | Acceptor | |
| Alanine | Ala | A | 88.6 | 115 | Methyl | Amide | Yes | Yes | |
| Alanine | Ala | A | 88.6 | 115 | Methyl | Amide | Yes | No | |
| Alanine | Ala | A | 88.6 | 115 | Methyl | Methyl | No | No | |
| Aspartic acid | Asp | D | 114.1 | 160 | Carboxyl | Carboxyl | No | Yes | |

**Three letter codes of amino acids that have side chains that could serve as hydrogen bond acceptors:**

**Natural join**

# Complex queries

What are the three letter codes of amino acids with volumes between 100 and 150 AND [(that can serve as hydrogen bond donors AND NOT serve as hydrogen bond acceptors) OR (that have surface areas greater than 120 A2 AND have distal methyl groups)].

The structured Query Language (SQL) is fairly well standardized syntax for probing relational databases with complex queries.

Complex queries containing logical connectivities are translatable into Codd's set of operations on tables.

**Syntax**

SELECT <3_letter_code> from <amino_acid _table)

WHERE (sidechain_volume between 100 and 150)

AND

(H-bond_donor = "yes" AND H-hond_acceptor = "no")

OR

(surface_area > 120 AND distal_group = "methyl"))

# Database collections

Nucleic acid research Database issue (First issue in every year). It is available for free access.

**http://nar.oupjournals.org/**

Listing of databases

http://www.oxfordjournals.org/nar/database/a/

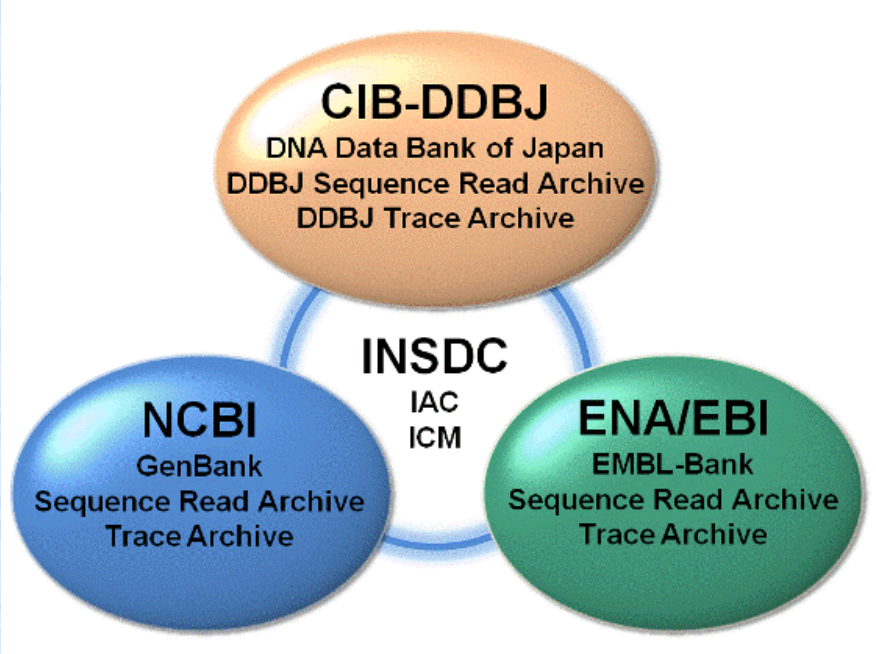
# Database categories

- Nucleotide Sequence Databases
- RNA sequence databases
- Protein sequence databases
- Structure Databases
- Thermodynamic databases
- Genomics Databases (non-vertebrate)
- Metabolic and Signaling Pathways
- Human and other Vertebrate Genomes
- Human Genes and Diseases
- Microarray Data and other Gene Expression Databases
- Proteomics Resources
- Other Molecular Biology Databases (PUBMED)
- Organelle databases
- Plant databases
- Immunological databases

# Nucleotide sequence databases

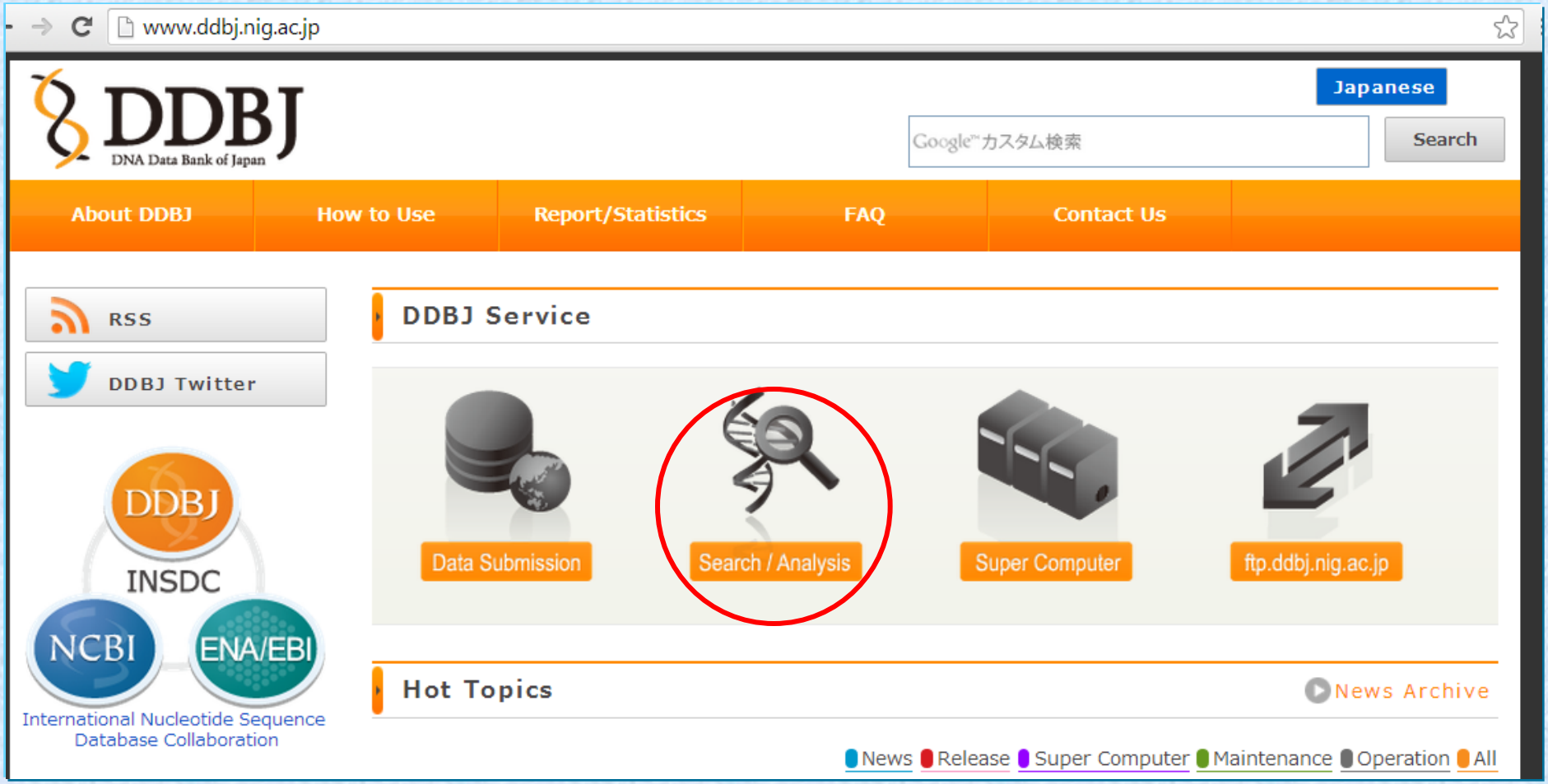
## International collaboration

1. DDBJ (DNA Data Bank of Japan)
2. EMBL (European Molecular Biology Laboratory)
3. Genbank (USA)



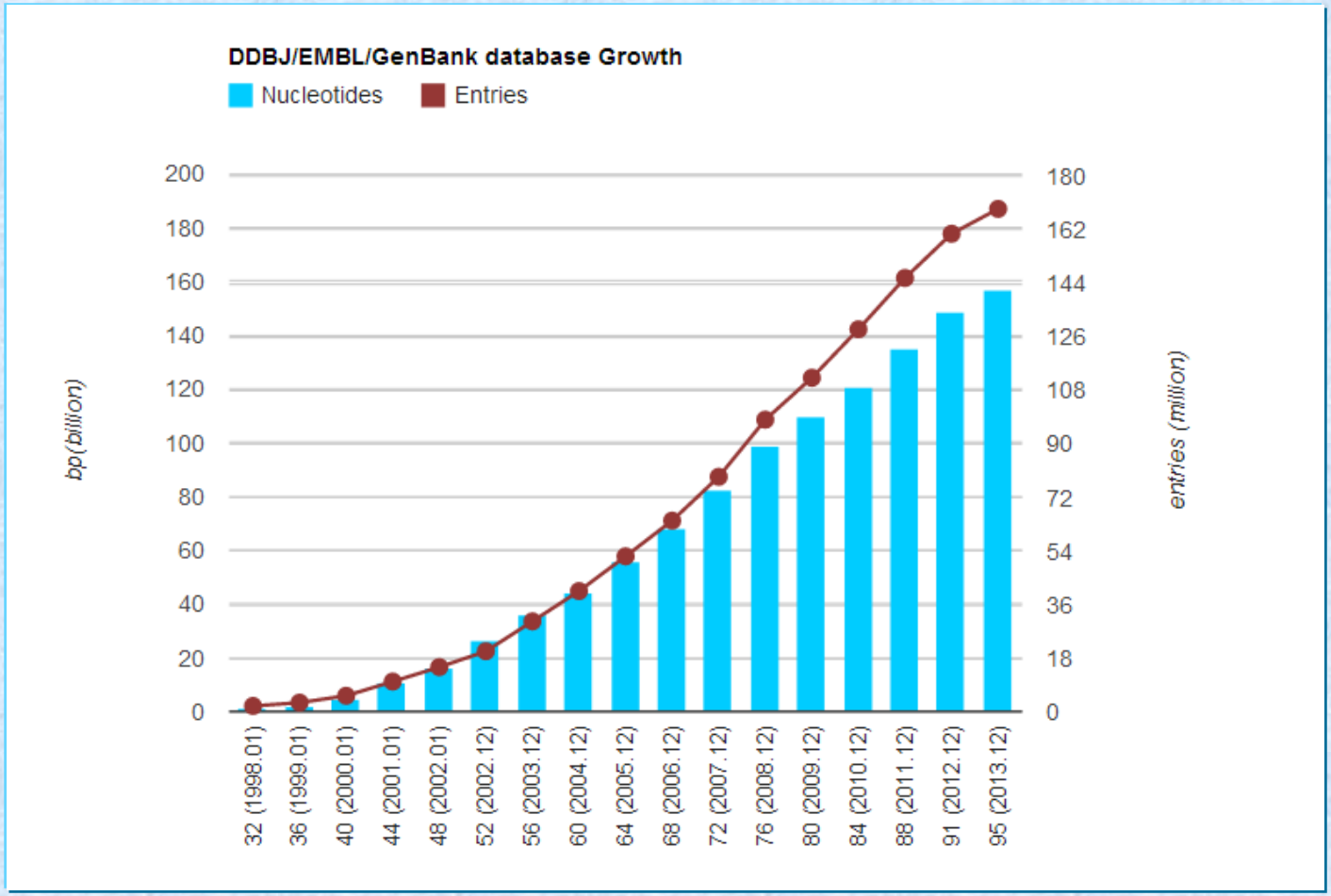

They exchange sequences via SINET3 Computer network
SINET: Science Information Network

# 1. DNA Data Bank of Japan




http://www.ddbj.nig.ac.jp/

# DDBJ: Growth




Dec 2016: > 2.1 Trillion nucleotides

Number of sequences: > 790 millions

# DDBJ: Topmost 10 organisms

| No. Organism | Nucleotides | Entries |
|---|---|---|
| 001 Homo sapiens | 17382318654 bp | 20522483 |
| 002 Mus musculus | 9988986985 bp | 9727377 |
| 003 Rattus norvegicus | 6525616319 bp | 2197920 |
| 004 Bos taurus | 5390887315 bp | 2202851 |
| 005 Zea mays | 5076664404 bp | 3963159 |
| 006 Sus scrofa | 4890614452 bp | 3289376 |
| 007 Danio rerio | 3120659920 bp | 1726789 |
| 008 Marine metagenome | 2482805950 bp | 3173890 |
| 009 Vitis vinifera | 1555395119 bp | 810152 |
| 010 Hordeum vulgare | 1455338435 bp | 1009480 |

# Search

## Search Condition

### Quick Search

Homo sapiens mRNA Glyceraldehyde-3-phosphate-dehydrogenase

Search AND

### Available Fields

## Search Result

### Facet

### List of Entries

1 - 1 entries / Number of founds: 1 FlatFile XML Fasta View selected Download selected Download All

| PrimaryAccessionNumber | Definition | SequenceLength | MolecularType | Organism |
|---|---|---|---|---|
| Z36833 | Definition:H.sapiens (xs4) mRNA, 315bp. | SequenceLength:315 | MolecularType:mRNA | Organism:Homo sapiens |

End of search results

PAGE TO

# Contents

Name
Source
Accession number
Keywords
Authors
Reference
PUBMED index
Nucleotide sequence
Number of bases
A, T, C and G
Protein sequence (translated)

```
LOCUS       Z36833                 315 bp    mRNA    linear   HUM 10-APR-1997
DEFINITION  H.sapiens (xs4) mRNA, 315bp.
ACCESSION   Z36833
VERSION     Z36833.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 315)
  AUTHORS   Mueller-Pillasch,F., Gress,T., Lehrach,H. and  Adler,G.
  TITLE     Differential gene expression in pancreatic cancer. Use of an
            automated approach for the large scale isolation and
            characterisation of cDNA clones containing differentially expressed
            sequences.
  JOURNAL   Unpublished.
REFERENCE   2  (bases 1 to 315)
  AUTHORS   Gress,T.
  JOURNAL   Submitted (16-AUG-1994) to the INSDC. Gress T., University of Ulm,
            Department of Internal Medicine I, Robert Koch Str.8, 89081 Ulm,
            Germany, 89081
REFERENCE   4  (bases 1 to 315)
  AUTHORS   Gress,T.M., Muller-Pillasch,F., Geng,M., Zimmerhackl,F.,
            Zehetner,G., Friess,H., Buchler,M., Adler,G. and  Lehrach,H.
  TITLE     A pancreatic cancer-specific expression profile
  JOURNAL   Oncogene 13(8), 1819-1830(1996).
   PUBMED   8895530
FE
```

```
BASE COUNT          80 a          74 c          82 g          76 t
ORIGIN
        1 atcagcggat ttgcgtcgta ttgggcgcct ggatcaccag ggctgctttt atctctggta
       61 aagtggatat tgttgacatc actgaccccc acattgacca catatacatg gtttacatgt
      121 tccaatatga ctccacccat gagatattcn atgacaccga caggggtgag aacgggnagc
      181 ttgacatcaa tggaaatccc acaccatctn cgaggagaga catcctccaa catcatgtgg
      241 cgagatgttg cgctgagtac gtcgtggagt cactgtgtct cacacatgag aggtgtgctc
      301 atggaggggg agcaa
//
```

# Advanced search

Advanced Search

Quick Search

Field

Show examples

Primary Accession Number

Search Result

Facet

List of Entries

1 - 30 entries / Number of founds: 26506 FlatFile

| PrimaryAccessionNumber | Definition | Se |
|---|---|---|
| AF217656 | Definition:Human papillomavirus is Organism:Human papillomavirus | |
| AF217657 | Definition:Human papillomavirus is Organism:Human papillomavirus | |
| AF217659 | Definition:Human papillomavirus is Organism:Human papillomavirus | |
| AF217658 | Definition:Human papillomavirus Organism:Human papillomavirus | |
| AF217660 | Definition:Human papillomavirus Organism:Human papillomavirus | |
| AF217661 | Definition:Human papillomavirus Organism:Human papillomavirus | |
| AF455142 | Definition:Human papillomavirus Organism:Human papillomavirus | |
| AF455144 | Definition:Human papillomavirus Organism:Human papillomavirus | |
| AF455146 | Definition:Human papillomavirus Organism:Human papillomavirus | |

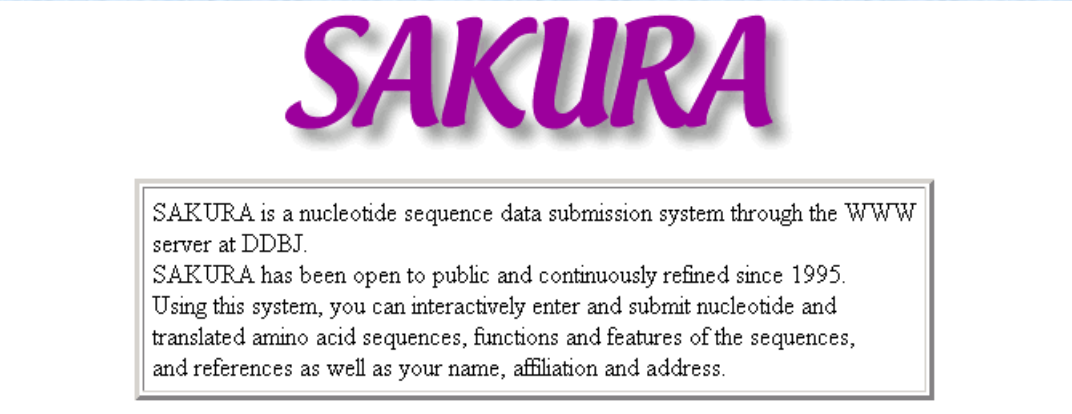
Reference PubmedID

```
LOCUS       AF217656                 434 bp    DNA     linear   VRL 30-NOV-2000
DEFINITION  Human papillomavirus isolate FA14 major capsid protein L1 gene,
            partial cds.
ACCESSION   AF217656
VERSION     AF217656.1
KEYWORDS    .
SOURCE      Human papillomavirus
  ORGANISM  Human papillomavirus
            Viruses; dsDNA viruses, no RNA stage; Papillomaviridae;
            unclassified Papillomaviridae.
REFERENCE   1  (bases 1 to 434)
  AUTHORS   Antonsson,A., Forslund,O., Ekberg,H., Sterner,G. and Hansson,B.G.
  TITLE     The ubiquity and impressive genomic diversity of human skin
```

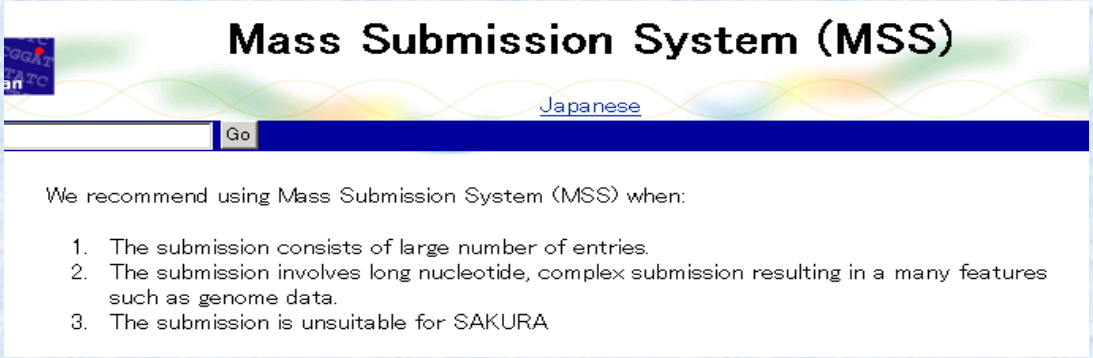
```
                     /translation="NIYNNQGTRLEVPKVSGNQHRVFRLKLPDPNRFALADMSVYNPD
                     KERLVWGLKGIEIGRGQPLGIGSSGHPLFNKVNDTENGNTYRNSSKDDRQNISFDPKQ
                     LQMFIIGCTPCIGEHWDRAPACVNDDQAGRCPPIELINSYIQ"
BASE COUNT          153 a           71 c           93 g          117 t
ORIGIN
        1 tcaatattta taacaatcaa ggcacacgat tggaggttcc taaagtatca ggaaatcaac
       61 acagggtatt tagattaaag ctaccagatc ctaataggtt tgcgttagct gacatgtcag
      121 tatataaccc tgacaaagaa agattagtat ggggtttgaa aggcatagaa ataggcaggg
      181 gccaaccttt aggaataggc agcagtggtc atccactgtt taataaggtt aatgatacag
      241 aaaatggcaa tacatatagg aactcctcta aggatgatag acaaaatatt tcatttgacc
      301 ccaagcagtt gcaaatgttt attattggct gtactccatg tataggagaa cattgggaca
      361 gagcaccagc atgtgttaat gatgatcaag ctggtagatg tcctcctata gagttaataa
      421 actcatatat acag
//
```

# DDBJ Data Submission

- Single sequence
- Multiple sequences
- Updates




Sequence, address, contact details, status of publications.



# Genbank

NCBI Resources How To

GenBank Nucleotide

GenBank Submit Genomes WGS HTGs EST/GSS Metagen

## GenBank Overview

### What is GenBank?

GenBank ® is the NIH genetic sequence database, an annotated collection of all publicly available DNA (*Nucleic Acids Research*, 2013 Jan;41(D1):D36-42). GenBank is part of the International Nucleotide S Collaboration , which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology l and GenBank at NCBI. These three organizations exchange data on a daily basis.

The complete release notes for the current version of GenBank are available on the NCBI ftp site. A ne every two months. GenBank growth statistics for both the traditional GenBank divisions and the WGS from each release.

An example of a GenBank record may be viewed for a *Saccharomyces cerevisiae* gene.

### Access to GenBank

There are several ways to search and retrieve data from GenBank.

- Search GenBank for sequence identifiers and annotations with Entrez Nucleotide , which is divided divisions: CoreNucleotide (the main collection), dbEST (Expressed Sequence Tags), and dbGSS ( Sequences).

**GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences**

***Nucleic Acids Research*, 2013 Jan;41(Database issue):D36-42**

# Genbank: contents

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrenc
  TITLE     Cloning and sequence of REV7, a gene whose fun
            DNA damage-induced mutagenesis in Saccharomyce
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requi
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology,
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
```

```
     gene            complement(3300..4037)
                     /gene="REV7"
     CDS             complement(3300..4037)
                     /gene="REV7"
                     /codon_start=1
                     /product="Rev7p"
                     /protein_id="AAA98667.1"
                     /db_xref="GI:1293616"
                     /translation="MNRWVEKWLRVYLKCYINLILFYRNVYPPQSFDYTTYQSFNLPQ
                     FVPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKKNDLCIEKYVLDFSELQHVD
                     KDDQIITETEVFDEFRSSLNSLIMHLEKLPKVNDDTITFEAVINAIELELGHKLDRNR
                     RVDSLEEKAEIERDSNWVKCQEDENLPDNNGFQPPKIKLTSLVGSDVGPLIIHQFSEK
                     LISGDDKILNGVYSQYEEGESIFGSLF"
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
      121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
      181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg
      241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa
      301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa
      361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat
      421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga
      481 gagtcgccct cctttgtcga gtaattttca cttttcatat gagaacttat tttcttattc
```

# Genbank: contents

The LOCUS field contains locus name, sequence length, molecule type, GenBank division, and modification date

Definition: Brief description of sequence; includes information such as source organism, gene name/protein name, or some description of the sequence's function

Accession: The unique identifier for a sequence record

Version: A nucleotide sequence identification number that represents a single, specific sequence in the GenBank database. GI: GenInfo identifier.

Keywords: Word or phrase describing the sequence

Source: organism name

Reference: Publications by the authors of the sequence that discuss the data reported in the record.

The GenBank database is divided into 18 divisions:

1. PRI - primate sequences
2. ROD - rodent sequences
3. MAM - other mammalian sequences
4. VRT - other vertebrate sequences
5. INV - invertebrate sequences
6. PLN - plant, fungal, and algal sequences
7. BCT - bacterial sequences
8. VRL - viral sequences
9. PHG - bacteriophage sequences
10. SYN - synthetic sequences
11. UNA - unannotated sequences
12. EST - EST sequences (expressed sequence tags)
13. PAT - patent sequences
14. STS - STS sequences (sequence tagged sites)
15. GSS - GSS sequences (genome survey sequences)
16. HTG - HTG sequences (high-throughput genomic sequences)
17. HTC - unfinished high-throughput cDNA sequencing
18. ENV - environmental sampling sequences

# Genbank: contents

Title: TMFunction: database for functional residues in membrane proteins.

Authors: Gromiha MM, Yabuki Y, Suresh MX, Thangakani AM, Suwa M, Fukui K.

Journal: Nucleic Acids Res. 2009 Jan;37(Database issue):D201-4.

PMID: 18842639

Features: Information about genes and gene products, as well as regions of biological significance reported in the sequence.

CDs: Coding sequence; region of nucleotides that corresponds with the sequence of amino acids in a protein (location includes start and stop codons).

**complete** feature is simply written as ***n..m***
Example: 687..3158

< indicates **partial on the 5' end**
Example: <1..206

> indicates **partial on the 3' end**
Example: 4821..5028>

**(complement)** indicates that the feature is on the complementary strand
Example: complement(3300..4037)
The feature extends from base 3300 through base 4037 but is actually on the complementary strand.

Translation: The amino acid translation corresponding to the nucleotide coding sequence

Base count: Frequency of occurrence of the different base types, A, C, G and T in the sequence.

**http://www.ncbi.nlm.nih.gov/genbank/**

# Genbank: search

NCBI **GenBank Overview**

PubMed Entrez BLAST OMIM Books Taxonomy Structure

Search Entrez for homo sapiens GAPD mRNA g Go

NCBI Home
NCBI Site Map
GenBank Submissions Handbook
Submit to GenBank
Submit an update
Search GenBank

**What is GenBank?**

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly availal DNA sequences (*Nucleic Acids Research*, 2011 Jan;39(Database issue):D32-7). There are approximately 126,551,501,141 bases in 135,440,924 sequence records in the traditional GenBank divisions and 191,401,393,188 bases in 62,715,288 sequence records in the WGS division as of April 2011.

The complete release notes for the current version of GenBank are available on the NCBI ftp site. A new release is made every two months. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the

NCBI Resources How To

Nucleotide
Alphabet of Life

Search: Nucleotide Save search Limits Advanced search

homo sapiens GAPD mRNA glyceraldehyde-3-phosphate dehydrogenase Searc

Display Settings: Summary, 20 per page, Sorted by Default order

**Are you looking for gene information?**

GAPDH (GAPD) glyceraldehyde-3-phosphate dehydrogenase [Homo sapiens]

gapd reference sequences Genomic (1) | Transcript (1) | All reference sequences

**Results: 19**

Homo sapiens glyceraldehyde-3-phosphate dehydrogenase (GAPDH), mRNA

NCBI

Entrez, The Life S

HOME SEARCH SITE MAP PubMed All Databases Human Genome

Search across databases homo sapiens GAPD mRNA glyceraldeh

- Result counts displayed in gray indicate one or more terms not found

656 **PubMed:** biomedical literature citations and abstracts
245 **PubMed Central:** free, full text journal articles
1 **Site Search:** NCBI web and FTP sites
none Bo
none ON

19 **Nucleotide:** Core subset of nucleotide sequence records
none **EST:** Expressed Sequence Tag records
none **GSS:** Genome Survey Sequence records
3 **Protein:** sequence database
none **Genome:** whole genome sequences
none **Structure:** three-dimensional macromolecular structures
none **Taxonomy:** organisms in GenBank
none dbGa
2 UniGe
none CDD:
none UniST
none PopS
none GEO
2 GEO

NCBI Resources How To

Nucleotide
Alphabet of Life

Search: Nucleotide Limits Advanced search Help

Search Clear

Display Settings: GenBank Send:

**Human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) mRNA, complete cds**

GenBank: M33197.1

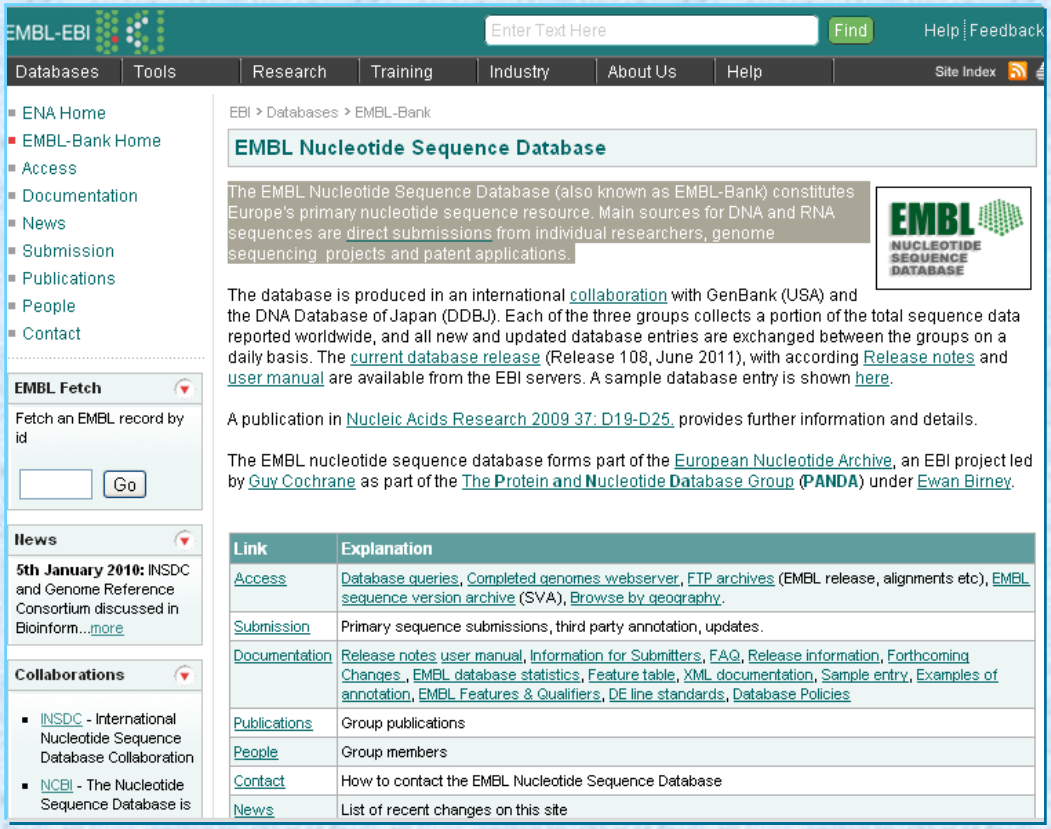
FASTA Graphics

Go to:

```
LOCUS       HUMGAPDH                 1268 bp    mRNA    linear   PRI 08-NOV-1994
DEFINITION  Human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) mRNA,
            complete cds.
ACCESSION   M33197
VERSION     M33197.1  GI:182976
KEYWORDS    glyceraldehyde-3-phosphate dehydrogenase.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1268)
  AUTHORS   Tokunaga,K., Nakamura,Y., Sakata,K., Fujimori,K., Ohkubo,M.,
            Sawada,K. and Sakiyama,S.
  TITLE     Enhanced expression of a glyceraldehyde-3-phosphate dehydrogenase
            gene in human lung cancers
  JOURNAL   Cancer Res. 47 (21), 5616-5619 (1987)
   PUBMED   3664468
COMMENT     Original source text: Human lung cancer cell, cDNA to mRNA.
FEATURES             Location/Qualifiers
     source          1..1268
                     /organism="Homo sapiens"
```

# EMBL: Nucleotide sequence database

**The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications.**

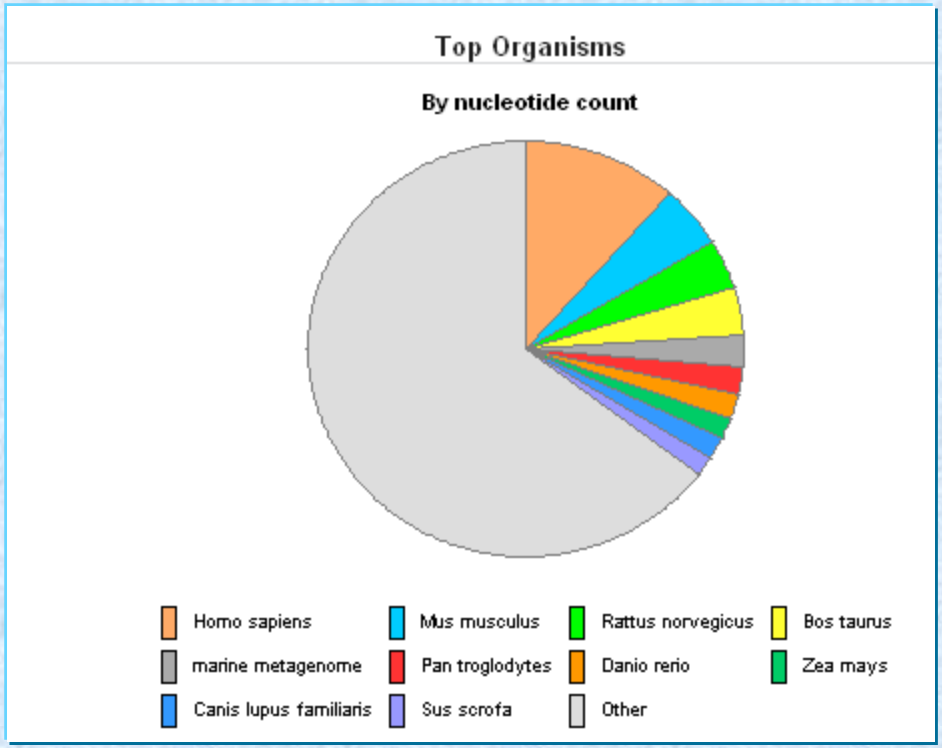


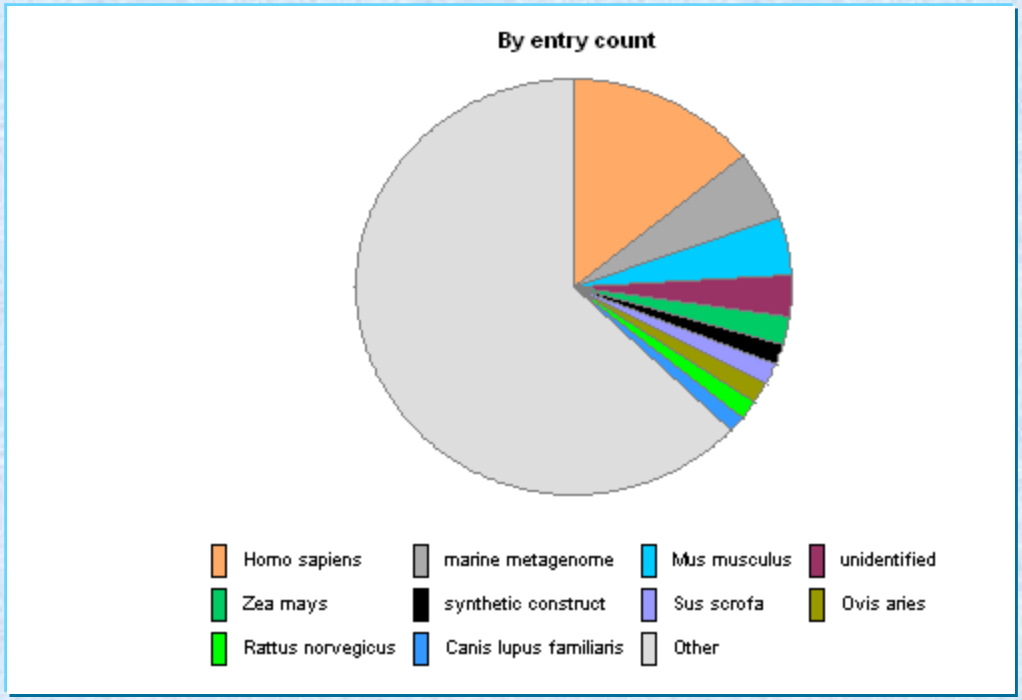

**http://www.ebi.ac.uk/embl/**

# EMBL: Statistics

Release on 16-Jan-2014.

The release contains 798.2 million sequence entries comprising 2004.9 billion nucleotides.



# Database entry

EMBL-EBI

Enter

Databases | Tools | Research | Training | Indust

EBI > Databases > Database Browsing > Dbfetch > EMBL-Bank: TRBG361

## EBI Dbfetch

```
ID   X56734; SV 1; linear; mRNA; STD; PLN; 1859 BP.
XX
AC   X56734; S46826;
XX
DT   12-SEP-1991 (Rel. 29, Created)
DT   25-NOV-2005 (Rel. 85, Last updated, Version 11)
XX
DE   Trifolium repens mRNA for non-cyanogenic beta-glu
XX
KW   beta-glucosidase.
XX
OS   Trifolium repens (white clover)
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophy
OC   Spermatophyta; Magnoliophyta; eudicotyledons; cor
OC   fabids; Fabales; Fabaceae; Papilionoideae; Trifol
XX
RN   [5]
RP   1-1859
RX   DOI; 10.1007/BF00039495
RX   PUBMED; 1907511.
RA   Oxtoby E., Dunn M.A., Pancoro A., Hughes M.A.;
RT   "Nucleotide and derived amino acid sequence of th
RT   beta-glucosidase (linamarase) from white clover (
RL   Plant Mol. Biol. 17(2):209-219(1991).
XX
RN   [6]
RP   1-1859
RA   Hughes M.A.;
RT   ;
RL   Submitted (19-NOV-1990) to the INSDC.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..1859
FT                   /organism="Trifolium repens"
FT                   /mol_type="mRNA"
FT                   /clone_lib="lambda gt10"
FT                   /clone="TRE361"
FT                   /tissue_type="leaves"
FT                   /db_xref="taxon:3899"
FT   CDS             14..1495
FT                   /product="beta-glucosidase"
FT                   /EC_number="3.2.1.21"
FT                   /note="non-cyanogenic"
FT                   /db_xref="GOA:P26204"
FT                   /db_xref="InterPro:IPR001360"
FT                   /db_xref="InterPro:IPR013781"
FT                   /db_xref="InterPro:IPR017853"
FT                   /db_xref="InterPro:IPR018120"
FT                   /db_xref="UniProtKB/Swiss-Prot:P26204"
FT                   /protein_id="CAA40058.1"
FT                   /translation="MDFIVAIFALFVISSFTITSTNAVEASTLLDIGNLSRSSFPRGFI
FT                   FGAGSSAYQFEGAVNEGGRGPSIWDTFTHKYPEKIRDGSNADITVDQYHRYKEDVGIMK
FT                   DQNMDSYRFSISWPRILPKGKLSGGINHEGIKYYNNLINELLANGIQPFVTLFHWDLPQ
FT                   VLEDEYGGFLNSGVINDFRDYTDLCFKEFGDRVRYWSTLNEPWVFSNSGYALGTNAPGR
FT                   CSASNVAKPGDSGTGPYIVTHNQILAHAEAVHVYKTKYQAYQKGKIGITLVSNWLMPLD
FT                   DNSIPDIKAAERSLDFQFGLFMEQLTTGDYSKSMRRIVKNRLPKFSKFESSLVNGSFDF
FT                   IGINYYSSSYISNAPSHGNAKPSYSTNPMTNISFEKHGIPLGPRAASIWIYVYPYMFIQ
FT                   EDFEIFCYILKINITILQFSITENGMNEFNDATLPVEEALLNTYRIDYYYRHLYYIRSA
FT                   IRAGSNVKGFYAWSFLDCNEWFAGFTVRFGLNFVD"
FT   mRNA            1..1859
FT                   /experiment="experimental evidence, no additional details
FT                   recorded"
XX
SQ   Sequence 1859 BP; 609 A; 314 C; 355 G; 581 T; 0 other;
     aaacaaacca aatatggatt ttattgtagc catatttgct ctgtttgtta ttagctcatt        60
     cacaattact tccacaaatg cagttgaagc ttctactctt cttgacatag gtaacctgag       120
     tcggagcagt tttcctcgtg gcttcatctt tggtgctgga tcttcagcat accaatttga       180
     aggtgcagta aacgaaggcg gtagaggacc aagtatttgg gataccttca cccataaata       240
```

# EMBL: Search

```
ID   M33197; SV 1; linear; mRNA; STD; H
XX
AC   M33197;
XX
DT   10-JUL-1990 (Rel. 24, Created)
DT   17-APR-2005 (Rel. 83, Last updated,
XX
DE   Human glyceraldehyde-3-phosphate d
XX
KW   glyceraldehyde-3-phosphate dehydro
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Cran
OC   Eutheria; Euarchontoglires; Primat
OC   Homo.
XX
RN   [1]
RP   1-1268
RX   PUBMED; 3664468.
RA   Tokunaga K., Nakamura Y., Sakata K
RA   Sakiyama S.;
RT   "Enhanced expression of a glycerald
RT   human lung cancers";
RL   Cancer Res. 47(21):5616-5619(1987)
XX
DR   Ensembl-Gn; ENSG00000111640; Homo_
DR   Ensembl-Tr; ENST00000229239; Homo_
DR   Ensembl-Tr; ENST00000396859; Homo_
DR   Ensembl-Tr; ENST00000396861; Homo_
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..1268
FT                   /organism="Homo sa
FT                   /map="12p13"
FT                   /mol_type="mRNA"
```

```
FT   mRNA            <1..1268
FT                   /note="GAPDH mRNA"
FT   CDS             61..1068
FT                   /codon_start=1
FT                   /gene="GAPD"
FT                   /note="glyceraldehyde-3-phosphate dehydrogenase (EC
FT                   1.2.1.12)"
FT                   /db_xref="GDB:119249"
FT                   /db_xref="GOA:P04406"
FT                   /db_xref="H-InvDB:HIT000195501"
FT                   /db_xref="HGNC:4141"
FT                   /db_xref="InterPro:IPR006424"
FT                   /db_xref="InterPro:IPR016040"
FT                   /db_xref="InterPro:IPR020828"
FT                   /db_xref="InterPro:IPR020829"
FT                   /db_xref="InterPro:IPR020830"
FT                   /db_xref="InterPro:IPR020831"
FT                   /db_xref="PDB:1U8F"
FT                   /db_xref="PDB:1ZNQ"
FT                   /db_xref="PDB:2FEH"
FT                   /db_xref="PDB:3GPD"
FT                   /db_xref="UniProtKB/Swiss-Prot:P04406"
FT                   /protein_id="AAA52518.1"
FT                   /translation="MGKVKVGVNGFGRIGRLVTRAAFNSGKVDIVAINDPFIDLNYMVY
FT                   MFQYDSTHGKFHGTVKAENGKLVINGNPITIFQERDPSKIKWGDAGAEYVVESTGVFTT
FT                   MEKAGAHLQGGAKRVIISAPSADAPMFVMGVNHEKYDNSLKIISNASCTTNCLAPLAKV
FT                   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGALQNIIPASTGAAKAVGKVIP
FT                   ELNGKLTGMAFRVPTANVSVVDLTCRLEKPAKYDDIKKVVKQASEGPLKGILGYTEHQV
FT                   VSSDFNSDTHSSTFDAGAGIALNDHFVKLISWYDNEFGYSNRVVDLMAHMASKE"
XX
SQ   Sequence 1268 BP; 295 A; 385 C; 326 G; 262 T; 0 other;
     gttcgacagt cagccgcatc ttcttttgcg tcgccagccg agccacatcg ctcagacacc      60
     atggggaagg tgaaggtcgg agtcaacgga tttggtcgta ttgggcgcct ggtcaccagg     120
     gctgctttta actctggtaa agtggatatt gttgccatca atgacccctt cattgacctc     180
     aactacatgg tttacatgtt ccaatatgat tccacccatg gcaaattcca tggcaccgtc     240
     aaggctgaga acgggaagct tgtcatcaat ggaaatccca tcaccatctt ccaggagcga     300
```

# Literature database

Chemical abstracts

Physical abstracts

Science citation index

PUBMED

Mainly for biological sciences and are freely available

Useful for getting the references of any work and related papers

# PUBMED

## Search with the keyword "DDBJ Nucleic acid sequence database"

www.ncbi.nlm.nih.gov/pubmed?term=DDBJ%20Nucleic%20acid%20sequence%20database

PubMed.gov
U.S. National Library of Medicine
National Institutes of Health

Search: PubMed | RSS | Save search | Limits | Advanced search | Help

DDBJ Nucleic acid sequence database | Search | Clear

Display Settings: Summary, 20 per page, Sorted by Recently Added | Send to:

Display Settings: Summary, 20 per page, Sorted by Recently Added

| Format | Items per page | Sort by |
|---|---|---|
| Summary | 5 | Recently Added |
| Summary (text) | 10 | Pub Date |
| Abstract | 20 | First Author |
| Abstract (text) | 50 | Last Author |
| MEDLINE | 100 | Journal |
| XML | 200 | Title |
| PMID List | | |

Apply

Send to: Filter your resul

Choose Destination

| | |
|---|---|
| File | Clipboard |
| Collections | E-mail |
| Order | My Bibliography |

The sequence read archive.

3. Leinonen R, Sugawara H, Shumway M; International Nucleotide Sequence Database Collaboration.
Nucleic Acids Res. 2011 Jan;39(Database issue):D19-21. Epub 2010 Nov 9.
PMID: 21062823 [PubMed - indexed for MEDLINE] Free PMC Article
Free full text Related citations

DDBJ progress report.

4. Kaminuma E, Kosuge T, Kodama Y, Aono H, Mashima J, Gojobori T, Sugawara H, Ogasawara O, Takagi T, Okubo K, Nakamura Y.
Nucleic Acids Res. 2011 Jan;39(Database issue):D22-7. Epub 2010 Nov 9.
PMID: 21062814 [PubMed - indexed for MEDLINE] Free PMC Article
Free full text Related citations

http://www.ncbi.nlm.nih.gov/pubmed/

# Abstract

NCBI Resources How To My NCBI Sign In

PubMed.gov
U.S. National Library of Medicine
National Institutes of Health

Search: PubMed Limits Advanced search Help

Search Clear

Display Settings: Abstract Send to:

OPEN ACCESS OXFORD JOURNALS FREE full text article in PubMed Central

Nucleic Acids Res. 2011 Jan;39(Database issue):D22-7. Epub 2010 Nov 9.

## DDBJ progress report.

Kaminuma E, Kosuge T, Kodama Y, Aono H, Mashima J, Gojobori T, Sugawara H, Ogasawara O, Takagi T, Okubo K, Nakamura Y.
Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization for Information and Systems, Yata, Mishima 411-8510, Japan.

**Abstract**

The DNA Data Bank of Japan (DDBJ, http://www.ddbj.nig.ac.jp) provides a nucleotide sequence archive database and accompanying database tools for sequence submission, entry retrieval and annotation analysis. The DDBJ collected and released 3,637,446 entries/2,272,231,889 bases between July 2009 and June 2010. A highlight of the released data was archive datasets from next-generation sequencing reads of Japanese rice cultivar, Koshihikari submitted by the National Institute of Agrobiological Sciences. In this period, we started a new archive for quantitative genomics data, the DDBJ Omics aRchive (DOR). The DOR stores quantitative data both from the microarray and high-throughput new sequencing platforms. Moreover, we improved the content of the DDBJ patent sequence, released a new submission tool of the DDBJ Sequence Read Archive (DRA) which archives massive raw sequencing reads, and enhanced a cloud computing-based analytical system from sequencing reads, the DDBJ Read Annotation Pipeline. In this article, we describe these new functions of the DDBJ databases and support tools.

PMID: 21062814 [PubMed - indexed for MEDLINE] PMCID: PMC3013661 Free PMC Article

**Related citations**

DDBJ launches a new archive database with analytical tools for nex [Nucleic Acids Res. 2010]

Biological databases at DNA Data Bank of Japan in the era of next- [Adv Exp Med Biol. 2010]

The sequence read archive. [Nucleic Acids Res. 2011]

Review The EMBL Nucleotide Sequence Database. Contributing ar [Mol Biotechnol. 1999]

Review From mapping to sequencing, post-sequencing and beyon [Plant Cell Physiol. 2005]

See reviews...

See all...

## What is Open Access?

# Full text

PubMed Central
Search | Journal List

Nucleic Acids Research

Journal List > Nucleic Acids Res > v.39(Database issue); Jan 2011

Formats: Abstract | Full Text | PDF (3.6M)

Nucleic Acids Res. 2011 January; 39(Database issue): D22–D27. PMCID: PMC3013661
Published online 2010 November 8. doi: 10.1093/nar/gkq1041.

Copyright © The Author(s) 2010. Published by Oxford University Press.

## DDBJ progress report

Eli Kaminuma, Takehide Kosuge, Yuichi Kodama, Hideo Aono, Jun Mashima, Takashi Gojobori, Hideaki Sugawara, Osamu Ogasawara, Toshihisa Takagi, Kousaku Okubo, and Yasukazu Nakamura*

Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization for Information and Systems, Yata, Mishima, 411-8510, Japan
*To whom correspondence should be addressed. Tel: Phone: +81 55 981 6859; Fax: +81 55 981 6889; Email: yanakamu@genes.nig.ac.jp

Received September 27, 2010; Revised October 7, 2010; Accepted October 11, 2010.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

### ABSTRACT

Other Sections

The DNA Data Bank of Japan (DDBJ, http://www.ddbj.nig.ac.jp) provides a nucleotide sequence archive database and accompanying database tools for sequence submission, entry retrieval and annotation analysis. The DDBJ collected and released 3 637 446 entries/2 272 231 889 bases between July 2009 and June 2010. A highlight of the released data was archive datasets from next-generation sequencing

Nucleic Acids Res

Nucle

### PubMed articles by these authors

- Kaminuma, E.
- Kosuge, T.
- Kodama, Y.
- Nakamura, Y.

### PubMed related articles

- DDBJ launches a new archive database with analytical tools for next-generation sequence data. [Nucleic Acids Res. 2010]
- Biological databases at DNA Data Bank of Japan in the era of next-generation sequencing technolo [Adv Exp Med Biol. 2010]
- The sequence read archive. [Nucleic Acids Res. 2011]
- Review The EMBL Nucleotide Sequence Database. Contributing and accessing data. [Mol Biotechnol. 1999]
- Review From mapping to sequencing, post-sequencing and beyond. [Plant Cell Physiol. 2005]

» See reviews... | » See all...

### Recent Activity

Turn Off Clear

DDBJ progress report

# Related articles

NCBI Resources How To My NCBI Sign In

PubMed.gov
U.S. National Library of Medicine
National Institutes of Health

Search: PubMed Limits Advanced search Help

Search Clear

Display Settings: Summary, 20 per page, Sorted by Link Ranking Send to:

**Results: 1 to 20 of 106**

<< First < Prev Page 1 of 6 Next > Last >>

1. DDBJ progress report.
   Kaminuma E, Kosuge T, Kodama Y, Aono H, Mashima J, Gojobori T, Sugawara H, Ogasawara O, Takagi T, Okubo K, Nakamura Y.
   Nucleic Acids Res. 2011 Jan;39(Database issue):D22-7. Epub 2010 Nov 9.
   PMID: 21062814 [PubMed - indexed for MEDLINE] **Free PMC Article**
   Free full text Related citations

2. DDBJ launches a new archive database with analytical tools for next-generation sequence data.
   Kaminuma E, Mashima J, Kodama Y, Gojobori T, Ogasawara O, Okubo K, Takagi T, Nakamura Y.
   Nucleic Acids Res. 2010 Jan;38(Database issue):D33-8. Epub 2009 Oct 22.
   PMID: 19850725 [PubMed - indexed for MEDLINE] **Free PMC Article**
   Free full text Related citations

3. Biological databases at DNA Data Bank of Japan in the era of next-generation sequencing technologies.
   Kodama Y, Kaminuma E, Saruhashi S, Ikeo K, Sugawara H, Tateno Y, Nakamura Y.
   Adv Exp Med Biol. 2010;680:125-35.
   PMID: 20865494 [PubMed - indexed for MEDLINE]
   Related citations

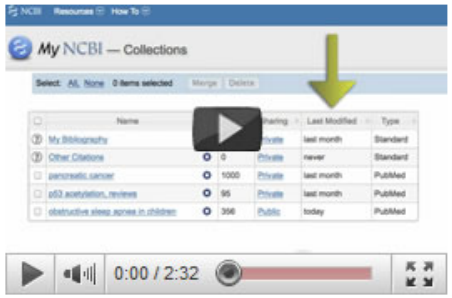
4. The sequence read archive.
   Leinonen R, Sugawara H, Shumway M; International Nucleotide Sequence Database Collaboration.
   Nucleic Acids Res. 2011 Jan;39(Database issue):D19-21. Epub 2010 Nov 9.

**Filter your results:**

All (106)
Free Full Text (97)
Review (3)

Manage Filters

**Save Results in Collections Tutorial**




See larger video at YouTube
See all NCBI YouTube video channel videos

# PUBMED: Features

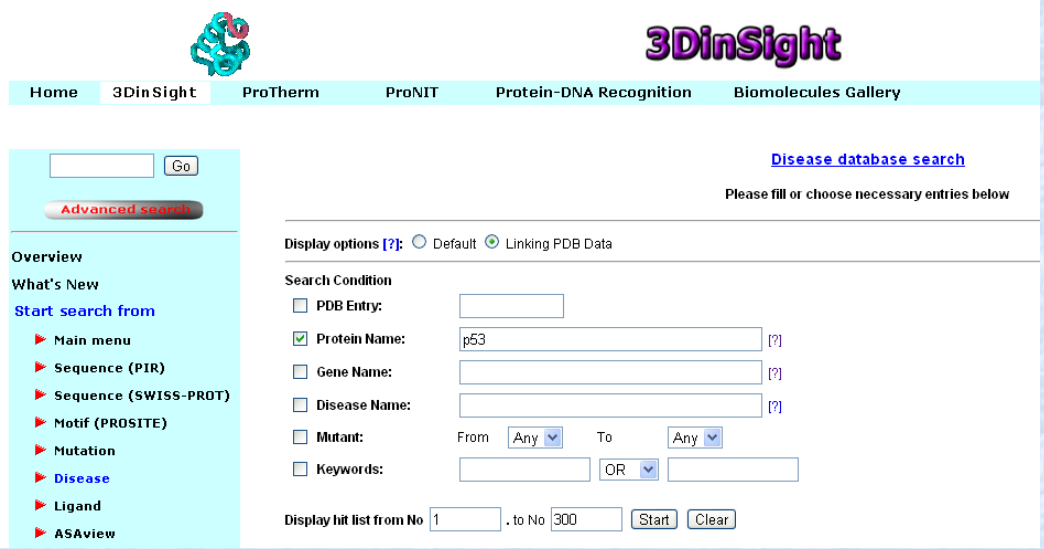
**Search with keywords display the results from the latest article.**

**Search with "related entries" show the output from the most relevant article.**

**One can also use the mesh terms to get a specific article. E.g. Nucleic Acids Res. 34, D204-D206 (2006)**

**204-206[PAGE] AND 34[VOL] AND 2006[PDAT]**

# Disease database



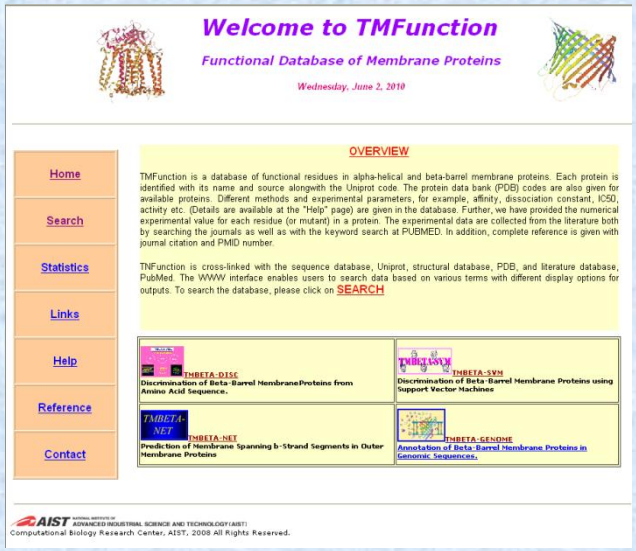

Alterations in proteins cause cancer.

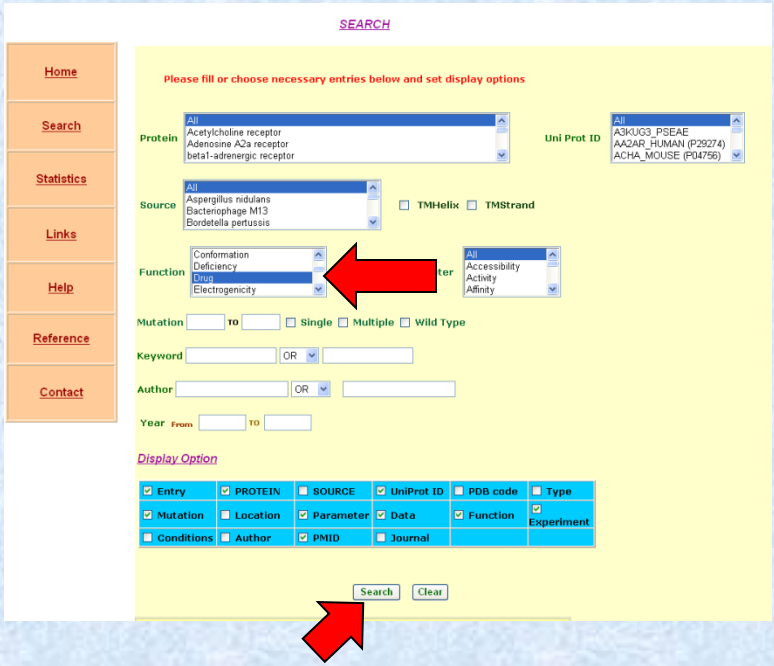
Dr. Akinoi Sarai's group in Japan

**Protein** P53

| PDBENTRY | SOURCE | MUTATION | PROTEIN | GENE_NAME | DISEASE | DISEASE_DB | SWISS |
|---|---|---|---|---|---|---|---|
| 1TUP | Homo sapiens | Gly 245 Arg | TUMOR SUPPRESSOR P53 | TP53 | ADENOCARCINOMA | OMIM | P53_HUMAN |
| | Homo sapiens | Pro 152 Leu | TUMOR SUPPRESSOR P53 | TP53 | ADRENOCORTICAL CARCINOMA | OMIM | P53_HUMAN |
| | Homo sapiens | Arg 158 His | TUMOR SUPPRESSOR P53 | TP53 | ADRENOCORTICAL CARCINOMA | OMIM | P53_HUMAN |
| | Homo sapiens | Arg 196 Ter | TUMOR SUPPRESSOR P53 | TP53 | ADRENOCORTICAL CARCINOMA | OMIM | P53_HUMAN |
| | Homo sapiens | Pro 219 Ser | TUMOR SUPPRESSOR P53 | TP53 | ADRENOCORTICAL CARCINOMA | OMIM | P53_HUMAN |
| | Homo sapiens | Asn 235 Asp | TUMOR SUPPRESSOR P53 | TP53 | ADRENOCORTICAL CARCINOMA | OMIM | P53_HUMAN |
| | Homo sapiens | Glu 286 Ala | TUMOR SUPPRESSOR P53 | TP53 | ADRENOCORTICAL CARCINOMA | OMIM | P53_HUMAN |
| | Homo sapiens | Pro 151 Ser | TUMOR SUPPRESSOR P53 | TP53 | ASTROCYTOMA | OMIM | P53_HUMAN |
| | Homo sapiens | Arg 283 His | TUMOR SUPPRESSOR P53 | TP53 | ASTROCYTOMA | OMIM | P53_HUMAN |
| | Homo sapiens | Arg 181 Cys | TUMOR SUPPRESSOR P53 | TP53 | BREAST CANCER | OMIM | P53_HUMAN |
| | Homo sapiens | Arg 181 His | TUMOR SUPPRESSOR P53 | TP53 | BREAST CANCER | OMIM | P53_HUMAN |
| | Homo sapiens | Arg 267 Gln | TUMOR SUPPRESSOR P53 | TP53 | BREAST CANCER | OMIM | P53_HUMAN |
| | Homo sapiens | Pro 278 Leu | TUMOR SUPPRESSOR P53 | TP53 | BREAST CANCER | OMIM | P53_HUMAN |

# Protein function database




## Search Results

## Important residues for protein function

## M.M. Gromiha et al.

## CBRC, AIST, Japan

**Search Conditions**

| | |
|---|---|
| Function | Drug |
| Parameter | All |
| Protein | All |
| Source | All |
| UniProt ID | All |

HIT: 216

| No. | Protein | UniProt ID | Mutation | Parameter | Data | Function | Experiment | PubMed ID |
|---|---|---|---|---|---|---|---|---|
| 187 | Multidrug resistance protein 1; MRP1 | MRP1_HUMAN (P33527) | S1233A | Relative resistance factor to vincristine | 12.5 | Drug resistance | site-directed mutagenesis | 11925441 |
| 188 | Multidrug resistance protein 1; MRP1 | MRP1_HUMAN (P33527) | S1235A | Relative resistance factor to vincristine | 9.3 | Drug resistance | site-directed mutagenesis | 11925441 |
| 189 | Multidrug resistance protein 1; MRP1 | MRP1_HUMAN (P33527) | Y1236F | Relative resistance factor to vincristine | 4.8 | Drug resistance(partial) | site-directed mutagenesis | 11925441 |
| 190 | Multidrug resistance protein 1; MRP1 | MRP1_HUMAN (P33527) | S1237A | Relative resistance factor to vincristine | 13.7 | Drug resistance | site-directed mutagenesis | 11925441 |
| 191 | Multidrug resistance protein 1; MRP1 | MRP1_HUMAN (P33527) | Q1239A | Relative resistance factor to vincristine | 17.3 | Drug resistance | site-directed mutagenesis | 11925441 |
| 192 | Multidrug resistance protein 1; MRP1 | MRP1_HUMAN (P33527) | T1241A | Relative resistance factor to vincristine | 4.4 | Drug resistance | site-directed mutagenesis | 11925441 |
| 193 | Multidrug resistance protein 1; MRP1 | MRP1_HUMAN (P33527) | Y1243F | Relative resistance factor to vincristine | 4.2 | Drug resistance(partial) | site-directed mutagenesis | 11925441 |
| 194 | Multidrug resistance protein 1; MRP1 | MRP1_HUMAN (P33527) | N1245A | Relative resistance factor to vincristine | 29.3 | Drug resistance(partial) | site-directed mutagenesis | 11925441 |
| 195 | Multidrug resistance protein 1; MRP1 | MRP1_HUMAN (P33527) | null | null | null | Drug resistance | site-directed mutagenesis | 11925441 |
| 196 | Multidrug resistance protein 1; MRP1 | MRP1_HUMAN (P33527) | null | null | null | Drug resistance | site-directed mutagenesis | 11925441 |
| 198 | Multidrug resistance protein 1; MRP1 | MRP1_HUMAN (P33527) | S1233A | Relative resistance factor to VP-16 | 13.3 | Drug resistance | site-directed mutagenesis | 11925441 |

**M.M. Gromiha et al. (2009) Nucleic Acids Res. 37, D201-204**

**TMFunction: database for functional residues in membrane proteins**

M. Michael Gromiha[1,*], Yukimitsu Yabuki[1], M. Xavier Suresh[1], A. Mary Thangakani[2], Makiko Suwa[1] and Kazuhiko Fukui[1]

[1]Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), AIST Tokyo Waterfront Bio-IT Research Building, 2-42 Aomi, Koto-ku, Tokyo 135-0064 and [2]Advanced Technology Inc., Tokyo, Japan

# Protein structure and function

***Proteins***

- ***Extremely versatile in their function***

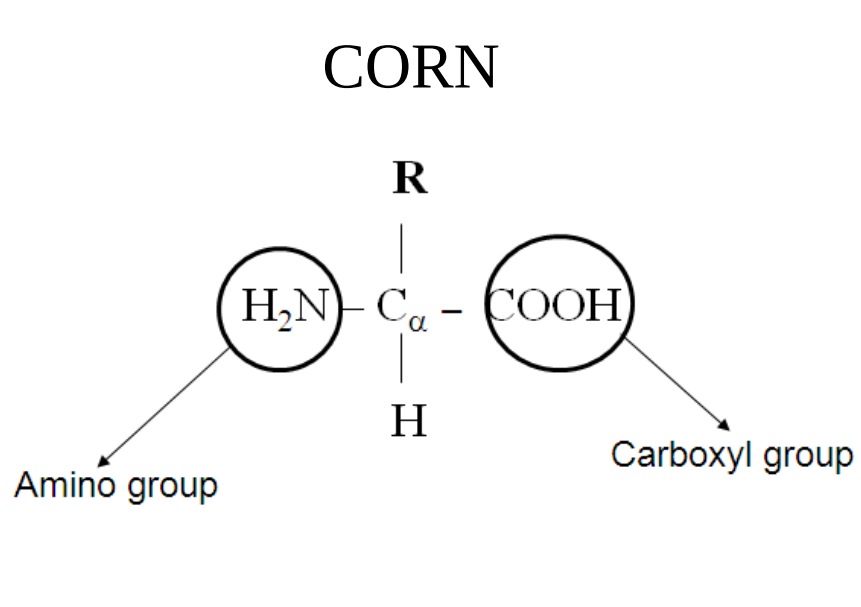
**Nucleic acids**

- **Carry genetic information**

**Carbohydrates**

- **Staple of human diet**
- **Structural and protective elements**

# Building blocks



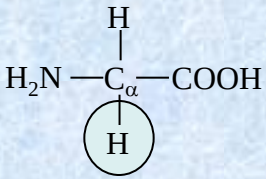
# 3 and 1 letter codes

| | | | | | |
|---|---|---|---|---|---|
| Alanine | Ala | A | Methionine | Met | M |
| Cysteine | Cys | C | Asparagine | Asn | N |
| Aspartic acid | Asp | D | Proline | Pro | P |
| Glutamic acid | Glu | E | Glutamine | Gln | Q |
| Phenylalanine | Phe | F | Arginine | Arg | R |
| Glycine | Gly | G | Serine | Ser | S |
| Histidine | His | H | Threonine | Thr | T |
| Isoleucine | Ile | I | Valine | Val | V |
| Lysine | Lys | K | Tryptophan | Trp | W |
| Leucine | Leu | L | Tyrosine | Tyr | Y |

B, J, O, U, X, Z

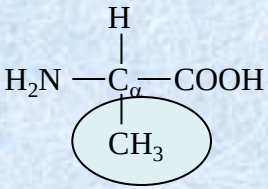
# Hydrophobic residues

***Hydrogen***

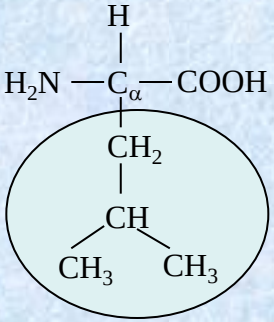



**Glycine**
**Gly, G**

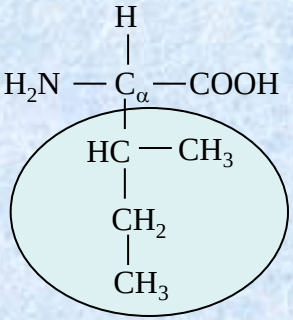
***Aliphatic***



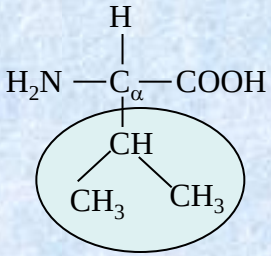

**Alanine**
**Ala, A**




**Leucine**
**Leu, L**

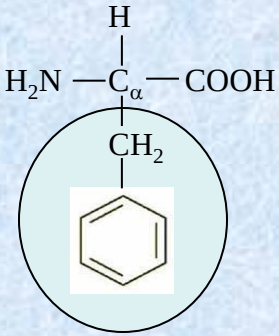



**Isoleucine**
**Ile, I**

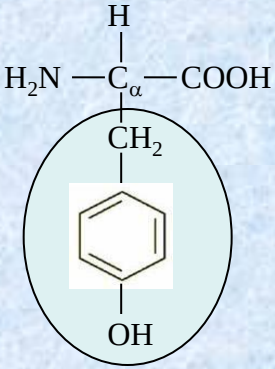



**Valine**
**Val, V**

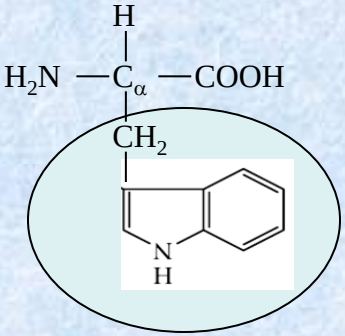
***Aromatic***




**Phenylalanine**
**Phe, F**

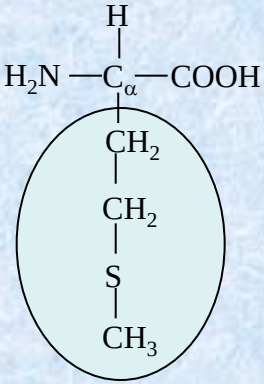



**Tyrosine**
**Tyr, Y**

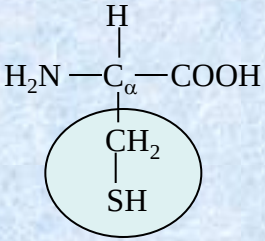



**Tryptophan**
**Trp, W**

***Sulfur containing***



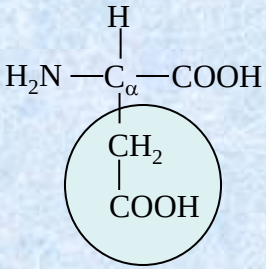

**Methionine**
**Met, M**



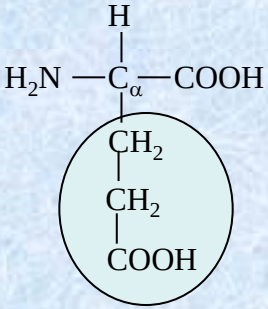

**Cysteine**
**Cys, C**

# Hydrophilic residues

***Negative charged***

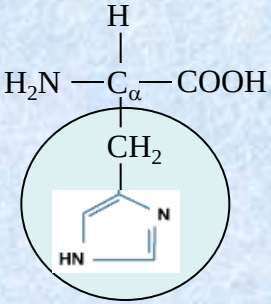



**Aspartic acid**
**Asp, D**

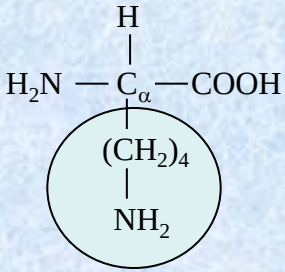



**Glutamic acid**
**Glu, E**

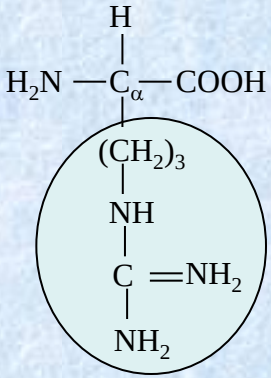
***Positive charged***




**Histidine**
**His, H**

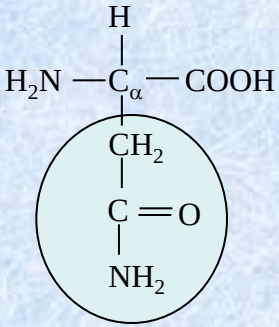



**Lysine**
**Lys, K**

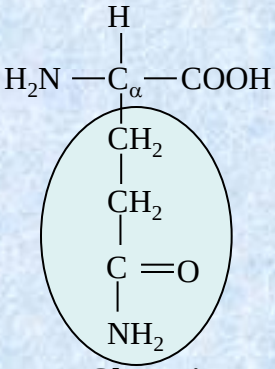



**Arginine**
**Arg, R**

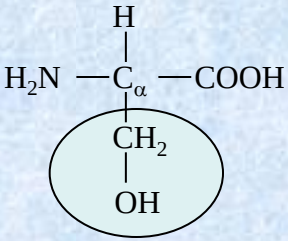
***Polar***



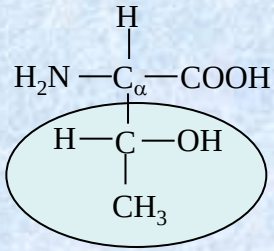

**Asparagine**
**Asn, N**



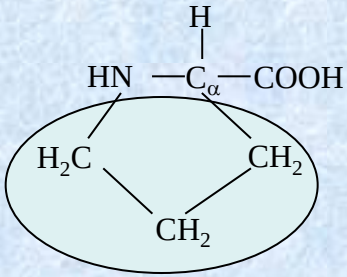

**Glutamine**
**Gln, Q**




**Serine**
**Ser, S**

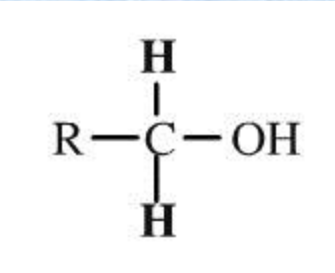



**Threonine**
**Thr, T**

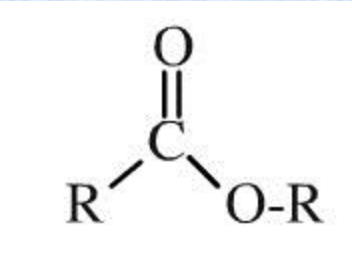



**Proline**
**Pro, P**

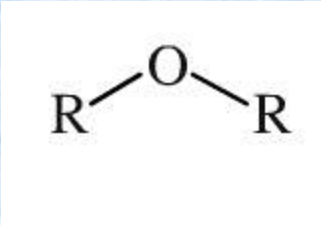
# Chemical groups



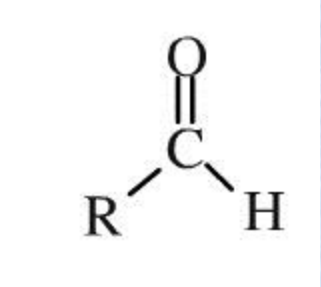

Alcohol



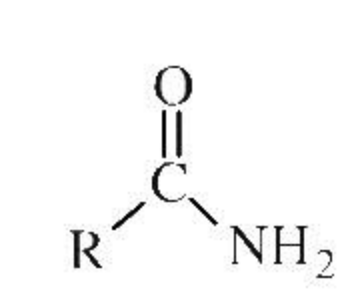

Ester




Ether



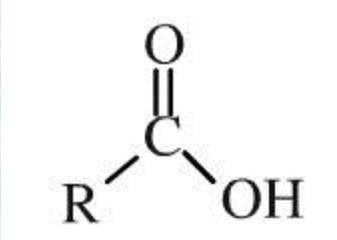

Aldehyde



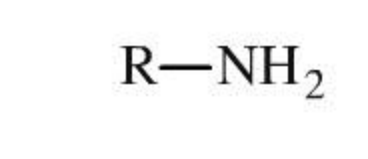

Amide

Alkane

Aliphatic
Aromatic



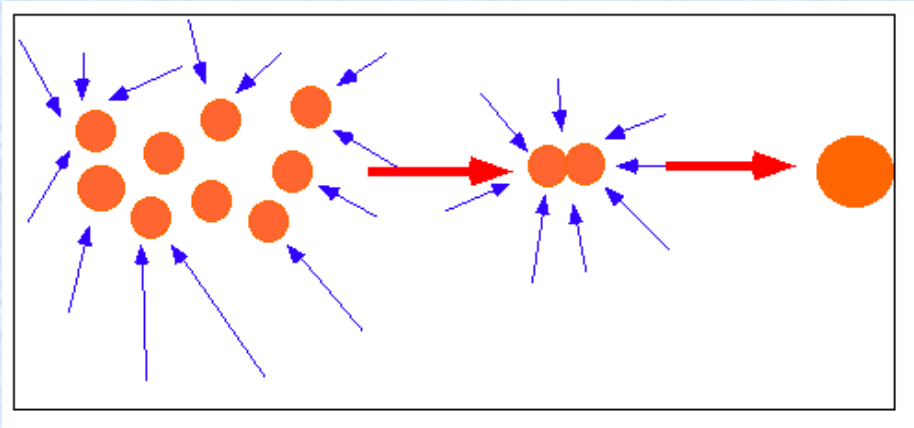

Acid




Amine

# Amino acids: properties

**1. Nonpolar (Hydrophobic; water hating; hydrocarbon alkyl groups (alkane branches) or aromatic (benzene rings)**



**2. Polar (hydrophilic; water loving): functional groups such as acids, amides, alcohols, and amines**

**Amide > Acid > Alcohol > Amine > Ether > Alkane**

**Neutral (glutamine, asparagine)**
**Acidic (glutamic acid, aspartic acid)**
**Basic (lysine, arginine)**

# Polarity

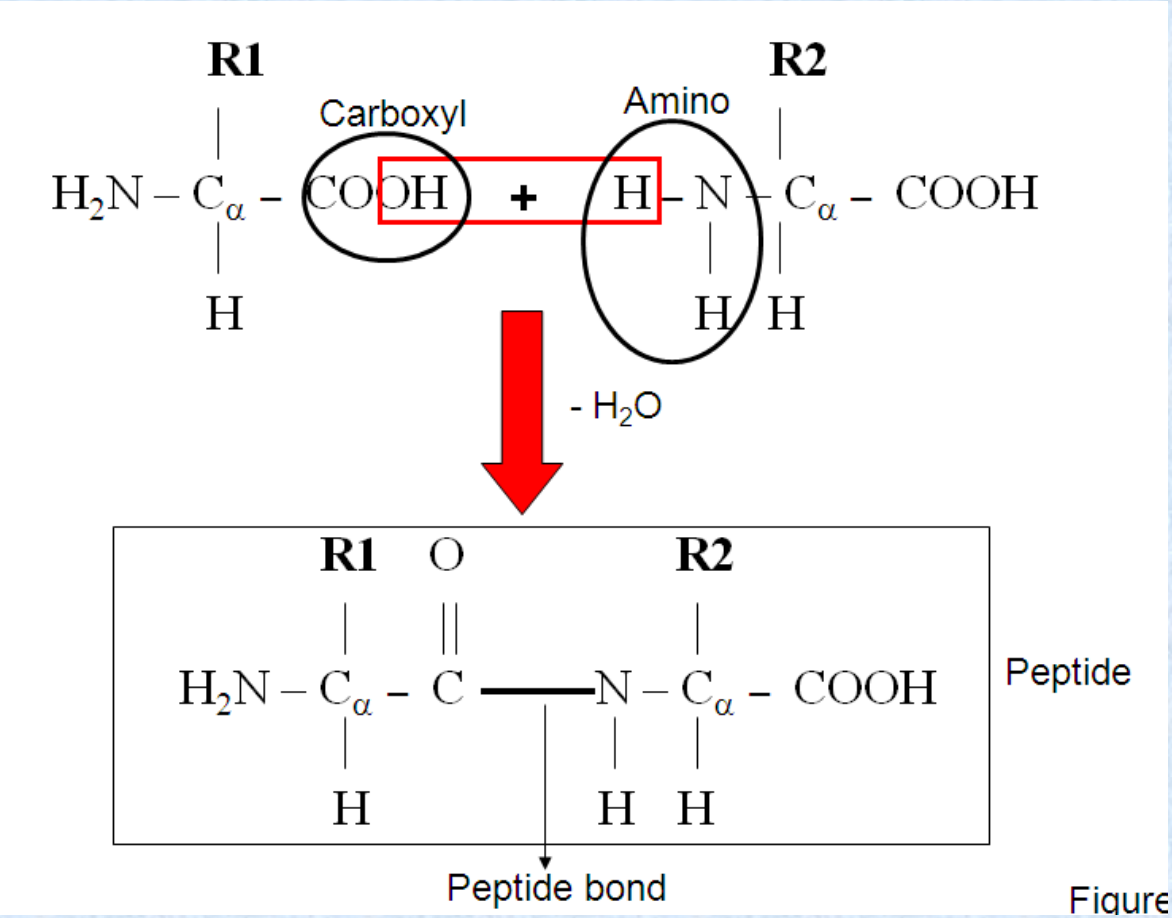
1. Rank the following according to increasing non-polarity i.e. 1 = least non-polar, 4 = most non-polar; leu; phe; val; ala (A, V, L, F)

2. Rank the following amino acids by increasing polarity. i.e.
   1 = more non-polar.
   ser ; glu ; asp ; lys ; ala ; gln (A, K, S, E, D, Q)

3. Which amino acid is most insoluble in water: isoleucine or alanine ? (I)

4. Which amino acid is most soluble in water: lys or ser? (S)

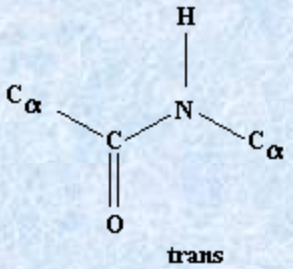
# Peptide bond

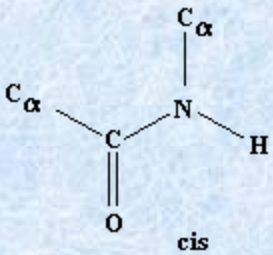



Combination of amino acids

Elimination of water

Peptide bond formation; strong



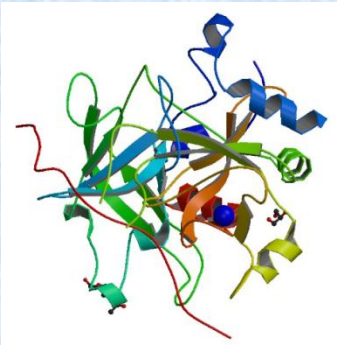



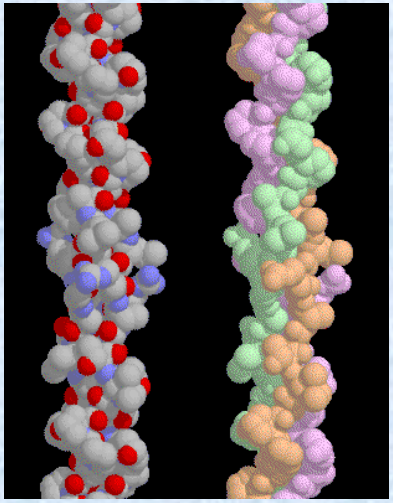
# Proteins

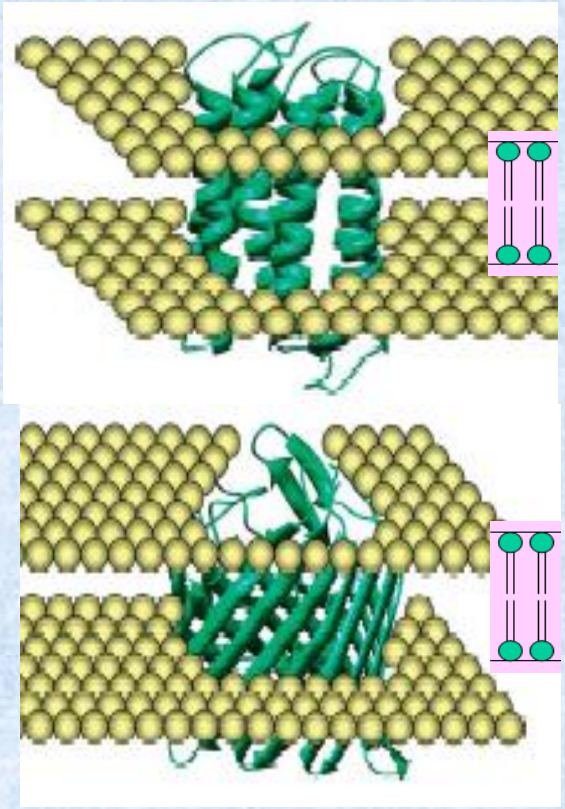
Globular proteins

Fibrous proteins

Membrane proteins







Cytoplasm

Inner membrane

α-helical

Periplasm

Outer membrane

β-barrel (TMβ)

Outer space

# Proteins: Functions

**Enzymes**

**Specialized proteins with catalytic activity (increase the rate of chemical reactions).**

**Only a small portion of the enzyme (3–4 amino acids) is directly involved in catalysis. The region that contains these catalytic residues, binds the substrate, and then carries out the reaction is known as the active site.**

**All the chemical reactions in cell are catalyzed by enzymes.**

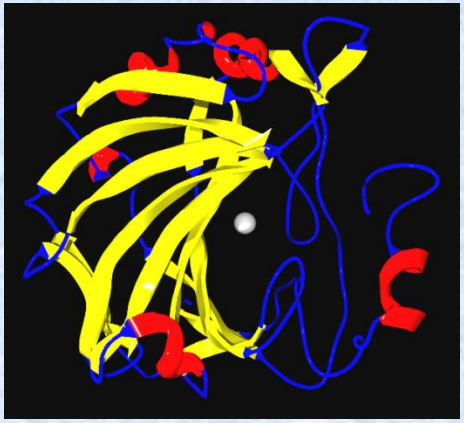
The **carbonic anhydrases** form a family of enzymes that catalyze the rapid interconversion of carbon dioxide and water to bicarbonate.

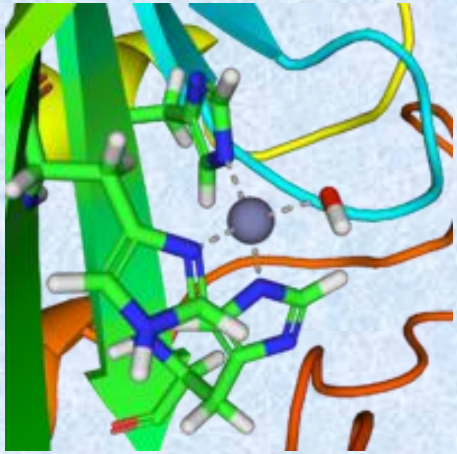
$$CO_2 + H_2O \xrightarrow{\text{Carbonic anhydrase}} H_2CO_3$$

(in tissues; high $CO_2$ concentration)

$$H_2CO_3 \xrightarrow{\text{Carbonic anhydrase}} CO_2 + H_2O$$

(in lungs; low $CO_2$ concentration)





three histidine residues and a hydroxide group coordinating (dashed lines) the zinc on at center.

**The active site of most carbonic anhydrases contains a zinc ion; they are therefore classified as metalloenzymes.**

# Proteins: Functions

In enzymatic reactions, the **molecules** at the beginning of the process are called **substrates**, and they are converted into different molecules, called the **products**

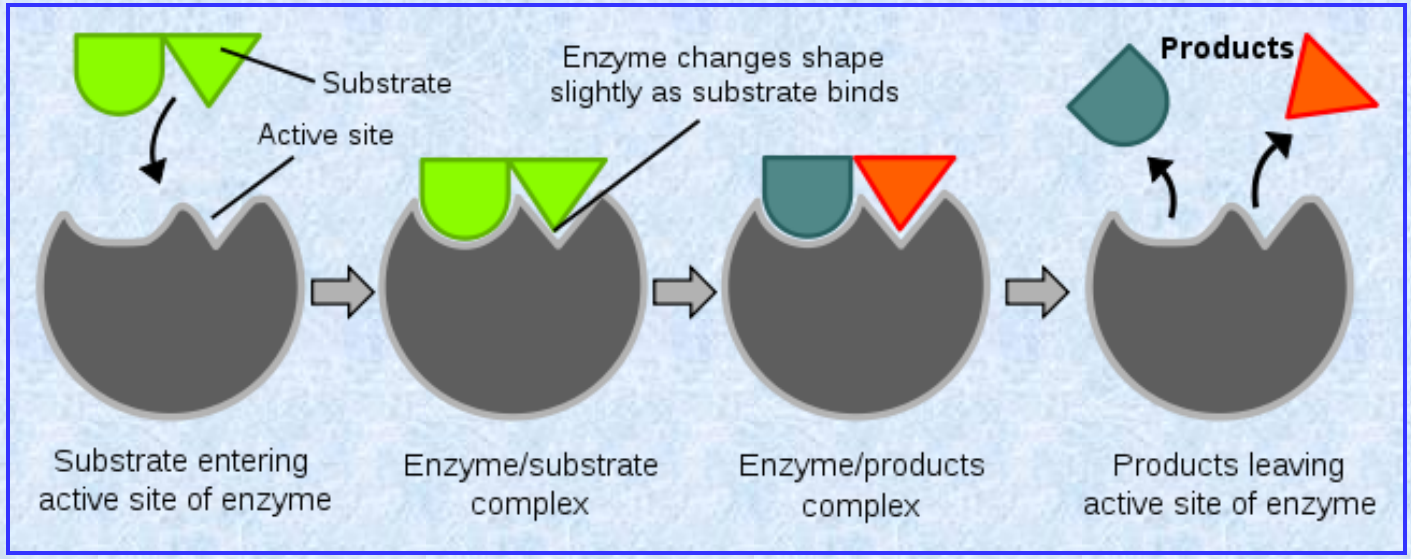
$$CO_2 + H_2O \xrightarrow{\text{Carbonic anhydrase}} H_2CO_3$$

(in tissues; high $CO_2$ concentration)

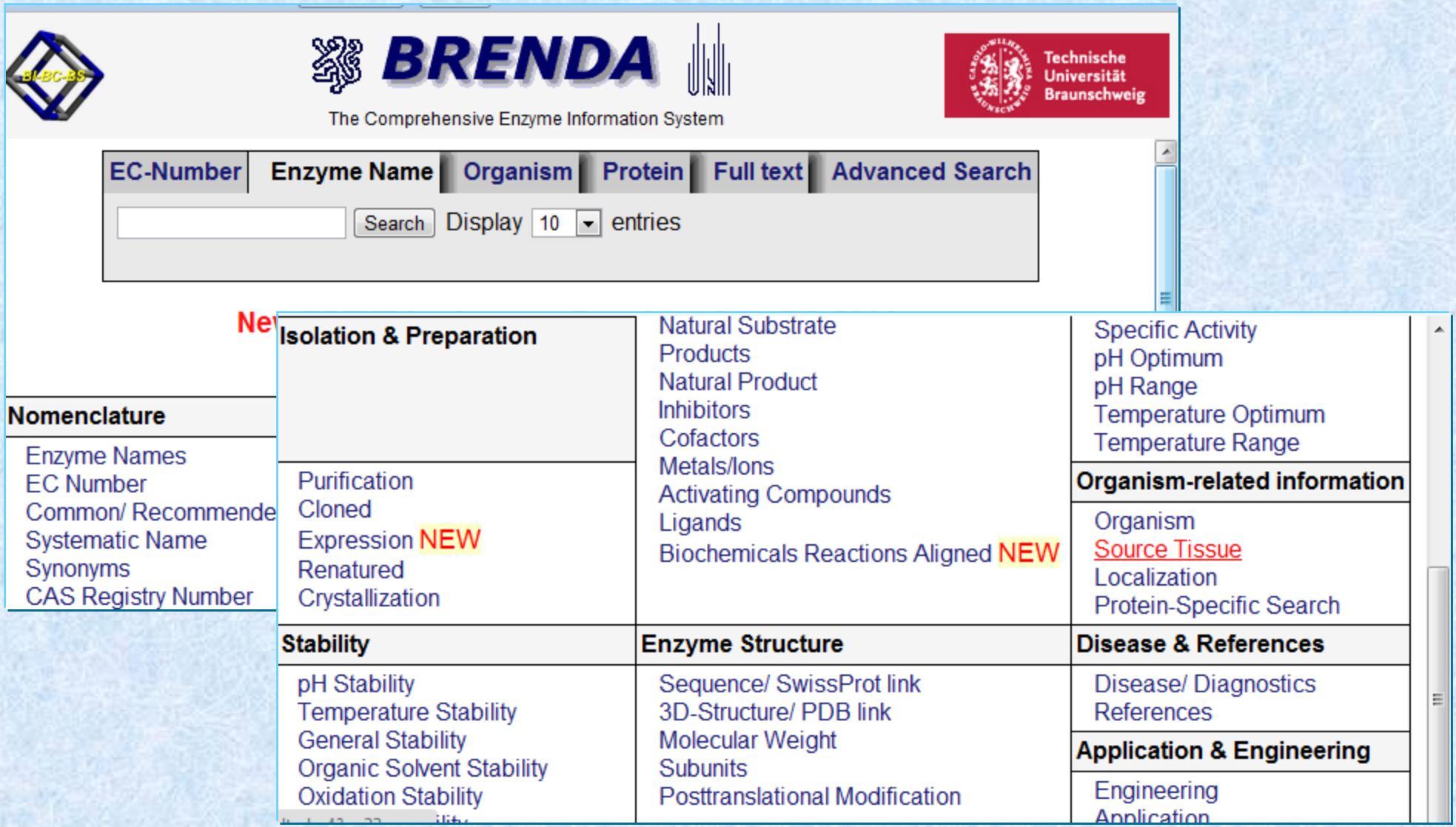
$$H_2CO_3 \xrightarrow{\text{Carbonic anhydrase}} CO_2 + H_2O$$

(in lungs; low $CO_2$ concentration)

## Mechanism




**Alcohol dehydrogenase**: facilitate to break down alcohol (toxic) to others (aldehade/ketone)
**Catalyzing oxidation**




**http://www.brenda-enzymes.org/**

# EC number: format

**Every enzyme code consists of the letters "EC" followed by four numbers separated by periods. Those numbers represent a progressively finer classification of the enzyme.**

- **For example, the tripeptide aminopeptidases have the code "EC 3.4.11.4", whose components indicate the following groups of enzymes:**

***EC 3*** **enzymes are hydrolases (enzymes that use water to break up some other molecule)**

***EC 3.4*** **are hydrolases that act on peptide bonds**

***EC 3.4.11*** **are those hydrolases that cleave off the amino-terminal amino acid from a polypeptide**

***EC 3.4.11.4*** **are those that cleave off the amino-terminal end from a tripeptide**

# Top level codes

**Top-level EC numbers**[4]

| Group | Reaction catalyzed | Typical reaction | Enzyme example(s) with trivial name |
|---|---|---|---|
| EC 1 *Oxidoreductases* | To catalyze oxidation/reduction reactions; transfer of H and O atoms or electrons from one substance to another | AH + B → A + BH (**reduced**)<br>A + O → AO (**oxidized**) | Dehydrogenase, oxidase |
| EC 2 *Transferases* | Transfer of a functional group from one substance to another. The group may be methyl-, acyl-, amino- or phosphate group | AB + C → A + BC | Transaminase, kinase |
| EC 3 *Hydrolases* | Formation of two products from a substrate by hydrolysis | AB + $H_2O$ → AOH + BH | Lipase, amylase, peptidase |
| EC 4 *Lyases* | Non-hydrolytic addition or removal of groups from substrates. C-C, C-N, C-O or C-S bonds may be cleaved | RCOCOOH → RCOH + $CO_2$<br>or [X-A-B-Y] → [A=B + X-Y] | Decarboxylase |
| EC 5 *Isomerases* | Intramolecule rearrangement, i.e. isomerization changes within a single molecule | AB → BA | Isomerase, mutase |
| EC 6 *Ligases* | Join together two molecules by synthesis of new C-O, C-S, C-N or C-C bonds with simultaneous breakdown of ATP | X + Y+ ATP → XY + ADP + Pi | Synthetase |

# BRENDA

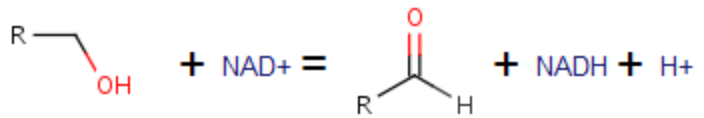
The Comprehensive Enzyme Information System

## EC 1.1.1.1 - alcohol dehydrogenase

| EC NUMBER | COMMENTARY |
|---|---|
| 1.1.1.1 | - |

| RECOMMENDED NAME | GeneOntology No. |
|---|---|
| alcohol dehydrogenase | GO:0004025 |

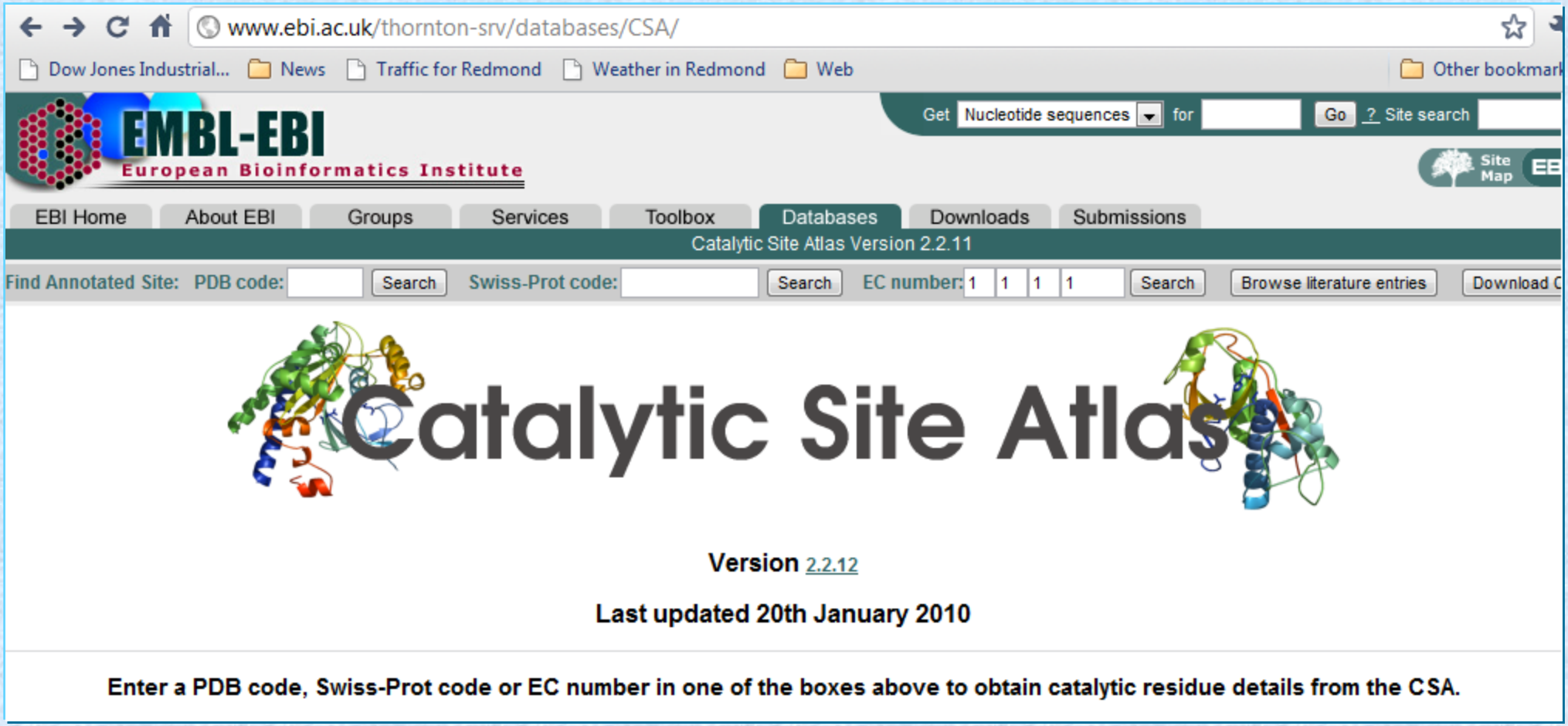
**Reaction catalyzed by alcohol dehydrogenase (1.1.1.1)**




A primary alcohol + NAD+ = an aldehyde + NADH + H+

| REACTION | REACTION DIAGRAM | COMMENTARY | ORGANISM | LITERATURE |
|---|---|---|---|---|
| A primary alcohol + NAD+ = an aldehyde + NADH + H+ | | ordered bi bi mechanism with cofactor adding first to form a binary enzyme complex | Homo sapiens | 285578 |
| A primary alcohol + NAD+ = an aldehyde + NADH + H+ | | Ser48 is involved in catalysis, isozyme gamma(2)gamma(2) | Homo sapiens | 654727 |

# Catalytic site atlas




**http://www.ebi.ac.uk/thornton-srv/databases/CSA/**

## EC: 1.1.1.1

**CSA Entries:**

1a4u, 1a71, 1a72, 1adb, 1adc,
1adf, 1adg, 1adh, 1agn, 1axe,
1axg, 1b14, 1b15, 1b16, 1b21,
1bto, 1cdo, 1d1s, 1d1t, 1dda,
1deh, 1e3e, 1e3i, 1e3l, 1ee2,
1h2b, 1hdx, 1hdy, 1hdz, 1het,
1heu, 1hf3, 1hld, 1hso, 1hsz,
1ht0, 1htb, 1j5r, 1ju9, 1jvb,
1lde, 1ldy, 1llu, 1m6h, 1m6w,
1ma0, 1mc5, 1mg0, 1mg5, 1mgo,
1mp0, 1n8k, 1n92, 1nto, 1nvg,
1o2d, 1otq, 1p1r, 1qlh, 1qlj,
1qv6, 1qv7, 1r37, 1rjw, 1sby,
1teh, 1u3t, 1u3u, 1u3v, 1u3w,
1vj0, 1ye3, 2adh, 2eer, 2fze,
2fzw, 2hcy, 2jhf, 2jhg, 2ohx,
2oxi, 3adh, 3bto, 3hud, 3i4c,
5adh, 6adh, 7adh, 8adh

## CSA entry for 8adh

Homologous Entry

| | | | |
|---|---|---|---|
| Title: | Oxidoreductase(nad(a)-choh(d)) | | |
| Compound: | Apo-liver alcohol dehydrogenase (e.c.1.1.99.8) | | |
| Mutant: | No | | |
| UniProt/Swiss-Prot: | P00327-ADHE_HORSE | EC Class: | 1.1.1.1 |
| Other CSA Entries: | Homologues of 8adh<br>Entries for UniProt/Swiss-Prot: P00327<br>Entries for EC: 1.1.1.1 | Other Databases: | PDB entry: 8adh<br>PDBsum entry: 8adh<br>UniProt/Swiss-Prot: P00327<br>IntEnz entry: 1.1.1.1<br>KEGG entry: 1.1.1.1<br>EzCatDB entry: D00001 |

**Sites:**

☑ Catalytic Site (Get help with this section)

Found by: PsiBLAST alignment on 1guf
1guf has EC code 0... whereas 8adh has EC code 1.1.1.1. The difference in function suggests that the transfer of annotation from 1guf to 8adh may be incorrect.
In addition to these residues, this site requires the cofactor(s) with PDB three-letter code(s) NDP for its catalytic activity.

| Residue | Chain | Number | UniProt number | Functional part |
|---|---|---|---|---|
| LEU | A | 57 | 58 | Sidechain |

☑ Catalytic Site (Get help with this section)

Found by: PsiBLAST alignment on 1qlh
In addition to these residues, this site requires the cofactor(s) with PDB three-letter code(s) ZN, NAD for its catalytic activity.

| Residue | Chain | Number | UniProt number | Functional part |
|---|---|---|---|---|
| SER | A | 48 | 49 | Sidechain |
| HIS | A | 51 | 52 | Sidechain |

Use the check-boxes to select site(s) to view on the 3D structure in RasMol, and press Display Protein

Catalytic residues are indicated in red.

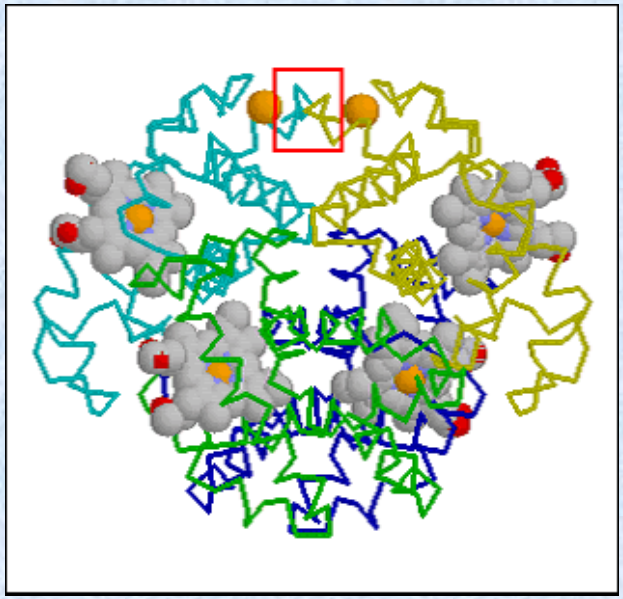
Chain: A

```
        10        20        30        40        50        60        70        80        90       100
         |         |         |         |         |         |         |         |         |         |
STAGKVIKCKAAVLWEEKKPFSIEEVEVAPPKAHEVRIKMVATGICRSDDHVVSGTLVTPLPVIAGHEAAGIVESIGEGVTTVRPGDKVIPLFTPQCGKC

       110       120       130       140       150       160       170       180       190       200
         |         |         |         |         |         |         |         |         |         |
RVCKHPEGNFCLKNDLSMPRGTMQDGTSRFTCRGKPIHHFLGTSTFSQYTVVDEISVAKIDAASPLEKVCLIGCGFSTGYGSAVKVAKVTQGSTCAVFGL

       210       220       230       240       250       260       270       280       290       300
         |         |         |         |         |         |         |         |         |         |
GGVGLSVIMGCKAAGAARIIGVDINKDKFAKAKEVGATECVNPQDYKKPIQEVLTEMSNGGVDFSFEVIGRLDTMVTALSCCQEAYGVSVIVGVPPDSQN

       310       320       330       340       350       360       370
         |         |         |         |         |         |         |
LSMNPMLLLSGRTWKGAIFGGFKSKDSVPKLVADFMAKKFALDPLITHVLPFEKINEGFDLLRSGESIRTILTF
```

# Proteins: Functions

**Transport proteins**

- Binds and carry specific molecules or ions from one organ to another
- Hemoglobin is present in red blood cells, which efficiently carries oxygen from the lungs to the tissues of the body. Hemoglobin also helps in the transportation of carbon dioxide and hydrogen ions back to the lungs.



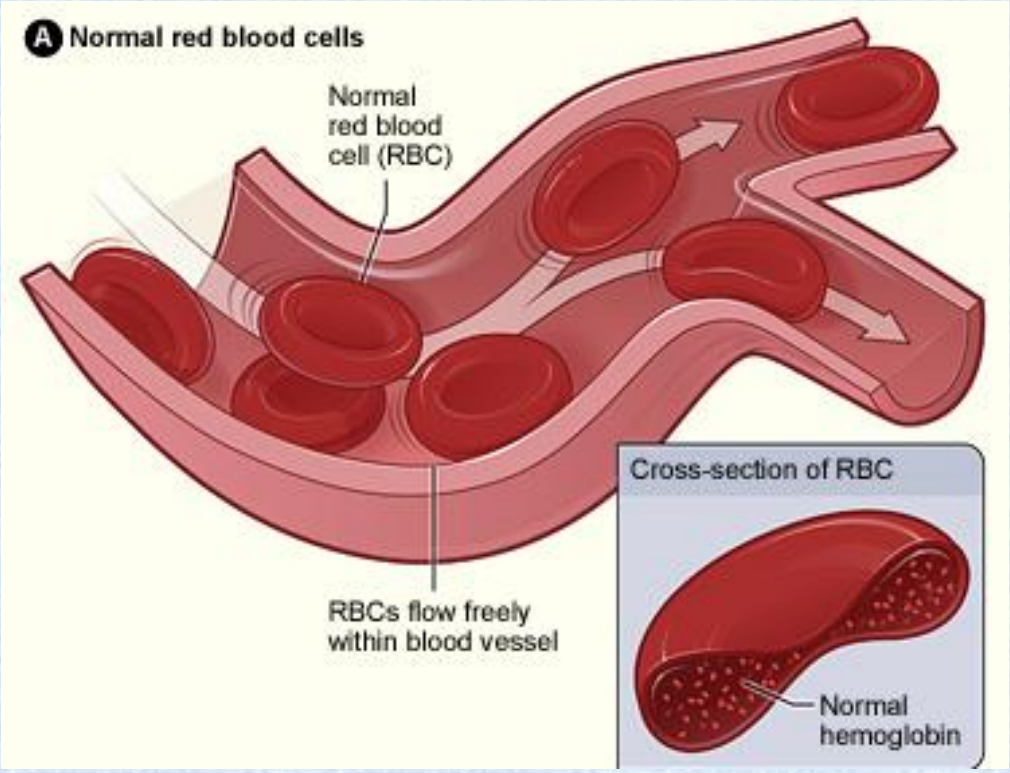
**HBB Sequence in Normal Adult Hemoglobin (Hb A):**

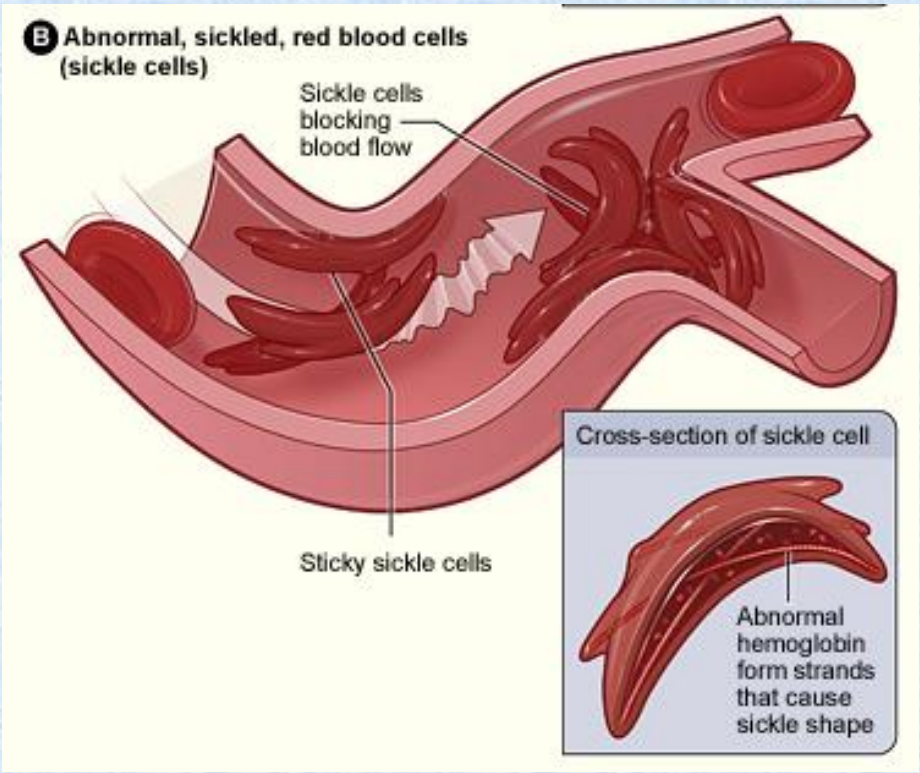
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Nucleotide | CTG | ACT | CCT | GAG | GAG | AAG | TCT |
| Amino Acid | Leu | Thr | Pro | Glu | Glu | Lys | Ser |
| | 3 | | | 6 | | | 9 |

**HBB Sequence in Mutant Adult Hemoglobin (Hb S):**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Nucleotide | CTG | ACT | CCT | GTG | GAG | AAG | TCT |
| Amino Acid | Leu | Thr | Pro | Val | Glu | Lys | Ser |
| | 3 | | | 6 | | | 9 |

# Proteins: Functions

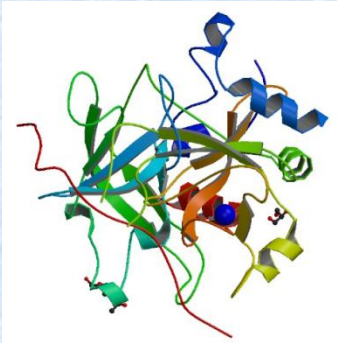


# Proteins: Functions

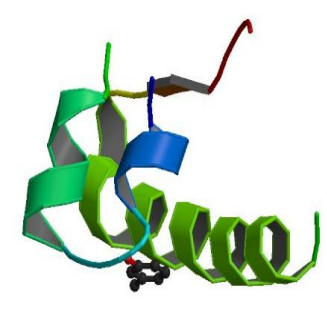
## Defense proteins

- **Antibodies** (immunoglobulins) can recognize and precipitate or neutralize invading bacteria or viruses from other species;
- Thrombin is a **blood clotting protein** that prevent loss of blood when a vascular system is injured.



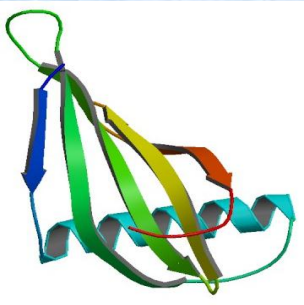
## Regulatory proteins

- Regulate cellular or physiological activity.
- Eg. insulin, which regulates **sugar metabolism**.



## Others

- Monellin has a intensely sweet taste, which is used to be a **sweetener**.



# Proteins: Functions

## Nutrient and strorage proteins

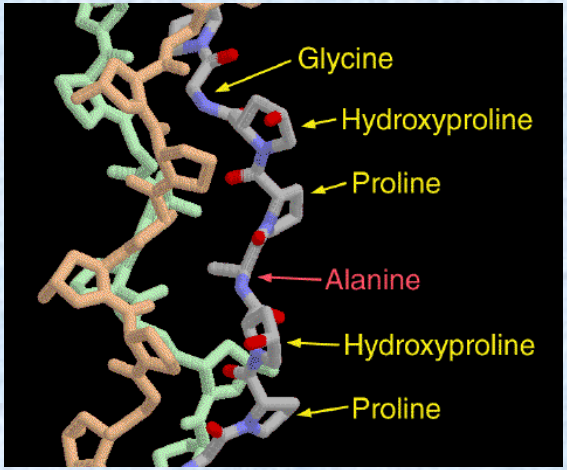
- Ovalbumin, major protein of egg white, caesin in milk

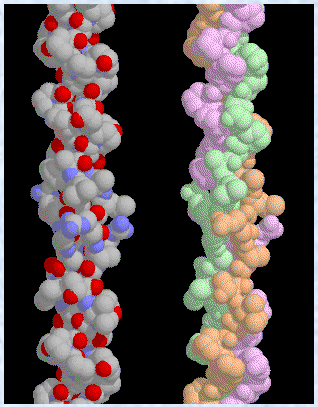
## Contractile or Motile proteins

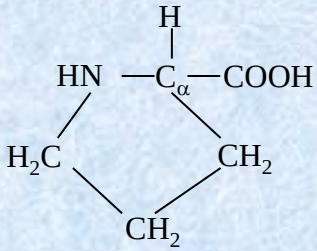
- Actin and myosin function in the contractile system of skeletal muscle and many non-muscle cells.

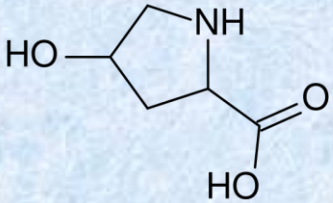
## Structural proteins

- Give biological **strength/ protection**. Major component of tendons and cartilage is the fibrous protein, collagen. Hairs, finger nails and feathers consists of keratin.



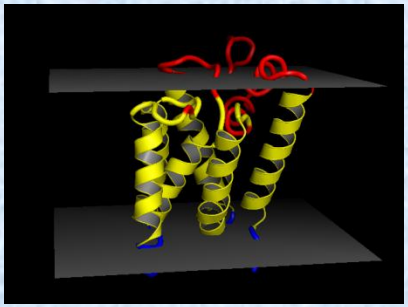


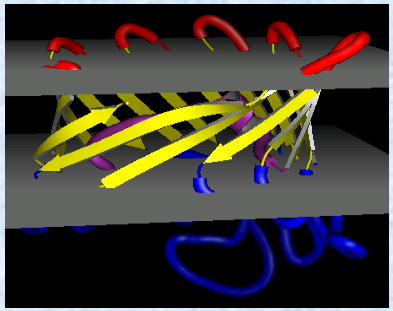




**Hydroxyproline is produced by hydroxylation of the amino acid proline by the enzyme prolyl hydroxylase following protein synthesis**

# Membrane proteins

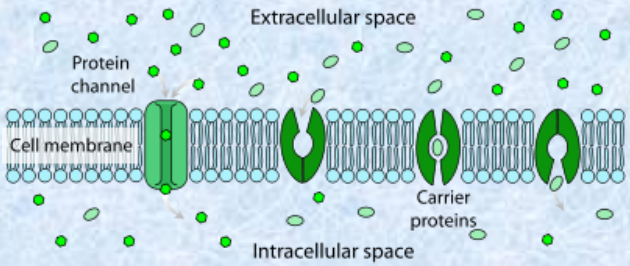
- Proteins that are embedded into membranes are called membrane proteins
- Structure: α-helical and β-barrel





- Function: Transporters, receptors, channels

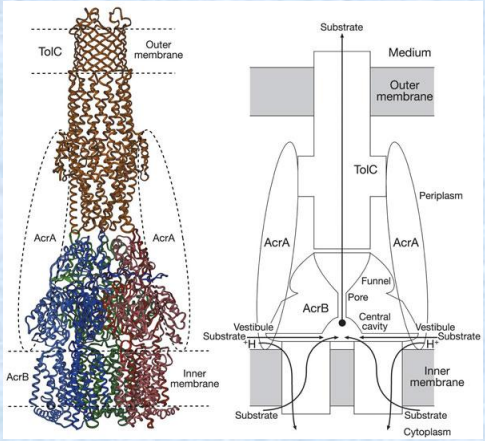
# Transporters




A membrane transport protein (transporter) is a membrane protein involved in the movement of ions, small molecules, or macromolecules such as another protein across a biological membrane.

The proteins may assist in the movement of substances by facilitated diffusion or active transport.

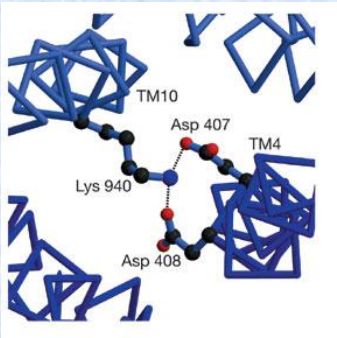
E.g. Multidrug efflux transporter AcrB




When substrate is transported AcR might recruit TolC to form direct transit pathway from the cytoplasm to extracellular.

Two pathways (through cytoplasm; central cavity; central trimer hole plays a role in substrate translocation)

Ion pairs are possible candidates for transmembrane proton translocation site




Asp are protonated, ion pairs are disturbed; helices TM4 and TM10: conformation change -> transduced to the pore region by remote conformational coupling and open the pore.

**Murakami et al. (2002) Nature 419, 587**

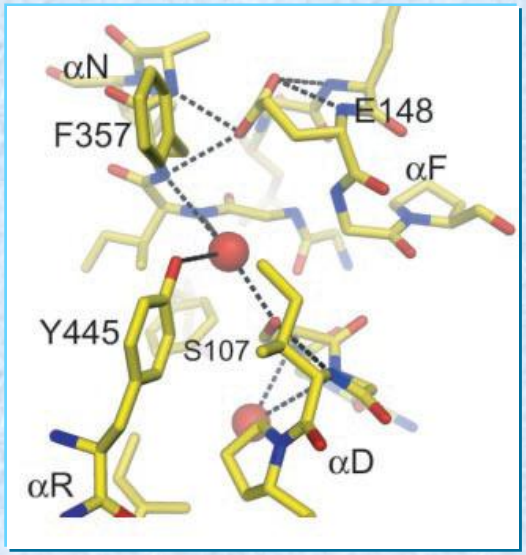
# Channels

**Ion channels** are membrane proteins that selectively allow a given species of ion to pass through them.

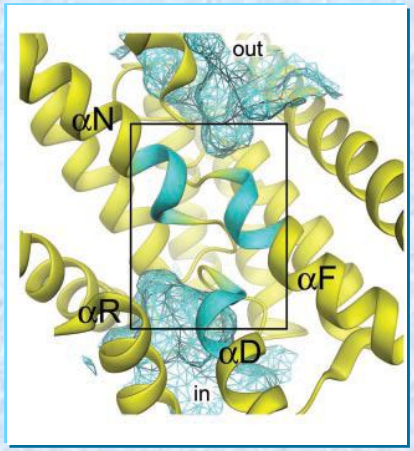
The two key properties are **selective ion conduction** and **gating**.

Selective conduction refers to a channel's ability to **select one ion species** among those present in the cellular environment.

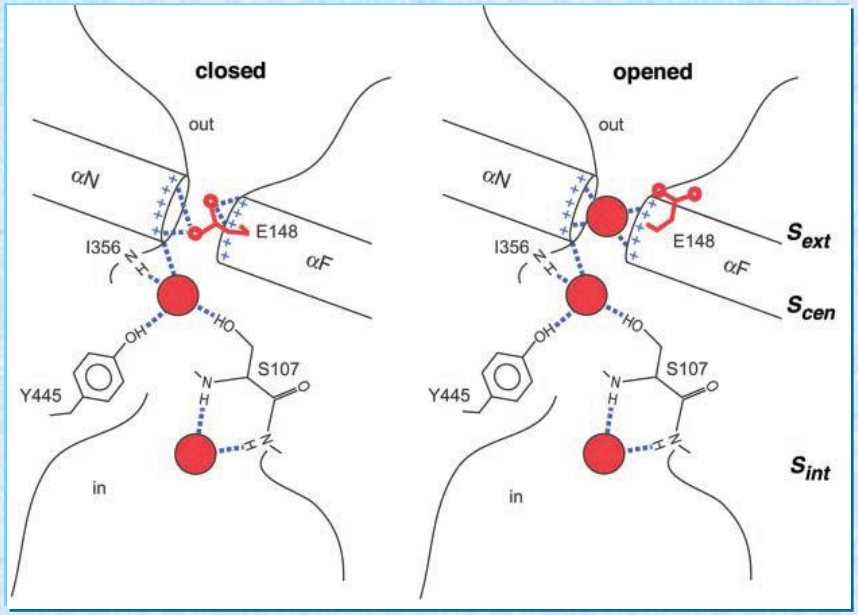
Gating refers to **opening and closing** the pore, the process by which ion conduction is turned on or off




Residues in the vicinity of chloride ions




Chloride channel
In: intracellular
Out: extracellular




Dutzler et al. (2003) Science 300, 108

# Olfactory receptors

**Olfaction is one of the senses involved in the perception of chemo-signals in the external environment.**

**The detection of odorant molecules involves specific binding to specialized receptor in olfactory system.**

**OR proteins belong to the G Protein-Coupled Receptor superfamily, which is characterized by the presence of hydrophobic transmembrane domains.**

**Each OR can recognize several chemically related molecules, and a specific odorant may bind to several ORs**

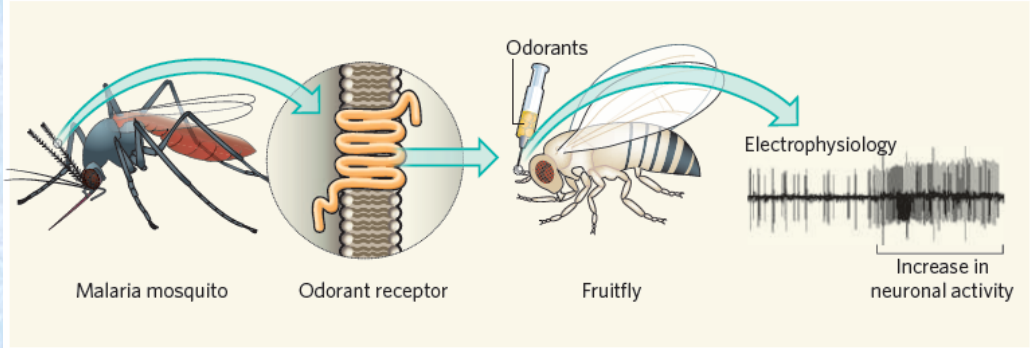



**Figure 1 | What's that smell?** Odorant receptors (ORs) in the antennas of the malaria mosquito *Anopheles gambiae* detect odorants produced by humans. The genes that might encode ORs have been identified from the insect's genome[4]. Carlson and colleagues[3] expressed 72 of these genes, one kind at a time, in the 'empty neurons' of the fruitfly *Drosophila melanogaster*. Of the putative ORs integrated into empty neurons, 50 turned out to be functional in the resulting mutant flies. The authors exposed these flies to a panel of 110 odorants, and measured the electrical activity of the OR-containing neurons. Odorants that bind to and activate ORs cause an increase in the number of spikes in the neuron's electrical activity. In this way, the authors identified which odorants activate (or, in some cases, inhibit) which ORs in *A. gambiae*.

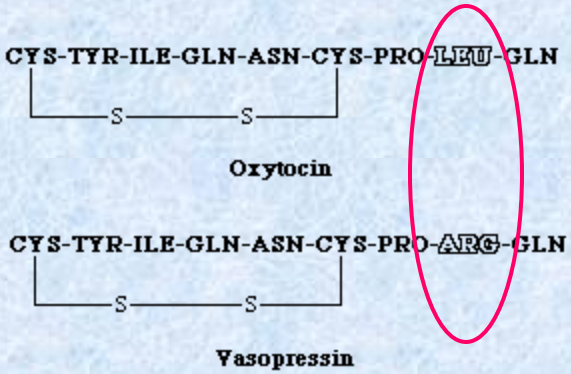
# Structure-Function Relationship

Protein structure → Protein function

Function of a protein is dictated by its structure

## Hormones

Simple change cause high blood pressure

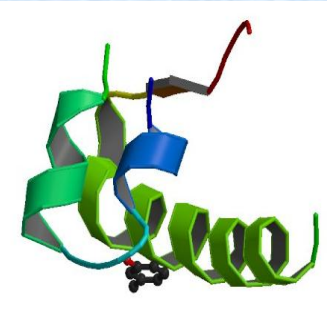



**HBB Sequence in Normal Adult Hemoglobin (Hb A):**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Nucleotide | CTG | ACT | CCT | GAG | GAG | AAG | TCT |
| Amino Acid | Leu | Thr | Pro | Glu | Glu | Lys | Ser |
| | 3 | | | 6 | | | 9 |

**HBB Sequence in Mutant Adult Hemoglobin (Hb S):**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Nucleotide | CTG | ACT | CCT | GTG | GAG | AAG | TCT |
| Amino Acid | Leu | Thr | Pro | Val | Glu | Lys | Ser |
| | 3 | | | 6 | | | 9 |



| MUTATION | | | PROTEIN | GENE_NAME | DISEASE |
|---|---|---|---|---|---|
| Phe | 24 | Ser | INSULIN | INS | DIABETES, MODY |
| His | 10 | Asp | INSULIN | INS | HYPERPROINSULINAEMIA |
| Phe | 24 | Leu | INSULIN | INS | HYPERPROINSULINAEMIA |
| Arg | 39 | His | INSULIN | INS | HYPERPROINSULINAEMIA |
| Arg | 39 | Leu | INSULIN | INS | HYPERPROINSULINAEMIA |
| Arg | 39 | Pro | INSULIN | INS | HYPERPROINSULINAEMIA |
| Val | 42 | Leu | INSULIN | INS | HYPERPROINSULINAEMIA |

http://wiz2.pharm.wayne.edu/biochem/prot.html

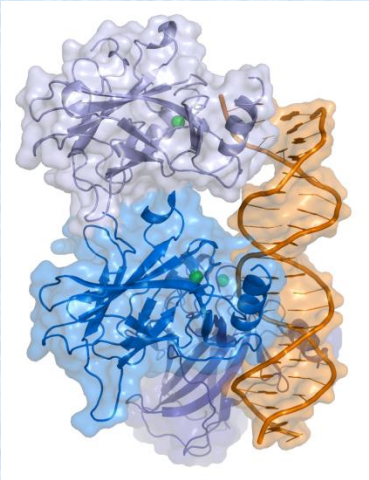
# Structure-Function Relationship

Mutations in the p53 causes human cancer.

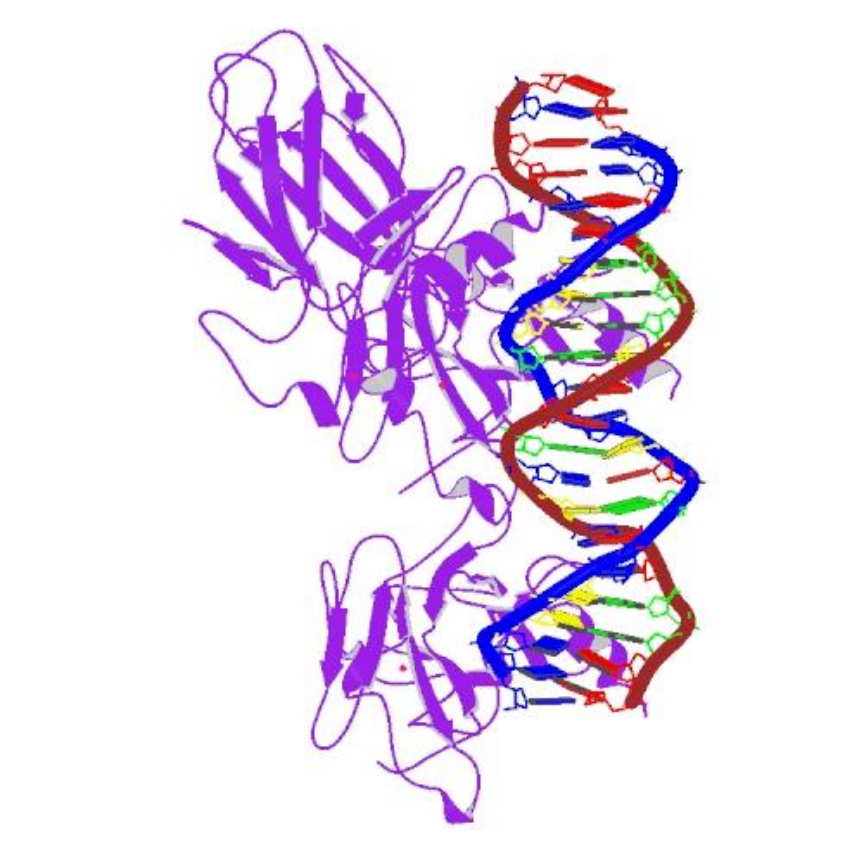
The majority of the mutations occur in the core domain (sequence-specific DNA binding activity of the p53 protein; residues 102-292)

Result in loss of DNA binding.

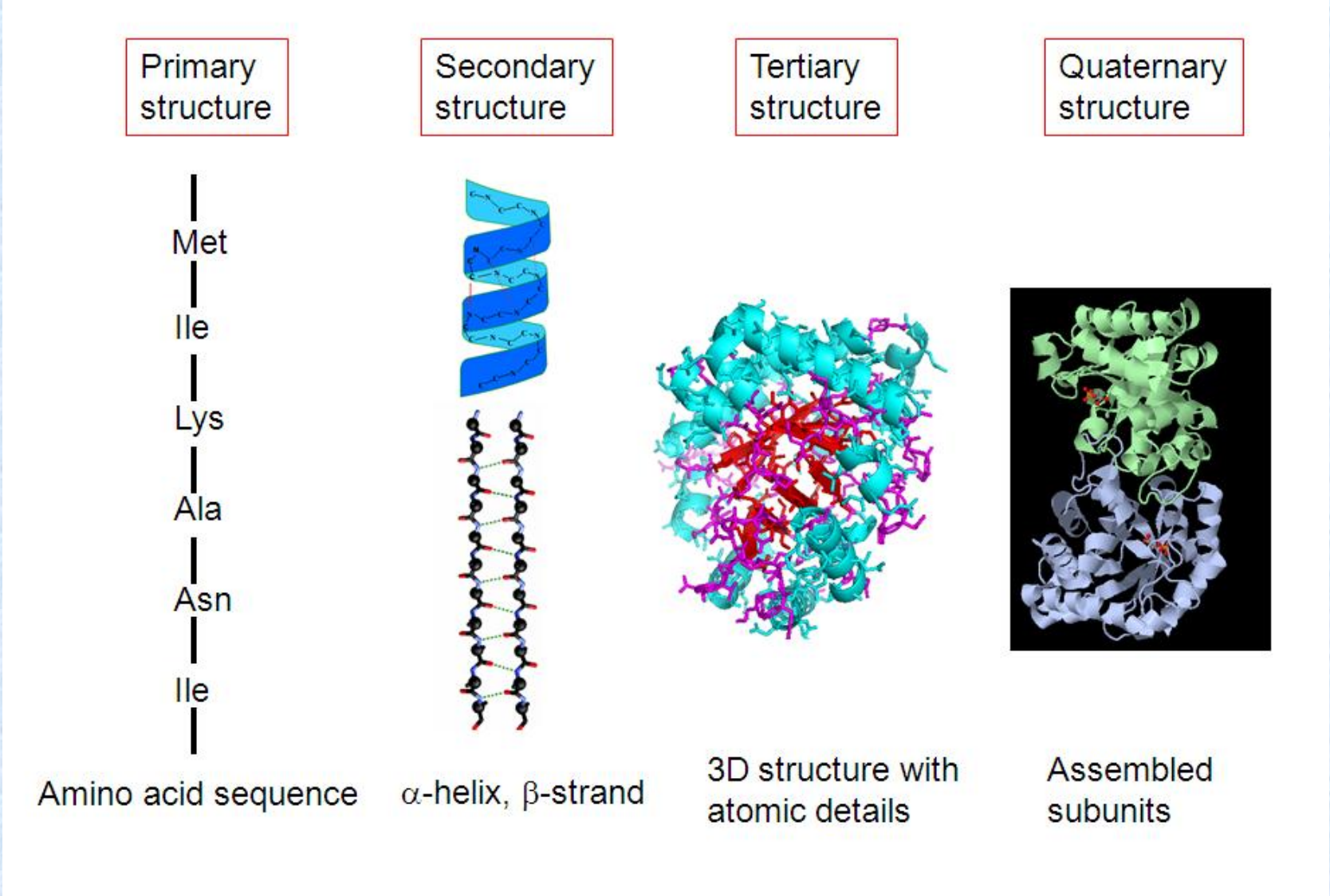
The structure supports the hypothesis that DNA binding is critical for the biological activity of p53



TUMOR SUPPRESSOR P53
COMPLEXED WITH DNA



# Protein Structure



# Primary structure: human hemoglobin

>sp|P68871|HBB_HUMAN Hemoglobin subunit beta OS=Homo sapiens

VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPD
AVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFR
LLGNVLVCVLAHHFG KEFTPPVQAAYQKVVAGVANALAHKYH

**Primary structure describes the linear sequence of amino acid residues in a protein.**

**It includes all covalent bonds between amino acids.**

**The relative arrangement of the linked amino acids is not specified.**

# Databases for protein sequences

EXProt

Munich Information Center for Protein Sequences (MIPS)

NCBI Protein database

PIR - Protein Information Resource (Georgetown University)

PIR-NREF

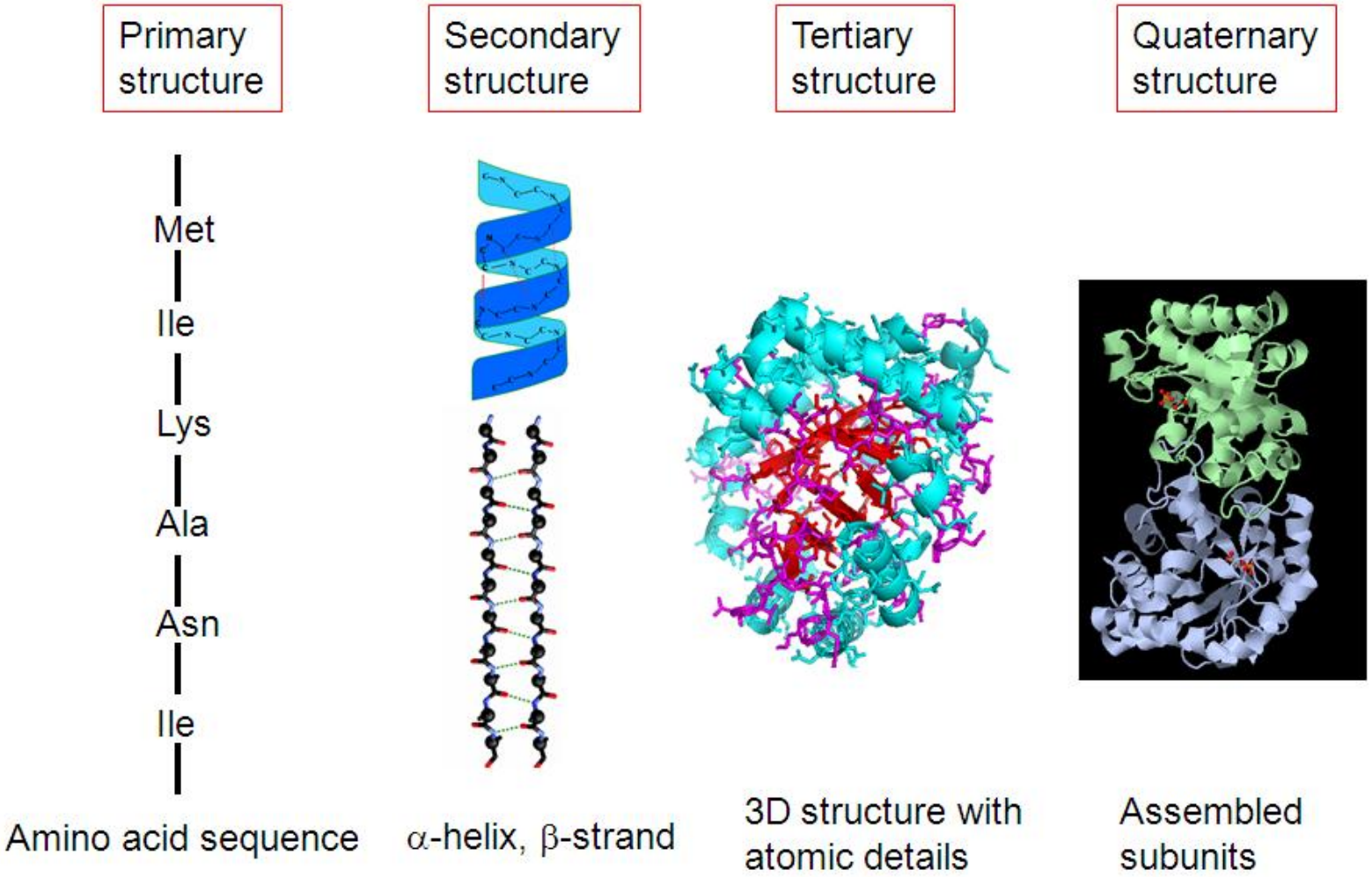
PRF

SWISS-PROT (Swiss Institute of Bioinformatics)

TrEMBL

**UniProt - The Universal Protein Knowledgebase**

# Protein Structure



# Primary structure: human hemoglobin

>sp|P68871|HBB_HUMAN Hemoglobin subunit beta OS=Homo sapiens

VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPD
AVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFR
LLGNVLVCVLAHHFG KEFTPPVQAAYQKVVAGVANALAHKYH

**Primary structure describes the linear sequence of amino acid residues in a protein.**

**It includes all covalent bonds between amino acids.**

**The relative arrangement of the linked amino acids is not specified.**

# Databases for protein sequences

EXProt

Munich Information Center for Protein Sequences (MIPS)

NCBI Protein database

PIR - Protein Information Resource (Georgetown University)

PIR-NREF

PRF

SWISS-PROT (Swiss Institute of Bioinformatics)

TrEMBL

**UniProt - The Universal Protein Knowledgebase**

# Protein Information Resource

**PRO: Protein family classification**

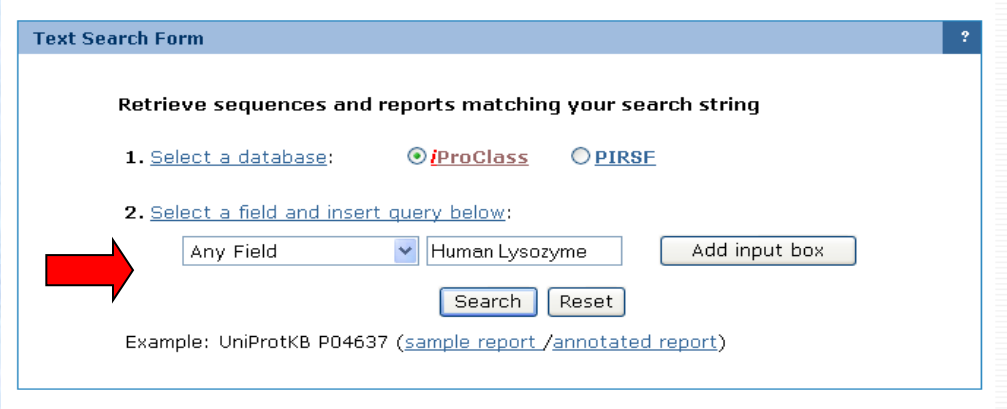
**iProClass: integrated protein knowledgebase**

**iProLink: literature, information and knowledge**

**http://pir.georgetown.edu/**

# Search with iProClass

**The iProClass database provides value-added information reports on protein sequences, structures, families, functions, interactions, expressions and modifications.**



1 selected (show) GO Slim / KEGG Pathway... • BLAST • FASTA • Pattern Match • Pairwise Alignment • Multiple Alignment • Domain Display

| Protein AC/ID | Protein Name | Length | Organism Name | PIRSF ID | Related Seq. + | Matched Fields |
|---|---|---|---|---|---|---|
| P79239/LYSC_PONPY ProClass UniProtKB/Swiss-Prot | Lysozyme C precursor BioThesaurus | 148 | Pongo pygmaeus (Bornean orangutan) | PIRSF001064 | 300 | Paper Title=>human lysozyme |
| P61628/LYSC_PANTR ProClass UniProtKB/Swiss-Prot | Lysozyme C precursor BioThesaurus | 148 | Pan troglodytes (Chimpanzee) | PIRSF001064 | 300 | Paper Title=>human lysozyme |
| P61627/LYSC_PANPA ProClass UniProtKB/Swiss-Prot | Lysozyme C precursor BioThesaurus | 148 | Pan paniscus (Pygmy chimpanzee) (Bonobo) | PIRSF001064 | 300 | Paper Title=>human lysozyme |
| P61626/LYSC_HUMAN ProClass UniProtKB/Swiss-Prot | Lysozyme C precursor BioThesaurus | 148 | Homo sapiens (Human) | PIRSF001064 | 300 | Paper Title=>human lysozyme; Paper Title=>human |
| P79179/LYSC_GORGO ProClass UniProtKB/Swiss-Prot | Lysozyme C precursor BioThesaurus | 148 | Gorilla gorilla gorilla (Lowland gorilla) | PIRSF001064 | 300 | Paper Title=>human lysozyme |
| P02788/TRFL_HUMAN ProClass UniProtKB/Swiss-Prot | Lactotransferrin precursor BioThesaurus | 710 | Homo sapiens (Human) | PIRSF002549; PIRSF500683 | 300 | Paper Title=>human lysozyme |
| Q6PCD2/Q6PCD2_HUMAN ProClass UniProtKB/TrEMBL | GABRE protein BioThesaurus | 365 | Homo sapiens (Human) | | 300 | Paper Title=>human lysozyme |
| B2R4C5/B2R4C5_HUMAN ProClass UniProtKB/TrEMBL | Lysozyme (Renal amyloidosis), isoform CRA_a BioThesaurus | 148 | Homo sapiens (Human) | | 300 | Paper Title=>human lysozyme; Paper Title=>human |
| Q876Z9/Q876Z9_ASPOR ProClass UniProtKB/TrEMBL | Predicted protein BioThesaurus | 600 | Aspergillus oryzae | PIRSF037788; PIRSF500676 | 300 | Paper Title=>human lysozyme |
| Q4R8K7/Q4R8K7_MACFA ProClass UniProtKB/TrEMBL | Testis cDNA clone: QtsA-12244, similar to human lysozyme homolog (LOC57151), BioThesaurus | 109 | Macaca fascicularis (Crab eating macaque) (Cynomolgus monkey) | | 300 | Protein Name=>human lysozyme |

Save Result As: TABLE | FASTA

1 selected (show) GO Slim / KEGG Pathway... • BLAST • FASTA • Pattern Match • Pairwise Alignment • Multiple Alignment • Domain Display

| Protein AC/ID | Protein Name | Length | Organism Name | PIRSF ID | Related Seq. + | Matched Fields |
|---|---|---|---|---|---|---|
| P61626/LYSC_HUMAN ProClass UniProtKB/Swiss-Prot | Lysozyme C precursor BioThesaurus | 148 | Homo sapiens (Human) | PIRSF001064 | 300 | UniProtKB AC=>P61626 |

PIR A UniProt CONSORTIUM MEMBER
Protein Information Resource

Text Search:

About PIR | Databases | Search/Analysis | Download | Support

## iProClass Summary Report for UniProtKB Entry: P61626

Related Sequences | BioThesaurus | ID Mapping

### GENERAL INFORMATION

**Protein Name and ID**

| UniProtKB ID | UniProtKB Accession | Protein Name |
|---|---|---|
| LYSC_HUMAN | P61626; P00695; Q13170; Q9UCF8 | Lysozyme C precursor |

PIR-PSD: LZHU
RefSeq: NP_000230.1
GenPept: AAA59535.1 ; AAC63078.1 ; EAW97222.1 ; AAA59536.1 ; CAA32175.1 ; AAH04147.1 ; EAW97221.1 ; ACO37637.1 ; AAA36188.1
IPI: IPI00019038

**Taxonomy**

*Source Organism:* Homo sapiens (Human)
*Taxon Group:* Euk/mammal
*NCBI Taxon:* 9606
*Lineage:* Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

**Gene Name**

LYZ; LZM

**Keywords**

3d-structure; amyloid; amyloidosis; antimicrobial; bacteriolytic enzyme; direct protein sequencing; disease mutation; disulfide bond; glycosidase; hydrolase; polymorphism; polysaccharide degradation; signal

**Function**

Lysozymes have primarily a bacteriolytic function; those in tissues and body fluids are associated with the monocyte- macrophage system and enhance the activity of immunoagents.

**Subunit**

Monomer.

### CROSS-REFERENCES

**Bibliography**

►View Bibliography Information ►Submit Bibliography
*Annotated references:*
PMID: 8105095; 10350481; 10469827; 10561612; 11887182; 11927576; 11986950 [PDB/GeneRIF]
More

*Other references:*
PMID: 11849445; 12675840; 15745733; 8765309; 9659355; 9745729; 18391951; 9359845; 8566845; 17353931; 9883972; 366724; 10534505; 12477932; 10558865; 18591461

**DNA Sequence**

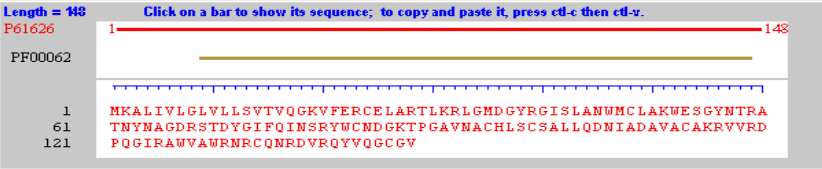
GenBank/EMBL/DDBJ: M21119; J03801; M19045; X14008; U25677; BC004147

| | |
|---|---|
| Structure | 1B5U: SCOP CATH FSSP MMDB PDBsum<br>1B5V: SCOP CATH FSSP MMDB PDBsum<br>1B5W: SCOP CATH FSSP MMDB PDBsum<br>1B5X: SCOP CATH FSSP MMDB PDBsum<br>1B5Y: SCOP CATH FSSP MMDB PDBsum<br>1B5Z: SCOP CATH FSSP MMDB PDBsum<br>1B7L: SCOP CATH FSSP MMDB PDBsum<br>1B7M: SCOP CATH FSSP MMDB PDBsum<br>1B7N: SCOP CATH FSSP MMDB PDBsum<br>1B7O: SCOP CATH FSSP MMDB PDBsum<br>1B7P: SCOP CATH FSSP MMDB PDBsum<br>1B7Q: SCOP CATH FSSP MMDB PDBsum<br>1B7R: SCOP CATH FSSP MMDB PDBsum<br>1B7S: SCOP CATH FSSP MMDB PDBsum<br>1BB3: SCOP CATH FSSP MMDB PDBsum<br>1BB4: SCOP CATH FSSP MMDB PDBsum<br>More |
| PIR Feature & Post Translational Modifications | FEAT1; active site: Glu, Asp (53,71) [predicted]<br>FEAT2; binding site: substrate (Asp) (120) [predicted]<br>FEAT3; disulfide bonds: (24-146,48-134,83-99,95-113) [experimental]<br>FEAT4; domain: signal sequence (1-18) [predicted]<br>FEAT5; product: lysozyme (19-148) [experimental]<br><br>Phosphosite: P61626 |

## FAMILY CLASSIFICATION

| | |
|---|---|
| UniRef | UniRef100_P61626; UniRef90_P61626; UniRef50_P61626 |
| PIRSF | PIRSF001064 lysozyme c |
| Pfam Domain | Pfam: PF00062: C-type lysozyme/alpha-lactalbumin family (19-146) |
| Prosite Motif | Prosite: PS00128: PDOC00119: Alpha-lactalbumin / lysozyme C signature.<br>Prosite: PS51348: PDOC00119: Alpha-lactalbumin / lysozyme C family profile. |
| InterPro | InterPro: LYSC_HUMAN<br>IPR001916: Glycoside hydrolase, family 22<br>IPR000974: Glycoside hydrolase, family 22, lysozyme |
| SCOP Fold | ►Class: Alpha and beta proteins (a+b); *Fold:* Lysozyme-like; *Superfamily:* Lysozyme-like; *Family:* C-type lysozyme [133L:A; 134L:A; 1B5U:A; 1B5V:A; 1B5W:A; 1B5X:A; 1B5Y:A; 1B5Z:A; 1B5Z:B; 1B7L:A; 1B7M:A; 1B7N:A; 1B7O:A; 1B7P:A; 1B7Q:A; 1B7R:A; 1B7S:A; 1BB3:A; 1BB3:B; 1BB4:A; 1BB4:B; 1BB5:A; 1BB5:B; 1C43:A; 1C45:A; 1C46:A; 1C7P:A; 1CJ6:A; 1CJ7:A; 1CJ8:A; 1CJ9:A; 1CKC:A; 1CKD:A; 1CKF:A; 1CKG:A; 1CKG:B; 1CKH:A; 1D6P:A; 1D6Q:A; 1DI3:A; 1DI4:A; 1DI5:A; 1EQ4:A; 1EQ5:A; 1EQE:A; 1GAY:A; 1GAZ:A; 1GB0:A; 1GB2:A; More] |
| Other Classification | BLOCKS: IPB000974 Lysozyme signature<br>PRINTS: PR00137 LYSOZYME<br>PRINTS: PR00135 LYZLACT<br>SMART: SM00263 LYZ1 |

## FEATURE & SEQUENCE DISPLAY



# Swiss-prot/Uniprot

**Annotated protein sequence database established in 1986 and maintained collaboratively, since 1987, by the Department of Medical Biochemistry of the University of Geneva and the EMBL Data Library.**

**It is a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domain structure, post-translational modifications and variants), a minimal level of redundancy and a high level of integration with other databases.**

**TrEMBL is a computer annotated supplement of SWISS-PROT that contains all the translations of EMBL nucleotide sequence entries not yet integrated in SWISS-PROT.**

**Currently, SWISS-PROT and TrEMBL have 0.55 and 73.7 million sequences, respectively.**

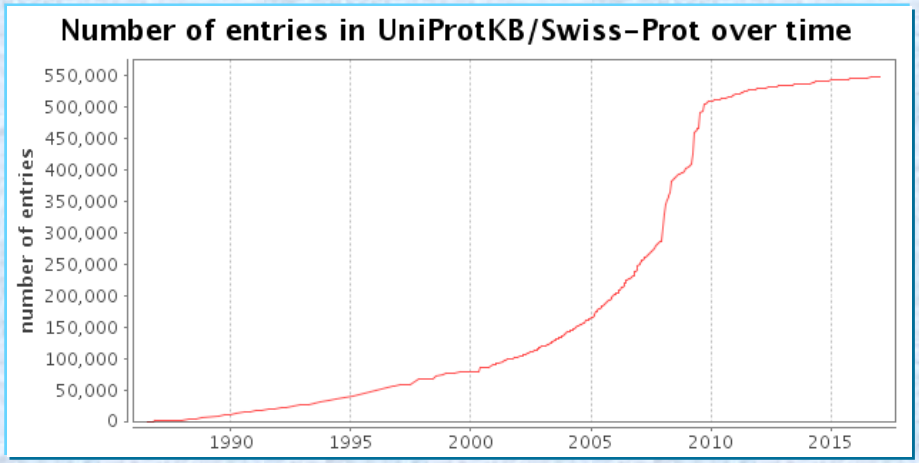
**Total: 74.4 million**

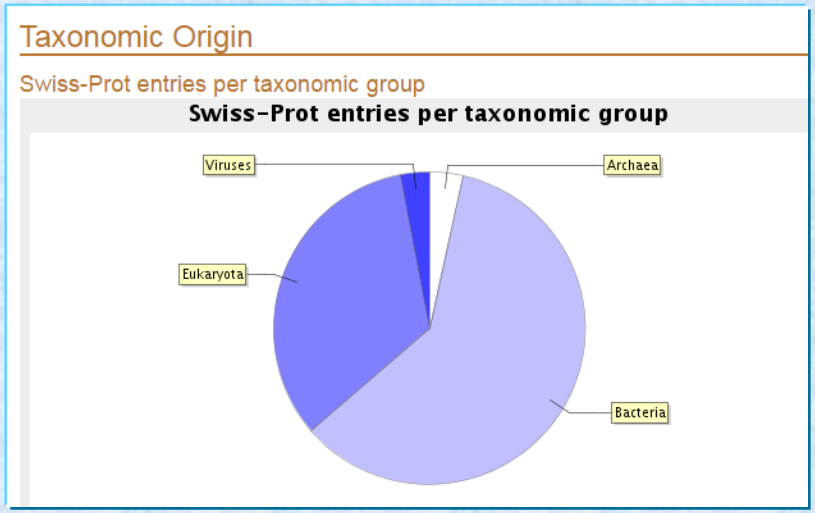
**http://www.ebi.ac.uk/swissprot/**

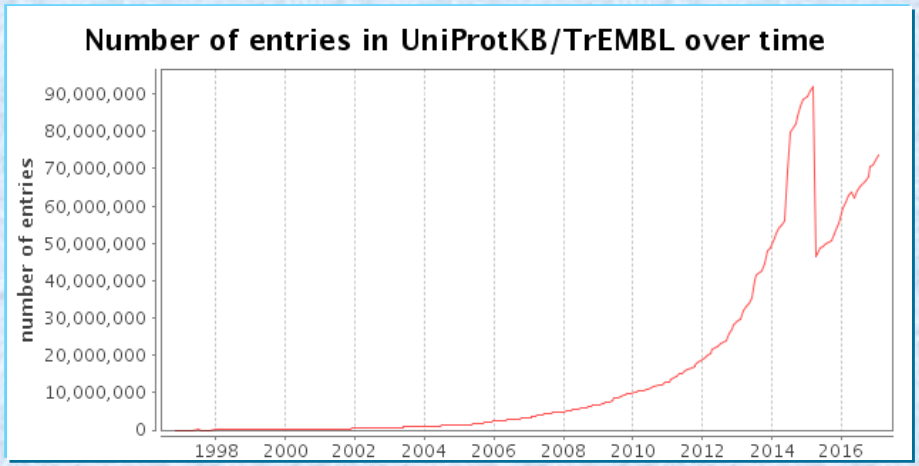
**http://www.uniprot.org/uniprot/**

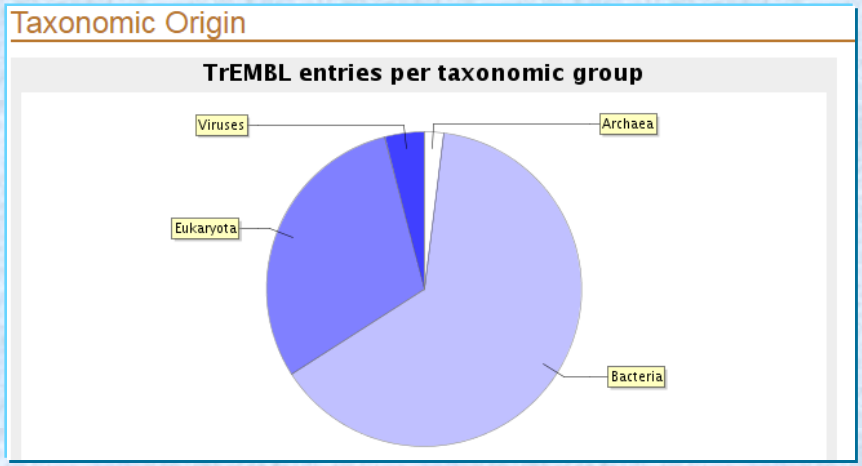
# Uniprot: statistics

Number of entries: ~74.4 million



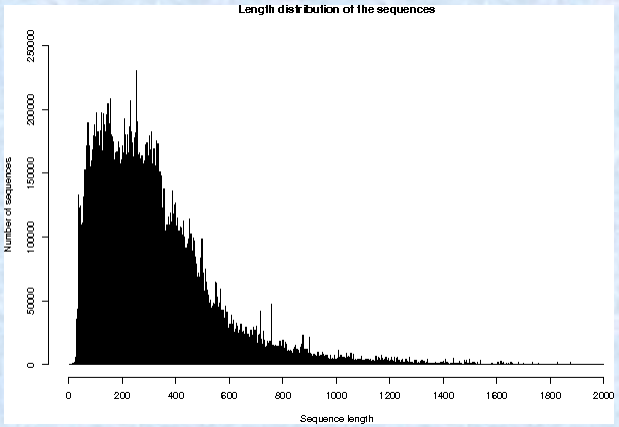



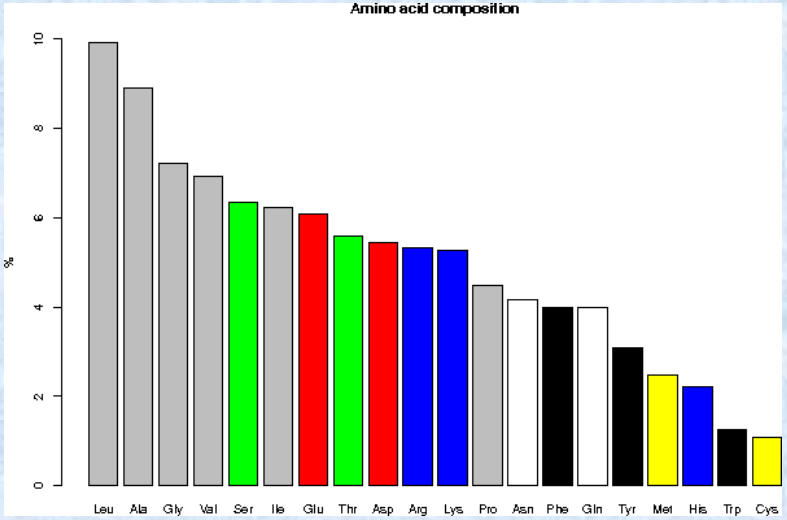



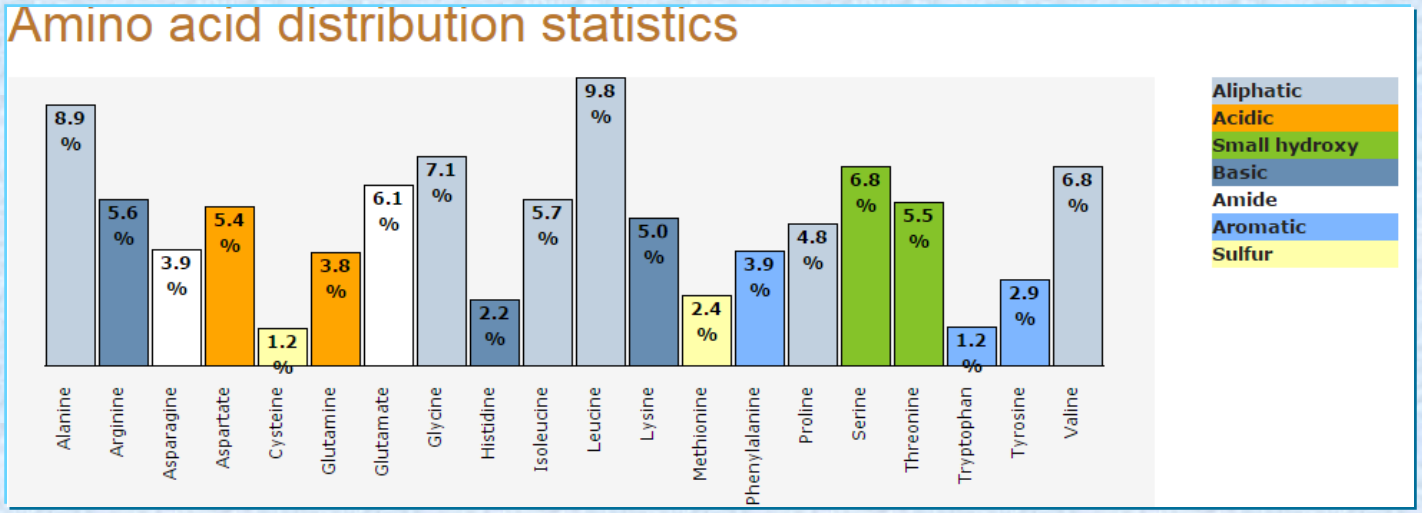



# Uniprot: statistics

Average sequence length:
315 amino acids



# Uniprot: contents

**Name and origin of the protein**

**protein attributes**

**general information**

**Ontologies**

**sequence annotation**

**amino acid sequence**

**bibliographic references**

**cross-references with sequence, structure and interaction databases**

**entry information.**

# Uniprot: search results

UniProt › UniProtKB

Search | Blast | Align | Retrieve | ID Mapping *

Search in: Protein Knowledgebase (UniProtKB)

Query: hemoglobin B chain

Search | Advanced Search » | Clear

1 - 25 of 1,243 results for **hemoglobin** ⊠ **AND B** ⊠ **AND chain** ⊠ in **UniProtKB** sorted by **score** descending ⊠

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50% | Download

Page 1 of 50 | Next »

## Results Customize

› Show only reviewed (825) (UniProtKB/Swiss-Prot) or unreviewed (418) (UniProtKB/TrEMBL) entries

› Quote terms: "hemoglobin b"

› Restrict term "hemoglobin" to protein family (24), gene name (3), gene ontology (883), keyword (9), protein name (970), web resource (5)

› Restrict term "b" to author (608), domain (4), gene name (15), gene ontology (5), keyword (4), protein name (58), organism (1), source (7), strain (5), taxonomy (6), tissue (2)

› Restrict term "chain" to author (48), gene name (2), gene ontology (7), keyword (7), protein name (740), annotation topic (943)

› Restrict term "chain" to pathway

| Accession | Entry name | Status | Protein names | Gene names | Organism | Length |
|---|---|---|---|---|---|---|
| P68871 | HBB_HUMAN | ★ | **Hemoglobin subunit beta** | **HBB** | Homo sapiens (Human) | 147 |
| P69905 | HBA_HUMAN | ★ | **Hemoglobin subunit alpha** | **HBA1** **HBA2** | Homo sapiens (Human) | 142 |
| P69892 | HBG2_HUMAN | ★ | **Hemoglobin subunit gamma-2** | **HBG2** | Homo sapiens (Human) | 147 |
| P69891 | HBG1_HUMAN | ★ | **Hemoglobin subunit gamma-1** | **HBG1** PRO2979 | Homo sapiens (Human) | 147 |
| P0A592 | TRHBN_MYCTU | ★ | **Group 1 truncated hemoglobin glbN** | **glbN** Rv1542c MT1594 MTCY48.23 | Mycobacterium tuberculosis | 136 |
| P02042 | HBD_HUMAN | ★ | **Hemoglobin subunit delta** | **HBD** | Homo sapiens (Human) | 147 |

# Uniprot: contents

## Names and origin

| | |
|---|---|
| Protein names | *Recommended name:*<br>**Hemoglobin subunit beta**<br>*Alternative name(s):*<br>Beta-globin<br>Hemoglobin beta chain<br>Cleaved into the following chain:<br>1. ***LVV-hemorphin-7*** |
| Gene names | Name: **HBB** |
| Organism | **Homo sapiens (Human)** [Complete proteome] |
| Taxonomic identifier | 9606 [NCBI] |
| Taxonomic lineage | Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Mammalia › Eutheria › Euarchontoglires › Primates › Haplorrhini › Catarrhini › Hominidae › Homo |

## Protein attributes

| | |
|---|---|
| Sequence length | 147 AA. |
| Sequence status | Complete. |
| Sequence processing | The displayed sequence is further processed into a mature form. |
| Protein existence | Evidence at protein level. |

# Uniprot: contents

## General annotation (Comments)

| | |
|---|---|
| Function | Involved in oxygen transport from the lung to the various peripheral tissues. Ref.35<br>LVV-hemorphin-7 potentiates the activity of bradykinin, causing a decrease in blood pressure. Ref.35 |
| Subunit structure | Heterotetramer of two alpha chains and two beta chains in adult hemoglobin A (HbA). |
| Tissue specificity | Red blood cells. |
| Post-translational modification | Glucose reacts non-enzymatically with the N-terminus of the beta chain to form a stable ketoamine linkage. This takes place slowly and continuously throughout the 120-day life span of the red blood cell. The rate of glycation is increased in patients with diabetes mellitus.<br>S-nitrosylated; a nitric oxide group is first bound to $Fe^{2+}$ and then transferred to Cys-94 to allow capture of $O_2$.<br>Acetylated on Lys-60, Lys-83 and Lys-145 upon aspirin exposure. Ref.34 reports the identification of HBB acetylated on Lys-145 in the cytosolic fraction of HeLa cells. This may have resulted from contamination of the sample. |
| Involvement in disease | Defects in HBB may be a cause of Heinz body anemias (HEIBAN) [MIM:140700]. This is a form of non-spherocytic hemolytic anemia of Dacie type 1. After splenectomy, which has little benefit, basophilic inclusions called Heinz bodies are demonstrable in the erythrocytes. Before splenectomy, diffuse or punctate basophilia may be evident. Most of these cases are probably instances of hemoglobinopathy. The hemoglobin demonstrates heat lability. Heinz bodies are observed also with the Ivemark syndrome (asplenia with cardiovascular anomalies) and with glutathione peroxidase deficiency. Ref.51 Ref.125 Ref.126 Ref.129<br>Defects in HBB are the cause of beta-thalassemia (B-THAL) [MIM:604131]. A form of thalassemia. Thalassemias are common monogenic diseases occurring mostly in Mediterranean and Southeast Asian populations. The hallmark of beta-thalassemia is an imbalance in globin-chain production in the adult HbA molecule. Absence of beta chain causes beta(0)-thalassemia, while reduced amounts of detectable beta globin causes $beta^{+}$-thalassemia. In the severe forms of beta-thalassemia, the excess alpha globin chains accumulate in the developing erythroid precursors in the marrow. Their deposition leads to a vast increase in erythroid apoptosis that in turn causes ineffective erythropoiesis and severe microcytic hypochromic anemia. Clinically, beta-thalassemia is divided into thalassemia major which is transfusion dependent, thalassemia intermedia (of intermediate severity), and thalassemia minor that is asymptomatic. Ref.52<br>Defects in HBB are the cause of sickle cell anemia (SKCA) [MIM:603903]; also known as sickle cell disease. Sickle cell anemia is characterized by abnormally shaped red cells resulting in chronic anemia and periodic episodes of pain, serious infections and damage to vital organs. Normal red blood cells are round and flexible and flow easily through blood vessels, but in sickle cell anemia, the abnormal hemoglobin (called Hb S) causes red blood cells to become stiff. They are C-shaped and resembles a sickle. These stiffer red blood cells can led to microvascular occlusion thus cutting off the blood supply to nearby tissues.<br>Defects in HBB are the cause of beta-thalassemia dominant inclusion body type (B-THALIB) [MIM:603902]. An autosomal dominant form of beta thalassemia characterized by moderate anemia, lifelong jaundice, cholelithiasis and splenomegaly, marked morphologic changes in the red cells, erythroid hyperplasia of the bone marrow with increased numbers of multinucleate red cell precursors, and the presence of large inclusion bodies in the normoblasts, both in the marrow and in the peripheral blood after splenectomy. Ref.52 |
| Miscellaneous | One molecule of 2,3-bisphosphoglycerate can bind to two beta chains per hemoglobin tetramer. |

# Uniprot: contents

## Ontologies

**Keywords**

| | |
|---|---|
| Biological process | Oxygen transport<br>Transport |
| Coding sequence diversity | Polymorphism |
| Disease | Congenital dyserythropoietic anemia<br>Disease mutation<br>Hereditary hemolytic anemia |
| Ligand | Heme<br>Iron<br>Metal-binding<br>Pyruvate |
| Molecular function | Hypotensive agent<br>Vasoactive |
| PTM | Acetylation<br>Glycation<br>Glycoprotein<br>Phosphoprotein<br>S-nitrosylation |
| Technical term | 3D-structure<br>Complete proteome<br>Direct protein sequencing |

**Gene Ontology (GO)**

| | |
|---|---|
| Biological process | blood coagulation<br>Traceable author statement. Source: Reactome<br>hydrogen peroxide catabolic process<br>Inferred from direct assay. Source: BHF-UCL<br>nitric oxide transport<br>Non-traceable author statement. Source: UniProtKB<br>positive regulation of cell death<br>Inferred from direct assay. Source: BHF-UCL<br>positive regulation of nitric oxide biosynthetic process<br>Non-traceable author statement. Source: UniProtKB<br>protein heterooligomerization<br>Inferred from direct assay. Source: BHF-UCL<br>regulation of blood pressure<br>Inferred from electronic annotation. Source: UniProtKB-KW<br>regulation of blood vessel size<br>Inferred from electronic annotation. Source: UniProtKB-KW |
| Cellular component | haptoglobin-hemoglobin complex<br>Inferred from direct assay. Source: BHF-UCL<br>hemoglobin complex<br>Non-traceable author statement Ref.33 Ref.71. Source: UniProtKB |
| Molecular function | heme binding<br>Inferred from electronic annotation. Source: InterPro<br>hemoglobin binding<br>Inferred from direct assay. Source: UniProtKB<br>oxygen binding<br>Inferred from direct assay. Source: UniProtKB<br>oxygen transporter activity<br>Non-traceable author statement Ref.71. Source: UniProtKB |

# Uniprot: contents

## Binary interactions

| With | Entry | #Exp. | IntAct | Not |
|---|---|---|---|---|
| HBA1 | P69905 | 1 | EBI-715554,EBI-714680 | |

## Sequence annotation (Features)

| | Feature key | Position(s) | Length | Descrip |
|---|---|---|---|---|
| **Molecule processing** | | | | |
| | Initiator methionine | 1 | 1 | Remov |
| | Chain | 2 – 147 | 146 | Hemog |
| | Peptide | 33 – 42 | 10 | LVV-he |
| **Sites** | | | | |
| | Metal binding | 64 | 1 | Iron (he |
| | Metal binding | 93 | 1 | Iron (he |
| | Binding site | 2 | 1 | 2,3-bis |
| | Binding site | 3 | 1 | 2,3-bis |
| | Binding site | 83 | 1 | 2,3-bis |
| | Binding site | 144 | 1 | 2,3-bis |

## Natural variations

| | | | | |
|---|---|---|---|---|
| | Natural variant | 2 | 1 | V → A in Raleigh; O(2) affinity down. [dbSNP:rs33949930] |
| | Natural variant | 3 | 1 | H → L in Graz. [dbSNP:rs35906307] Ref.76 |
| | Natural variant | 3 | 1 | H → Q in Okayama; O(2) affinity up. [dbSNP:rs713040] |
| | Natural variant | 3 | 1 | H → R in Deer Lodge; O(2) affinity up. [dbSNP:rs33983205] |
| | Natural variant | 3 | 1 | H → Y in Fukuoka. [dbSNP:rs35906307] |
| | Natural variant | 6 | 1 | P → R in Warwickshire. [dbSNP:rs34769005] |
| | Natural variant | 7 | 1 | E → A in G-Makassar. |
| | Natural variant | 7 | 1 | E → K in C. Ref.3 Ref.49 |
| | Natural variant | 7 | 1 | E → Q in Machida. [dbSNP:rs33930165] |
| | Natural variant | 7 | 1 | E → V in S; sickle cell anemia. [dbSNP:rs334] Ref.10 Ref.39 |
| | Natural variant | 8 | 1 | E → G in G-San Jose; mildly unstable. [dbSNP:rs34948328] |
| | Natural variant | 8 | 1 | E → K in G-Siriraj. [dbSNP:rs34948328] |
| | Natural variant | 9 | 1 | K → E in N-Timone. [dbSNP:rs33932981] Ref.101 |
| | Natural variant | 9 | 1 | K → Q in J-Luhe. [dbSNP:rs33926764] |
| | Natural variant | 9 | 1 | K → T in Rio Grande. Ref.114 |
| | Natural variant | 10 | 1 | S → C in Porto Alegre; O(2) affinity up. [dbSNP:rs33918131] |
| | Natural variant | 11 | 1 | A → D in Ankara. [dbSNP:rs33947457] Ref.55 |

# Uniprot contents

**Secondary structure**

1 — 147

Helix | Strand | Turn

Details...

## Cross-references

### Sequence databases

- (•) EMBL
- ( ) GenBank
- ( ) DDBJ

### 3D structure databases

- (•) PDBe
- ( ) RCSB PDB
- ( ) PDBj

| Entry | Method | Resol |
|---|---|---|
| 1A00 | X-ray | 2 |
| 1A01 | X-ray | 1 |
| 1A0U | X-ray | 2 |
| 1A0Z | X-ray | 2 |
| 1A3N | X-ray | 1 |
| 1A3O | X-ray | 1 |
| 1ABW | X-ray | 2 |
| 1ABY | X-ray | 2 |
| 1AJ9 | X-ray | 2 |
| 1B86 | X-ray | 2 |
| 1BAB | X-ray | 1 |
| 1BBB | X-ray | 1 |
| 1BIJ | X-ray | 2 |
| 1BUW | X-ray | 1 |
| 1BZ0 | X-ray | 1 |
| 1BZ1 | X-ray | 1 |
| 1BZZ | X-ray | 1 |
| 1C7B | X-ray | 1 |
| 1C7C | X-ray | 1 |
| 1C7D | X-ray | 1 |
| 1CBL | X-ray | 1 |
| 1CBM | X-ray | 1.74 |
| 1CH4 | X-ray | 2.50 |

| | |
|---|---|
| ProteinModelPortal | P68871. |
| SMR | P68871. Positions 2-147. |
| ModBase | Search... |

### Protein-protein interaction databases

| | |
|---|---|
| IntAct | P68871. 24 interactions. |
| MINT | MINT-5000306. |
| STRING | P68871. |

### PTM databases

| | |
|---|---|
| PhosphoSite | P68871. |

### 2-D gel databases

| | |
|---|---|
| SWISS-2DPAGE | P68871. |
| PMMA-2DPAGE | P68871. |
| REPRODUCTION-2DPAGE | IPI00654755.<br>P68871. |
| Siena-2DPAGE | P68871. |
| UCD-2DPAGE | P02023.<br>P68871. |

### Proteomic databases

| | |
|---|---|
| PeptideAtlas | P68871. |
| PRIDE | P68871. |

ding resistance to malaria."

# Uniprot: contents

**Genome annotation databases**

| | |
|---|---|
| Ensembl | ENST00000335295; ENSP |
| GeneID | 3043. |
| KEGG | hsa:3043. |
| UCSC | uc001mae.1, human |

**Organism-specific**

| | |
|---|---|
| CTD | |
| GeneCards | |
| H-InvDB | |
| HGNC | |
| HPA | |
| MIM | |
| neXtProt | |
| Orphanet | |
| PharmGKB | |
| GenAtlas | |

**Phylogenomic databases**

| | |
|---|---|
| HOVERGEN | HBG009709. |

**Enzyme and pathway databases**

| | |
|---|---|
| Reactome | REACT_604. Hemostasis. |

**Gene expression databases**

| | |
|---|---|
| ArrayExpress | P68871. |
| DrugBank | DB00893. Iron Dextran. |
| NextBio | 12048. |
| PMAP-CutDB | P68871. |

1 hit.

## Entry information

| | |
|---|---|
| Entry name | HBB_HUMAN |
| Accession | Primary (citable) accession number: **P68871**<br>Secondary accession number(s): A4GX73 Q9UCP9 |
| Entry history | Integrated into UniProtKB/Swiss-Prot: July 21, 1986<br>Last sequence update: January 23, 2007<br>Last modified: June 28, 2011<br>This is version 97 of the entry and version 2 of the sequence. [Complete history] |
| Entry status | Reviewed (UniProtKB/Swiss-Prot) |
| Annotation program | Chordata Protein Annotation Program |
| **Disclaimer** | Any medical or genetic information present in this entry is provided for research, e<br>used as a substitute for professional medical advice, diagnosis, treatment or care |

## Relevant documents

Human chromosome 11
Human chromosome 11: entries, gene names and cross-references to MIM

Human entries with polymorphisms or disease mutations
List of human entries with polymorphisms or disease mutations

# Uniprot: search results

UniProt › UniProtKB

Search | Blast | Align | Retrieve | ID Mapping *

Search in: Protein Knowledgebase (UniProtKB)

Query: hemoglobin B chain

Search | Advanced Search » | Clear

1 - 25 of 1,243 results for **hemoglobin** ⊠ **AND B** ⊠ **AND chain** ⊠ in **UniProtKB** sorted by **score** descending ⊠

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50% | Download

Page 1 of 50 | Next »

## Results Customize

› Show only reviewed (825) (UniProtKB/Swiss-Prot) or unreviewed (418) (UniProtKB/TrEMBL) entries

› Quote terms: "hemoglobin b"

› Restrict term "hemoglobin" to protein family (24), gene name (3), gene ontology (883), keyword (9), protein name (970), web resource (5)

› Restrict term "b" to author (608), domain (4), gene name (15), gene ontology (5), keyword (4), protein name (58), organism (1), source (7), strain (5), taxonomy (6), tissue (2)

› Restrict term "chain" to author (48), gene name (2), gene ontology (7), keyword (7), protein name (740), annotation topic (943)

› Restrict term "chain" to pathway

| Accession | Entry name | Status | Protein names | Gene names | Organism | Length |
|---|---|---|---|---|---|---|
| P68871 | HBB_HUMAN | ★ | **Hemoglobin subunit beta** | **HBB** | Homo sapiens (Human) | 147 |
| P69905 | HBA_HUMAN | ★ | **Hemoglobin subunit alpha** | **HBA1** **HBA2** | Homo sapiens (Human) | 142 |
| P69892 | HBG2_HUMAN | ★ | **Hemoglobin subunit gamma-2** | **HBG2** | Homo sapiens (Human) | 147 |
| P69891 | HBG1_HUMAN | ★ | **Hemoglobin subunit gamma-1** | **HBG1** PRO2979 | Homo sapiens (Human) | 147 |
| P0A592 | TRHBN_MYCTU | ★ | **Group 1 truncated hemoglobin glbN** | **glbN** Rv1542c MT1594 MTCY48.23 | Mycobacterium tuberculosis | 136 |
| P02042 | HBD_HUMAN | ★ | **Hemoglobin subunit delta** | **HBD** | Homo sapiens (Human) | 147 |

**Search in** Sequence Clusters (UniRef)

**Query** uniprot:(hemoglobin B chain) identity:0.5 | Search | Advanced Search » | Clear

**1 - 25 of 240 results for uniprot:(hemoglobin AND B AND chain) AND identity:0.5 in UniRef**

Browse by taxonomy | Map to UniProtKB or UniParc

Download

Page 1 of 10 Next »

## Results Customize

| | Accession | Status | UniRefCluster name | Size | Members | Organisms | Length | Identity |
|---|---|---|---|---|---|---|---|---|
| ☐ | UniRef50_A1B9S1 | ★ | Cluster: TonB-dependent heme/hemoglobin receptor family protein | 1 | A1B9S1 | Paracoccus denitrificans (strain Pd 1222) | 686 | 50% |
| ☐ | UniRef50_A1U6I3 | ★ | Cluster: TonB-dependent heme/hemoglobin receptor family protein | 4 | A1U6I3<br>F2N3G5<br>C0N4R6<br>A4XQE8 | Marinobacter aquaeolei (strain ATCC 700491 / DSM 11845 / VT8) (Marinobacter hydrocarbonoclasticus (strain DSM 11845))<br>Pseudomonas stutzeri DSM 4166<br>Methylophaga thiooxydans DMS010<br>Pseudomonas mendocina (strain ymp) | 657 | 50% |
| ☐ | UniRef50_A1WMF8 | ★ | Cluster: TonB-dependent hemoglobin/transferrin/lactoferrin family receptor | 1 | A1WMF8 | Verminephrobacter eiseniae (strain EF01-2) | 730 | 50% |
| ☐ | UniRef50_A6QC57 | ★ | Cluster: Globin | 4 | A6QC57<br>E6WZ29<br>A6Q1V1<br>Q30RW0 | Sulfurovum sp. (strain NBC37-1)<br>Nitratifractor salsuginis (strain DSM 16511 / JCM 12458 / E9I37-1)<br>Nitratiruptor sp. (strain SB155-2)<br>Sulfurimonas denitrificans (strain ATCC 33889 / DSM 1251) (Thiomicrospira denitrificans (strain ATCC 33889 / DSM 1251)) | 173 | 50% |
| ☐ | UniRef50_A6UBU8 | ★ | Cluster: TonB-dependent hemoglobin/transferrin/lactoferrin family receptor | 3 | A6UBU8<br>A9D265<br>B9JZN4 | Sinorhizobium medicae (strain WSM419) (Ensifer medicae)<br>Hoeflea phototrophica DFL-43<br>Agrobacterium vitis (strain S4 / ATCC BAA-846) (Rhizobium vitis (strain S4)) | 734 | 50% |

**240** results for **uniprot:(hemoglobin ☒ AND B ☒ AND chain ☒) ☒ AND identity:0.5 ☒** in **UniRef**

› Download data compressed or **uncompressed**

**Tab-Delimited**

Summary information from the result view.

[ Download | Open | Open first 10 ]

**Excel**

Summary information from the result view for MS Excel™.

[ Download | Open | Open first 10 ]

**FASTA**

Sequence data in FASTA format.

[ Download (200 KB*) | Open | Open first 10 ]

**XML**

Complete data in XML format.

[ Download (400 KB*) | Open | Open first 10 ]

**RDF/XML**

Complete data in RDF format.

[ Download (500 KB*) | Open | Open first 10 ]

**List**

List of accession numbers.

[ Download (5 KB*) | Open | Open first 10 ]

```
>UniRef50_A1B9S1 TonB-dependent heme/hemoglobin receptor family
MPRHSIRGALLAGTACLTALTFTAPLLAQERAGADSAQSTYVLDQIVLRAGKPKVASEVP
QSVSVVDSRQLEDIAPIHIGEVLATVPGVAGVGSGSFFGQGFNIRGFGSSGAAASESGIV
QLIDGEEKYYESYRQGALFVEPDFLRQVEVLRGPGSSTLYGSGALGGVIAMETIEAGDLI
AEGQTFGGRTKLGYASNPDTVLGSVALGWRPAEDFEALAAFAWRKLGDTKDADGNTTVRA
NSKTPNLLLKAKKTFGDQYVAFSYQHLEAKGDDQDFNQLEGAQVGLFPGFPGWGVGDITT
RDQTARFIWGWNPEDNRYVDLTATLSYTNTLKDVRQGDDPDEPIMDSLLGERDYRLWKFK
LSNVADLSGAGYDHHLTTGAEVLKQDRSSSVPSSSHPEAYTRAWAAYALSELTWGDLTIN
SGLRYEKQRTEPKSSVTVTDDTYDADSVEPQVAAIYRLNDSLSVFGSVAFVNRMPTVDEL
YDSFMGGAPSGDLKDEKGKNIELGLSYHGSGILTASDEAVVKLTLFRNHIDDMIVRTNAP
APMPAYVNIDRAYLRGGELEATYSVAAWEFGAAFSVVNGVDQDGADLDTLPNNRVTLQAI
WQASDALRLGLRSTLADGRDKPNGTHRAGYGVHDVFATWVPQGGAAAGIEVHVGVDNVTD
RDYTPATWLSGPAPGRNFKLSVSRSF
>UniRef50_A1U6I3 TonB-dependent heme/hemoglobin receptor family
MANSSPMKQPRRFRRNTLWLALMAAPLAHAQPVSLDPIQVTADREADADTVVDAETIERF
QADDLEDVFAGQPDVSVGGSNSIAQKVYVRGFDDPLLNVSVDGATQAGALFHHSGRLSVE
PELLKQVEVNAGAGRATEGAGALGGSIRFVTKDPDDLLRPGESAGALVKFGSFSNTDGYK
ASGTAFGRLSDNWSTLVSVSQSDHEPFKDGSGDRIAGSDARQQLGFAKLVGQLPADQTIK
LSHEVRTDEGERPQRPQWVVSSFNRLYELDGRRDTTTLNYGYAPAGNALVDLEATVYHTE
SDIEQNVEDRWGRYFGFSRNIGGDLRNTSRFGGHSLTYGVDYREDKVNAGYQEDKRAEQQ
TGEVLGVYLQGDLWLTSRLLFSAGARYDDYRLKDNDDQRFSEDEVSPNANLAWEVVDGLT
LKAGYAEAFRGPTTQDAFKLEGSENDPDLEGEKARNTEVGFDYRYETFRLSAEVYRSEIK
DAIADPLLPFRESIYKNIGDLESDGYLISAGYQWQALSAGLSFHSNDAEVDGQPLTVYEH
NLLGNTMGDTWIADLAYRWDRNLEFGWQGRFVEGIDNLDTSVGTIDKPGYGVHDLYLHWL
PTGNEDLRLSLTIKNVGDKQYLAHASNADYQHIEDYEGIVGMPEPGRDIRVGLAMRF
```

# Question

**Obtain the sequences of "transcription factors" with less than 50% sequence identity.**

**Find the amino acid sequence of human mitochondrial beta barrel membrane protein VDAC**

# Dataset for transcription factors

UniProt › UniProtKB  Downloads · Contact · Documentation/Help

Search in: Protein Knowledgebase (UniProtKB)  Query: transcription factors  Search  Clear  Fields »

Search | Blast | Align | Retrieve | ID Mapping *

1 - 25 of 10,963 results for **transcription** AND **factors** in **UniProtKB** sorted by **score** descending

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50% | Customize display  Download...

› Show only reviewed (UniProtKB/Swiss-Prot) or unreviewed (UniProtKB/TrEMBL) entries
› Quote terms: "transcription factors"
› Restrict term "transcription" to protein family, gene ontology, keyword, protein name
› Restrict term "factors" to protein family, gene ontology, protein name, web resource

Page 1 of 439 | Next »

| All | Accession | Entry name | Status | Protein names | Gene names | Organism | Length |
|---|---|---|---|---|---|---|---|
| ☑ | P0A4H2 | BVGA_BORPE | ★ | **Virulence factors putative positive transcription regulator bvgA** | **bvgA** (BP1878) | Bordetella pertussis | 209 |
| ☑ | P0A4H3 | BVGA_BORBR | ★ | **Virulence factors putative positive transcription regulator bvgA** | **bvgA** (BB2994) | Bordetella bronchiseptica (Alcaligenes bronchisepticus) | 209 |
| ☑ | P0A4H4 | BVGA_BORPA | ★ | **Virulence factors putative positive transcription regulator bvgA** | **bvgA** (BPP3028) | Bordetella parapertussis | 209 |
| ☑ | P18850 | ATF6A_HUMAN | ★ | **Cyclic AMP-dependent transcription factor ATF-6 alpha** (cAMP-dependent transcription factor ATF-6 alpha) (Activating transcription factor 6 alpha) (ATF6-alpha) [Cleaved into: Processed cyclic AMP-dependent transcription factor ATF-6 alpha] | **ATF6** | Homo sapiens (Human) | 670 |
| ☑ | A6H8Y1 | BDP1_HUMAN | ★ | **Transcription factor TFIIIB component B" homolog** (Transcription factor IIIB 150) (TFIIIB150) (Transcription factor-like nuclear regulator) | **BDP1** (KIAA1241) (KIAA1689) (TFNR) | Homo sapiens (Human) | 2,624 |
| ☑ | Q04636 | POB3_YEAST | ★ | **FACT complex subunit POB3** (Facilitates chromatin transcription complex subunit POB3) | **POB3** (YML069W) | Saccharomyces cerevisiae (Baker's yeast) | 552 |

```
>sp|P18485|1A12_SOLLC 1-aminocyclopropane-1-carboxylate synthase 2 OS=Solanum lycopersicum GN=ACS2 PE=1 SV=2
MGFEIAKTNSILSKLATNEEHGENSPYFDGWKAYDSDPFHPLKNPNGVIQMGLAENQLCL
DLIEDWIKRNPKGSICSEGIKSFKAIANFQDYHGLPEFRKAIAKFMEKTRGGRVRFDPER
VVMAGGATGANETIIFCLADPGDAFLVPSPYYPAFNRDLRWRTGVQLIPIHCESSNNFKI
TSKAVKEAYENAQKSNIKVKGLILTNPSNPLGTTLDKDTLKSVLSFTNQHNIHLVCDEIY
AATVFDTPQFVSIAEILDEQEMTYCNKDLVHIVYSLSKDMGLPGFRVGIIYSFNDDVVNC
ARKMSSFGLVSTQTQYFLAAMLSDEKFVDNFLRESAMRLGKRHKHFTNGLEVVGIKCLKN
NAGLFCWMDLRPLLRESTFDSEMSLWRVIINDVKLNVSPGSSFECQEPGWFRVCFANMDD
GTVDIALARIRRFVGVEKSGDKSSSMEKKQQWKKNNLRLSFSKRMYDESVLSPLSSPIPP
SPLVR
>sp|P16375|7UP1_DROME Steroid receptor seven-up, isoforms B/C OS=Drosophila melanogaster GN=svp PE=1 SV=1
MCASPSTAPGFFNPRPQSGAELSAFDIGLSRSMGLGVPPHSAWHEPPASLGGHLHAASAG
PGTTTGSVATGGGGTTPSSVASQQSAVIKQDLSCPSLNQAGSGHHPGIKEDLSSSLPSAN
GGSAGGHHSGSGSGSGSGSGVNPGHGSDMLPLIKGHGQDMLTSIKGQPTGCGSTTPSSQANS
SHSQSSNSGSQIDSKQNIECVVCGDKSSGKHYGQFTCEGCKSFFKRSVRRNLTYSCRGSR
NCPIDQHHRNQCQYCRLKKCLKMGMRREAVQRGRVPPTQPGLAGMHGQYQIANGDPMGIA
GFNGHSYLSSYISLLLRAEPYPTSRYGQCMQPNNIMGIDNICELAARLLFSAVEWAKNIP
FFPELQVTDQVALLRLVWSELFVLNASQCSMPLHVAPLLAAAGLHASPMAADRVVAFMDH
IRIFQEQVEKLKALHVDSAEYSCLKAIVLFTTDACGLSDVTHIESLQEKSQCALEEYCRT
QYPNQPTRFGKLLLRLPSLRTVSSQVIEQLFFVRLVGKTPIETLIRDMLLSGNSFSWPYL
PSM
>sp|P16376|7UP2_DROME Steroid receptor seven-up, isoform A OS=Drosophila melanogaster GN=svp PE=2 SV=3
MCASPSTAPGFFNPRPQSGAELSAFDIGLSRSMGLGVPPHSAWHEPPASLGGHLHAASAG
PGTTTGSVATGGGGTTPSSVASQQSAVIKQDLSCPSLNQAGSGHHPGIKEDLSSSLPSAN
GGSAGGHHSGSGSGSGSGSGVNPGHGSDMLPLIKGHGQDMLTSIKGQPTGCGSTTPSSQANS
SHSQSSNSGSQIDSKQNIECVVCGDKSSGKHYGQFTCEGCKSFFKRSVRRNLTYSCRGSR
NCPIDQHHRNQCQYCRLKKCLKMGMRREAVQRGRVPPTQPGLAGMHGQYQIANGDPMGIA
GFNGHSYLSSYISLLLRAEPYPTSRYGQCMQPNNIMGIDNICELAARLLFSAVEWAKNIP
FFPELQVTDQVALLRLVWSELFVLNASQCSMPLHVAPLLAAAGLHASPMAADRVVAFMDH
IRIFQEQVEKLKALHVDSAEYSCLKAIVLFTTGKLLDILYKDVPALLTKVSALLGKGSTA
SNDDVLAVVRDHLDELNRQEQESQAQQQAPLHLAAFMNCVAGVEAAVQQAEQAQVPTSSA
SASVSAPLVPSAGSAFSSCQAKSAGSEMDLLASLYAQAQATPPSSGGGDASGHNNSSGLG
ASLPTQSQSGSSSRNLTASPLSTSLATAPAPASASAPAPVPTSSVAQVPVPAPVPVTSSA
SSSSLGGGAYQTPSAAAAAAAMFHYQTPPRAAFGSAFDMFHHSTPFGVGVGHAHALAHSS
GSGSASFGSPSYRYSPYSLAGSRWQL
```

**Search in**

Protein Knowledgebase (UniProtKB)

**Query**

```
mitochondrial beta barrel membrane protein
AND human and VDAC
```

Search | Advanced Search » | Clear

3 results for **mitochondrial** ⊠ **AND beta** ⊠ **AND barrel** ⊠ **AND membrane** ⊠ **AND protein** ⊠ **AND human** ⊠ **AND VDAC** ⊠ in **UniProtKB** sorted by **score** descending ⊠

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50% | Download

Page 1 of 1

## Results Customize

› Restrict term "mitochondrion" to protein family (3), gene ontology (3), keyword (3), location (3), protein name (3)
› Restrict term "beta" to keyword (3), annotation topic (1)
› Restrict term "membrane" to gene ontology (3), keyword (3), location (3), protein name (3)
› Restrict term "protein" to gene ontology (3), keyword (2), protein name (3)
› Restrict term "mitochondrion" to pathway
› Restrict term "beta" to pathway
› Restrict term "membrane" to pathway
› Restrict term "protein" to pathway

| | Accession | Entry name | Status | Protein names | Gene names | Organism | Length |
|---|---|---|---|---|---|---|---|
| ☐ | P21796 | VDAC1_HUMAN | ★ | **Voltage-dependent anion-selective channel pro...** | **VDAC1** VDAC | Homo sapiens (Human) | 283 |
| ☐ | P45880 | VDAC2_HUMAN | ★ | **Voltage-dependent anion-selective channel pro...** | **VDAC2** | Homo sapiens (Human) | 294 |
| ☐ | Q9Y277 | VDAC3_HUMAN | ★ | **Voltage-dependent anion-selective channel pro...** | **VDAC3** | Homo sapiens (Human) | 283 |

## Sequences

| Sequence | | Length | Mass (Da) | Tools |
|---|---|---|---|---|
| P21796 [UniParc]. | FASTA | 283 | 30,773 | Blast go |

Last modified January 23, 2007. Version 2.
Checksum: 89BA3378B04020D5

```
        10         20         30         40         50         60
MAVPPTYADL GKSARDVFTK GYGFGLIKLD LKTKSENGLE FTSSGSANTE TTKVTGSLET

        70         80         90        100        110        120
KYRWTEYGLT FTEKWNTDNT LGTEITVEDQ LARGLKLTFD SSFSPNTGKK NAKIKTGYKR

       130        140        150        160        170        180
EHINLGCDMD FDIAGPSIRG ALVLGYEGWL AGYQMNFETA KSRVTQSNFA VGYKTDEFQL

       190        200        210        220        230        240
HTNVNDGTEF GGSIYQKVNK KLETAVNLAW TAGNSNTRFG IAAKYQIDPD ACFSAKVNNS

       250        260        270        280
SLIGLGYTQT LKPGIKLTLS ALLDGKNVNA GGHKLGLGLE FQA
```

```
>sp|P21796|VDAC1_HUMAN Voltage-dependent anion-selective channel protein 1 OS=Homo sapiens GN=VDAC1 PE=1 SV=2
MAVPPTYADLGKSARDVFTKGYGFGLIKLDLKTKSENGLEFTSSGSANTETTKVTGSLET
KYRWTEYGLTFTEKWNTDNTLGTEITVEDQLARGLKLTFDSSFSPNTGKKNAKIKTGYKR
EHINLGCDMDFDIAGPSIRGALVLGYEGWLAGYQMNFETAKSRVTQSNFAVGYKTDEFQL
HTNVNDGTEFGGSIYQKVNKKLETAVNLAWTAGNSNTRFGIAAKYQIDPDACFSAKVNNS
SLIGLGYTQTLKPGIKLTLSALLDGKNVNAGGHKLGLGLEFQA
```

# Pairwise alignment

**Alignment between two or more protein or nucleic acid sequences represents an explicit hypothesis regarding the evolutionary history of those sequences. As a direct result, it facilitates many recent advances in understanding the information content and function of genetic sequences**

>sp|P68871|HBB_HUMAN Hemoglobin subunit beta OS=Homo sapiens
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPD
AVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFR
LLGNVLVCVLAHHFG KEFTPPVQAAYQKVVAGVANALAHKYH

>sp|P02112|HBB_CHICK Hemoglobin subunit beta OS=Gallus gallus
VHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSP
TAILGNPMVRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENF
RLLGDILIIVLAAHFS KDFTPECQAAWQKLVRVVAHALARKYH

# Sequence comparison

| Different types of pairwise comparisons | |
|---|---|
| ***Method name*** | ***Situation*** |
| Dot plot | **General exploration of your sequence**<br>Discovering repeats<br>Finding long insertions and deletions<br>Extracting portions of sequences to make a multiple alignment |
| Local alignments | **Comparing sequences with partial homology**<br>Making high-quality alignments<br>Making residue-per-residue analysis |
| Global alignments | **Comparing two sequences over their entire length**<br>Identifying long insertions and deletions<br>Checking the quality of your data<br>Identifying every mutation in your sequences |

# Dot Plots

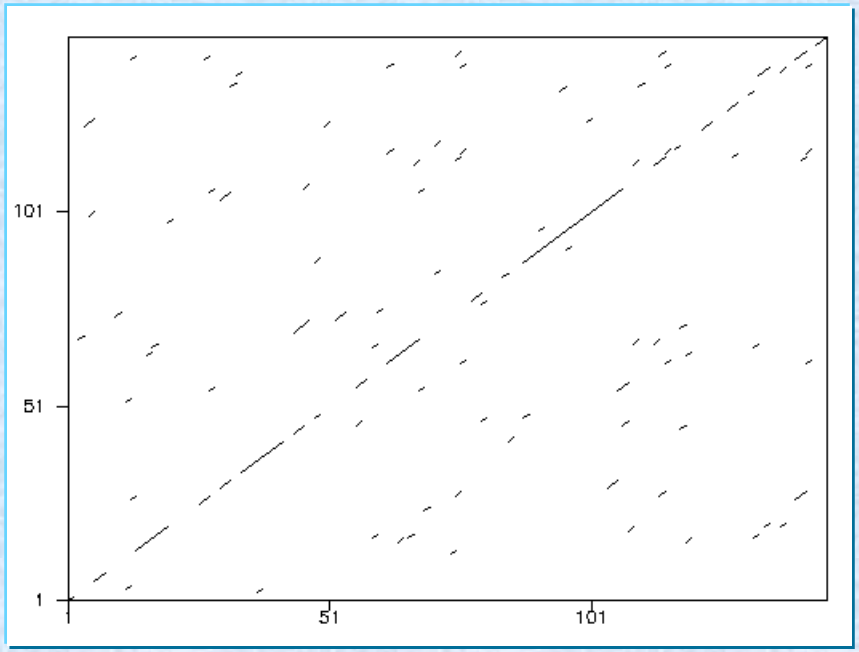
**One of the simplest methods for evaluating similarity between two sequences is to visualize regions of similarity using dot plots.**

**To construct a simple dot plot, the first sequence to be compared is assigned to the horizontal axis of a plot space and the second is then assigned to the vertical axis.**

**Dots are then placed in the plot space at each position where both of the sequence elements are identical.**

**Adjacent regions of identity between two sequences give rise to diagonal lines of dots in the plot.**

**Human and chicken hemoglobin**




**VHLTPEEKSAVTALWGKVNV**

**VHWTAEEKQLITGLWGKVNV**

Window size and cutoff (or threshold from substitution matrix)
E.g. 5 and 3

# Simple alignment

An alignment between two sequences is simply a pairwise match between the characters in each sequence.

A true alignment of nucleotide or amino acid sequences is one that reflects evolutionary relationship between two or more homologues (sequences that share a common ancestor).

The three kinds of changes between sequences are

(i) a mutation that replaces one character with another

(ii) an insertion that adds one or more positions

(iii) a deletion that deletes one or more positions

Insertions and deletions occur less frequently than mutations. In these cases, gaps in alignments are commonly added.

## Examples for mutation, insertion and deletion

# Example

## AATCTATA and AAGATA

**Dot plots are useful for visual inspection of the regions of similarity between sequences.**

**A numeric system for evaluating sequence similarity has obvious advantages for objective determination of optimal alignments.**

**Scoring alignment without gap: credit for aligned residues and penalty for mismatches**

```
AATCTATA        AATCTATA        AATCTATA
AAGATA          AAGATA          AAGATA
```

**match score =1, if seq1(i) = seq2(i)**

**match score = 0, if seq1(i)  seq2(i)**

Identity matrix

| | A | T | C | G |
|---|---|---|---|---|
| A | 1 | 0 | 0 | 0 |
| T | 0 | 1 | 0 | 0 |
| C | 0 | 0 | 1 | 0 |
| G | 0 | 0 | 0 | 1 |

**The alignment scores are 4, 1, 3**

# Gaps

The insertions and deletions complicate sequence alignments by increasing the number of possible alignments between two or more sequences.

E.g. AATCTATA and AAGATA can be aligned in only three different ways.

These sequences can be aligned in 28 different ways using gaps.

```
AATCTATA        AATCTATA        AATCTATA
AAG-AT-A        AA-G-ATA        AAG--ATA
```

An alignment that includes gaps, an additional term, the gap penalty must be included in the scoring function.

Gap penalty = -1, if seq1(i) = "-" or seq2(i) = "-"

match score =1, if seq1(i) = seq2(i)

match score = 0, if seq1(i)  seq2(i)

The scores are 1, 3, 3

# Sequence alignment

Position

Alignment 1

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Seq A: | V | – | E | I | T | G | E | I | S | T |
| Seq B: | P | R | E | – | T | E | R | I | – | T |
| Score: | 0 | –1 | 1 | –1 | 1 | 0 | 0 | 1 | –1 | 1 |
| Total: | 1 | | | | | | | | | |

Alignment 2

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Seq A: | V | E | I | T | G | E | I | S | T |
| Seq B: | P | R | E | T | – | E | R | I | T |
| Score: | 0 | 0 | 0 | 1 | –1 | 1 | 0 | 0 | 1 |
| Total: | 2 | | | | | | | | |

Alignment 3

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Seq A: | – | V | E | I | T | G | E | – | I | S | T |
| Seq B: | P | R | E | – | T | – | E | R | I | – | T |
| Score: | –1 | 0 | 1 | –1 | 1 | –1 | 1 | –1 | 1 | –1 | 1 |
| Total: | 0 | | | | | | | | | | |

VEITGEIST
PRETERIT

# Origination and length penalties

One method to further distinguish between alignment is to differentiate between the alignments that contain many isolated gaps and those that contain fewer, but longer, sequence of gaps. E.g.

Consider two arbitrary sequences of lengths 12 and 9. Any alignment will have a shortage of 3 gaps in the shorter sequence.

Assuming the two sequences are homologous from the beginning to the end, then the differences are due to insertions in the longer or deletions in the shorter or both. Such events are called indel (insertion/deletion).

Considering this situation, we can bias our alignment scoring function to reward alignments that are more likely form an evolutionary perspective.

Gap penalty is divided into origination penalty (for starting new series of gaps on one of the sequences being aligned) and length penalty (number of sequential missing characters)

# Origination and length penalties

| | | |
|---|---|---|
| AATCTATA | AATCTATA | AATCTATA |
| AAG-AT-A | AA-G-ATA | AAG--ATA |

**E.g. Origination penalty: -2; length penalty: -1, match score: +1 and mismatch score: 0**

**Case 1: 2x-2 + 2x-1 + 3x1 + 3x0 = -4 -2 +3 + 0 = -3**

**Case 2: -1**

**Case 3: 1x-2 + 2x-1 + 5x1 + 1x0 = -2 -2 + 5 + 0 = +1**

**Case 2 and case 3 are different; they were same in previous scoring method**

# Scoring matrices

In previous alignment, for non-gap positions scores are given as 1 for match and 0 for mismatch.

These scores can be further refined based on the substitutions.

For example Alanine is replaced with Valine or Lysine

Once the alignment score for each possible pair of nucleotides/amino acid residues are determined, the resulting scoring matrix is used to score each non-gap position in the alignment.

For nucleotides scoring is simple:

In BLAST, same nucleotides are given a score of +5 and different ones have -4

Case 2: matching nucleotides: mild reward (+1)

Transitions (purine to purine, A or G)/ pyramidine to pyramidine (C or T): mild penalty (-1)

Transversions (purine to pyramidine or vice versa); severe penalty (-5)

**BLAST matrix**

| | A | T | C | G |
|---|---|---|---|---|
| A | 1 | 0 | 0 | 0 |
| T | 0 | 1 | 0 | 0 |
| C | 0 | 0 | 1 | 0 |
| G | 0 | 0 | 0 | 1 |

**Transition Transversion matrix**

| | A | T | C | G |
|---|---|---|---|---|
| A | 1 | -5 | -5 | -1 |
| T | -5 | 1 | -1 | -5 |
| C | -5 | -1 | 1 | -5 |
| G | -1 | -5 | -5 | 1 |

# Scoring matrix: amino acids

Different criteria can be considered when devising a scoring matrix for amino acid sequence alignments

Most common ones are based on observed physical/chemical similarity and observed substitution frequencies

E.g. pairing two amino acids that both have aromatic functional groups might receive a good positive score,

pairing an amino acid that has a nonpolar functional group with one that has a charged functional group might result in a scoring penalty.

Scoring matrices have been derived based on residue hydrophobicity, charge and size

Another option is based on genetic code: minimum number of nucleotide substitutions are necessary to convert a codon from one residue to other

# Scoring matrix: amino acids

A common method for deriving scoring matrices is to observe the actual substitution rates among various amino acid residues in nature.

If substitution between two amino acid residues is observed frequently, then positions in which these residues are aligned favorably.

Likewise, alignments between residues that are not observed to interchange frequently in natural evolution is penalized.

One commonly used scoring matrix based on observed substitution rates is the point accepted mutation (PAM) matrix.

The scores in a PAM matrix are computed by observing the substitutions that occur in alignments between similar sequences.

# Development of PAM matrix

1. Alignment is constructed with very high sequence identity (usually >85%).

2. The relative mutatbility, $m_j$, for each amino acid is computed. It is the number of times the amino acid was substituted by any other amino acids. E.g. Ala to others

3. Pair of amino acids, $A_{ij}$, the number of times amino acid j was replaced by amino acid i, tallied for each amino acid pairs i and j. E.g. $A_{cm}$ is the number of time Met is replaced with cysteine.

4. The substitution tallies are divided by relative mutability.

5. Normalize with the frequency of occurrence of each amino acid

6. Take log of each resulting entries in the PAM-1 matrix (PAM-1 means 1 substitution per 100 residues or 1 PAM unit) . This matrix is also called log odds matrix, since the entries are based on the log of the substitution probability for each amino acid.

PAM-1 matrix is appropriate to compare sequences are closely related. PAM-1000 matrix might be used to compare sequences with distant relationships. Usually PAM-250 is used for sequence alignment.

# Calculation of a PAM matrix

PAM matrix is a 20x20 matrix for all pairs

Assumption:

Substitutions are equal is both directions (A to G and G to A)

E.g.: Element GA

Frequency of pairs, $F_{G,A} = 3$

Relative mutability, $m_A = 4$

Normalizing factor = number of mutations in the entire tree times 2, times relative frequency of A residues multiplied by 100 (1 substitution per 100 residues)

i.e., 6 x 2 x (10/63) x 100 = 190.4762

Hence, normalized relative mutability,

$m_A = 4/190.4762 = 0.021$

Consider a multiple sequence alignment

```
1.ACGCTAFKI
2.GCGCTAFKI (1: A  G)
3.ACGCTAFKL (1: I  L
4.GCGCTGFKI (2: A  G)
5.GCGCTLFKI (2: A  L)
6.ASGCTAFKL (3: C  S)
7.ACACTAFKL (3: G  A)
```

Construct tree

# Calculation of a PAM matrix

$m_A = 4/190.4762 = 0.021$

Mutation probability, $M_{ij} = m_j F_{ij} / f_{ij}$

$M_{G,A} = 0.021 \times 3 / 4$

$= 0.0157$

$f_{ij}$, total number of substitutions involving A

$R_{ij} = \log(M_{ij}/f_i) = \log(M_{GA}/f_G)$

$f_G = 10/63 = 0.1587$

$R_{GA} = \log(0.0157/0.1587) = \log(0.0989)$

$R_{GA} = -1.005$

Repeat for all off-diagonal elements.

For diagonal elements: $M_{jj} = 1 - m_j$

Calculate $R_{jj}$

Consider a multiple sequence alignment

```
1.ACGCTAFKI
2.GCGCTAFKI (1: A  G)
3.ACGCTAFKL (1: I  L
4.GCGCTGFKI (2: A  G)
5.GCGCTLFKI (2: A  L)
6.ASGCTAFKL (3: C  S)
7.ACACTAFKL (3: G  A)
```

Calculate the element $R_{AA}$

Calculate the element $R_{IL}$

# PAM-120 mutation matrix

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

# PAM 250 mutation matrix

| | Cys | Gly | Pro | Ser | Ala | Thr | Asp | Glu | Asn | Gln | His | Lys | Arg | Val | Met | Ile | Leu | Phe | Tyr | Trp |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cys | 12 | | | | | | | | | | | | | | | | | | | |
| Gly | -3 | 5 | | | | | | | | | | | | | | | | | | |
| Pro | -3 | -1 | 6 | | | | | | | | | | | | | | | | | |
| Ser | 0 | 1 | 1 | 1 | | | | | | | | | | | | | | | | |
| Ala | -2 | 1 | 1 | 1 | 2 | | | | | | | | | | | | | | | |
| Thr | -2 | 0 | 0 | 1 | 1 | 3 | | | | | | | | | | | | | | |
| Asp | -5 | 1 | -1 | 0 | 0 | 0 | 4 | | | | | | | | | | | | | |
| Glu | -5 | 0 | -1 | 0 | 0 | 0 | 3 | 4 | | | | | | | | | | | | |
| Asn | -4 | 0 | -1 | 1 | 0 | 0 | 2 | 1 | 2 | | | | | | | | | | | |
| Gln | -5 | -1 | 0 | -1 | 0 | -1 | 2 | 2 | 1 | 4 | | | | | | | | | | |
| His | -3 | -2 | 0 | -1 | -1 | -1 | 1 | 1 | 2 | 3 | 6 | | | | | | | | | |
| Lys | -5 | -2 | -1 | 0 | -1 | 0 | 0 | 0 | 1 | 1 | 0 | 5 | | | | | | | | |
| Arg | -4 | -3 | 0 | 0 | -2 | -1 | -1 | -1 | 0 | 1 | 2 | 3 | 6 | | | | | | | |
| Val | -2 | -1 | -1 | -1 | 0 | 0 | -2 | -2 | -2 | -2 | -2 | -2 | -2 | 4 | | | | | | |
| Met | -5 | -3 | -2 | -2 | -1 | -1 | -3 | -2 | 0 | -1 | -2 | 0 | 0 | 2 | 6 | | | | | |
| Ile | -2 | -3 | -2 | -1 | -1 | 0 | -2 | -2 | -2 | -2 | -2 | -2 | -2 | 4 | 2 | 5 | | | | |
| Leu | -6 | -4 | -3 | -3 | -2 | -2 | -4 | -3 | -3 | -2 | -2 | -3 | -3 | 2 | 4 | 2 | 6 | | | |
| Phe | -4 | -5 | -5 | -3 | -4 | -3 | -6 | -5 | -4 | -5 | -2 | -5 | -4 | -1 | 0 | 1 | 2 | 9 | | |
| Tyr | 0 | -5 | -5 | -3 | -3 | -3 | -4 | -4 | -2 | -4 | 0 | -4 | -5 | -2 | -2 | -1 | -1 | 7 | 10 | |
| Trp | -8 | -7 | -6 | -2 | -6 | -5 | -7 | -7 | -4 | -5 | -3 | -3 | 2 | -6 | -4 | -5 | -2 | 0 | 0 | 17 |

# Scoring matrix: amino acids

**Different criteria can be considered when devising a scoring matrix for amino acid sequence alignments**

**Most common ones are based on observed physical/chemical similarity and observed substitution frequencies**

**E.g. pairing two amino acids that both have aromatic functional groups might receive a good positive score,**

**pairing an amino acid that has a nonpolar functional group with one that has a charged functional group might result in a scoring penalty.**

**Scoring matrices have been derived based on residue hydrophobicity, charge and size**

**Another option is based on genetic code: minimum number of nucleotide substitutions are necessary to convert a codon from one residue to other**

# Scoring matrix: amino acids

A common method for deriving scoring matrices is to observe the actual substitution rates among various amino acid residues in nature.

If substitution between two amino acid residues is observed frequently, then positions in which these residues are aligned favorably.

Likewise, alignments between residues that are not observed to interchange frequently in natural evolution is penalized.

One commonly used scoring matrix based on observed substitution rates is the point accepted mutation (PAM) matrix.

The scores in a PAM matrix are computed by observing the substitutions that occur in alignments between similar sequences.

# Development of PAM matrix

1. Alignment is constructed with very high sequence identity (usually >85%).

2. The relative mutatbility, $m_j$, for each amino acid is computed. It is the number of times the amino acid was substituted by any other amino acids. E.g. Ala to others

3. Pair of amino acids, $A_{ij}$, the number of times amino acid j was replaced by amino acid i, tallied for each amino acid pairs i and j. E.g. $A_{cm}$ is the number of time Met is replaced with cysteine.

4. The substitution tallies are divided by relative mutability.

5. Normalize with the frequency of occurrence of each amino acid

6. Take log of each resulting entries in the PAM-1 matrix (PAM-1 means 1 substitution per 100 residues or 1 PAM unit) . This matrix is also called log odds matrix, since the entries are based on the log of the substitution probability for each amino acid.

PAM-1 matrix is appropriate to compare sequences are closely related. PAM-1000 matrix might be used to compare sequences with distant relationships. Usually PAM-250 is used for sequence alignment.

# Calculation of a PAM matrix

**PAM matrix is a 20x20 matrix for all pairs**

**Assumption:**

**Substitutions are equal is both directions (A to G and G to A)**

**E.g.: Element GA**

**Frequency of pairs, $F_{G,A} = 3$**

**Relative mutability, $m_A = 4$**

**Normalizing factor = number of mutations in the entire tree times 2, times relative frequency of A residues multiplied by 100 (1 substitution per 100 residues)**

**i.e., 6 x 2 x (10/63) x 100 = 190.4762**

**Hence, normalized relative mutability,**

**$m_A = 4/190.4762 = 0.021$**

**Consider a multiple sequence alignment**

```
1.ACGCTAFKI
2.GCGCTAFKI (1: A  G)
3.ACGCTAFKL (1: I  L
4.GCGCTGFKI (2: A  G)
5.GCGCTLFKI (2: A  L)
6.ASGCTAFKL (3: C  S)
7.ACACTAFKL (3: G  A)
```

Construct tree

# Calculation of a PAM matrix

$m_A = 4/190.4762 = 0.021$

Mutation probability, $M_{ij} = m_j F_{ij} / f_{ij}$

$M_{G,A} = 0.021 \times 3 / 4$

$= 0.0157$

$f_{ij}$, total number of substitutions involving A

$R_{ij} = \log(M_{ij}/f_i) = \log(M_{GA}/f_G)$

$f_G = 10/63 = 0.1587$

$R_{GA} = \log(0.0157/0.1587) = \log(0.0989)$

$R_{GA} = -1.005$

Repeat for all off-diagonal elements.

For diagonal elements: $M_{jj} = 1 - m_j$

Calculate $R_{jj}$

Consider a multiple sequence alignment

```
1.ACGCTAFKI
2.GCGCTAFKI (1: A  G)
3.ACGCTAFKL (1: I  L
4.GCGCTGFKI (2: A  G)
5.GCGCTLFKI (2: A  L)
6.ASGCTAFKL (3: C  S)
7.ACACTAFKL (3: G  A)
```

Calculate the element $R_{AA}$

Calculate the element $R_{IL}$

# PAM-120 mutation matrix

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

# PAM 250 mutation matrix

| | Cys | Gly | Pro | Ser | Ala | Thr | Asp | Glu | Asn | Gln | His | Lys | Arg | Val | Met | Ile | Leu | Phe | Tyr | Trp |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cys | 12 | | | | | | | | | | | | | | | | | | | |
| Gly | -3 | 5 | | | | | | | | | | | | | | | | | | |
| Pro | -3 | -1 | 6 | | | | | | | | | | | | | | | | | |
| Ser | 0 | 1 | 1 | 1 | | | | | | | | | | | | | | | | |
| Ala | -2 | 1 | 1 | 1 | 2 | | | | | | | | | | | | | | | |
| Thr | -2 | 0 | 0 | 1 | 1 | 3 | | | | | | | | | | | | | | |
| Asp | -5 | 1 | -1 | 0 | 0 | 0 | 4 | | | | | | | | | | | | | |
| Glu | -5 | 0 | -1 | 0 | 0 | 0 | 3 | 4 | | | | | | | | | | | | |
| Asn | -4 | 0 | -1 | 1 | 0 | 0 | 2 | 1 | 2 | | | | | | | | | | | |
| Gln | -5 | -1 | 0 | -1 | 0 | -1 | 2 | 2 | 1 | 4 | | | | | | | | | | |
| His | -3 | -2 | 0 | -1 | -1 | -1 | 1 | 1 | 2 | 3 | 6 | | | | | | | | | |
| Lys | -5 | -2 | -1 | 0 | -1 | 0 | 0 | 0 | 1 | 1 | 0 | 5 | | | | | | | | |
| Arg | -4 | -3 | 0 | 0 | -2 | -1 | -1 | -1 | 0 | 1 | 2 | 3 | 6 | | | | | | | |
| Val | -2 | -1 | -1 | -1 | 0 | 0 | -2 | -2 | -2 | -2 | -2 | -2 | -2 | 4 | | | | | | |
| Met | -5 | -3 | -2 | -2 | -1 | -1 | -3 | -2 | 0 | -1 | -2 | 0 | 0 | 2 | 6 | | | | | |
| Ile | -2 | -3 | -2 | -1 | -1 | 0 | -2 | -2 | -2 | -2 | -2 | -2 | -2 | 4 | 2 | 5 | | | | |
| Leu | -6 | -4 | -3 | -3 | -2 | -2 | -4 | -3 | -3 | -2 | -2 | -3 | -3 | 2 | 4 | 2 | 6 | | | |
| Phe | -4 | -5 | -5 | -3 | -4 | -3 | -6 | -5 | -4 | -5 | -2 | -5 | -4 | -1 | 0 | 1 | 2 | 9 | | |
| Tyr | 0 | -5 | -5 | -3 | -3 | -3 | -4 | -4 | -2 | -4 | 0 | -4 | -5 | -2 | -2 | -1 | -1 | 7 | 10 | |
| Trp | -8 | -7 | -6 | -2 | -6 | -5 | -7 | -7 | -4 | -5 | -3 | -3 | 2 | -6 | -4 | -5 | -2 | 0 | 0 | 17 |

# BLOSUM matrix

**BLOSUM (Blocks Substitution Matrix) is another popular scoring matrix obtained with statistical clustering techniques.**

**Clustering approach helps to avoid some statistical problems that can occur when the observed substitution rate is very low for a particular pair of amino acids.**

**BLOSUM considers mainly conserved regions.**

**BLOSUM matrices can also be derived for alignments with different sequence identities.**

**Lower numbered PAM matrices are appropriate for comparing closely related sequences.**

**Lower numbered BLOSUM matrices are appropriate for comparing distantly related sequences.**

**E.g. BLOSUM-62 matrix is appropriate for comparing sequences of approximately 62% sequence similarity.**

# BLOSUM-62 matrix

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 4 | −1 | −2 | −2 | 0 | −1 | −1 | 0 | −2 | −1 | −1 | −1 | −1 | −2 | −1 | 1 | 0 | −3 | −2 | 0 |
| R | −1 | 5 | 0 | −2 | −3 | 1 | 0 | −2 | 0 | −3 | −2 | 2 | −1 | −3 | −2 | −1 | −1 | −3 | −2 | −3 |
| N | −2 | 0 | 6 | 1 | −3 | 0 | 0 | 0 | 1 | −3 | −3 | 0 | −2 | −3 | −2 | 1 | 0 | −4 | −2 | −3 |
| D | −2 | −2 | 1 | 6 | −3 | 0 | 2 | −1 | −1 | −3 | −4 | −1 | −3 | −3 | −1 | 0 | −1 | −4 | −3 | −3 |
| C | 0 | −3 | −3 | −3 | 9 | −3 | −4 | −3 | −3 | −1 | −1 | −3 | −1 | −2 | −3 | −1 | −1 | −2 | −2 | −1 |
| Q | −1 | 1 | 0 | 0 | −3 | 5 | 2 | −2 | 0 | −3 | −2 | 1 | 0 | −3 | −1 | 0 | −1 | −2 | −1 | −2 |
| E | −1 | 0 | 0 | 2 | −4 | 2 | 5 | −2 | 0 | −3 | −3 | 1 | −2 | −3 | −1 | 0 | −1 | −3 | −2 | −2 |
| G | 0 | −2 | 0 | −1 | −3 | −2 | −2 | 6 | −2 | −4 | −4 | −2 | −3 | −3 | −2 | 0 | −2 | −2 | −3 | −3 |
| H | −2 | 0 | 1 | −1 | −3 | 0 | 0 | −2 | 8 | −3 | −3 | −1 | −2 | −1 | −2 | −1 | −2 | −2 | 2 | −3 |
| I | −1 | −3 | −3 | −3 | −1 | −3 | −3 | −4 | −3 | 4 | 2 | −3 | 1 | 0 | −3 | −2 | −1 | −3 | −1 | 3 |
| L | −1 | −2 | −3 | −4 | −1 | −2 | −3 | −4 | −3 | 2 | 4 | −2 | 2 | 0 | −3 | −2 | −1 | −2 | −1 | 1 |
| K | −1 | 2 | 0 | −1 | −3 | 1 | 1 | −2 | −1 | −3 | −2 | 5 | −1 | −3 | −1 | 0 | −1 | −3 | −2 | −2 |
| M | −1 | −1 | −2 | −3 | −1 | 0 | −2 | −3 | −2 | 1 | 2 | −1 | 5 | 0 | −2 | −1 | −1 | −1 | −1 | 1 |
| F | −2 | −3 | −3 | −3 | −2 | −3 | −3 | −3 | −1 | 0 | 0 | −3 | 0 | 6 | −4 | −2 | −2 | 1 | 3 | −1 |
| P | −1 | −2 | −2 | −1 | −3 | −1 | −1 | −2 | −2 | −3 | −3 | −1 | −2 | −4 | 7 | −1 | −1 | −4 | −3 | −2 |
| S | 1 | −1 | 1 | 0 | −1 | 0 | 0 | 0 | −1 | −2 | −2 | 0 | −1 | −2 | −1 | 4 | 1 | −3 | −2 | −2 |
| T | 0 | −1 | 0 | −1 | −1 | −1 | −1 | −2 | −2 | −1 | −1 | −1 | −1 | −2 | −1 | 1 | 5 | −2 | −2 | 0 |
| W | −3 | −3 | −4 | −4 | −2 | −2 | −3 | −2 | −2 | −3 | −2 | −3 | −1 | 1 | −4 | −3 | −2 | 11 | 2 | −3 |
| Y | −2 | −2 | −2 | −3 | −2 | −1 | −2 | −3 | 2 | −1 | −1 | −2 | −1 | 3 | −3 | −2 | −2 | 2 | 7 | −1 |
| V | 0 | −3 | −3 | −3 | −1 | −2 | −2 | −3 | −3 | 3 | 1 | −2 | 1 | −1 | −2 | −2 | 0 | −3 | −1 | 4 |

# BLAST

Basic Local Alignment Search Tool

## BLAST: Process the Query Sequence and Database

Divide the query sequence into all "words" of length K=2 (default 3 for proteins)

**Step 1:**
**Hash table for sequence A**

```
Query                    Database
1234567                  123456
QLNFSAGW                 NLNYTPW
QL                       NL
 LN                       LN
  NF                       NY
   FS                       YT
    SA                       TP
     AG                       PW
      GW
```

# BLAST

Basic Local Alignment Search Tool

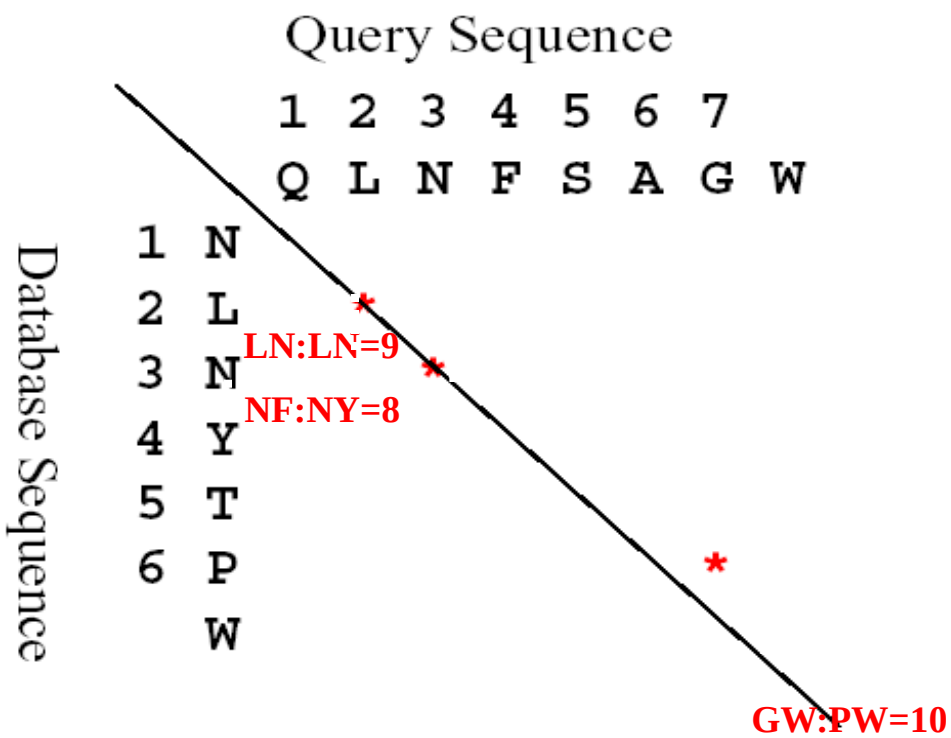
## Identify Word Matches

**Step 2:**

**Use all of the 2-letter words in query sequence to scan against database sequence and mark those with score > 8**

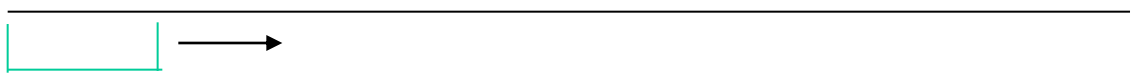
**Note:**

**Marked points can be on the diagonal and off-diagonal**



# BLAST

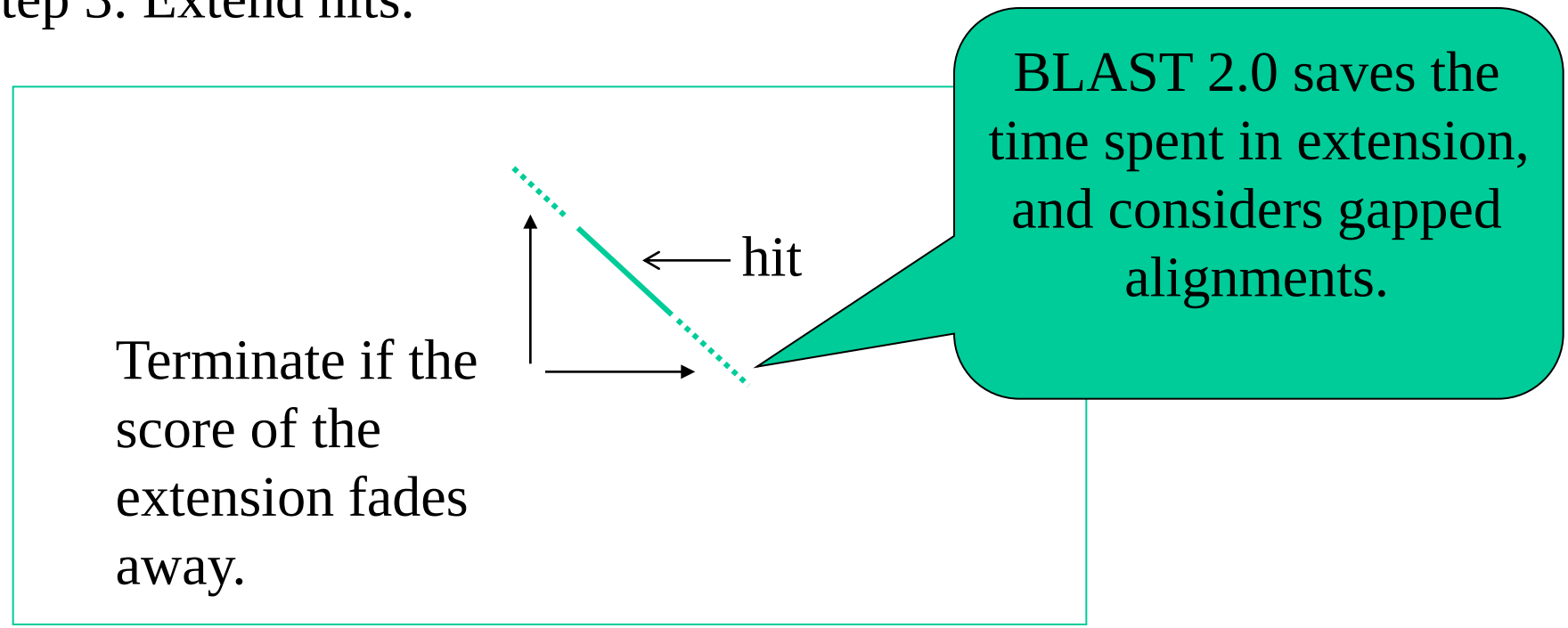
Step2: Scan sequence b for hits.



# BLAST

Step2: Scan sequence b for hits.

Step 3: Extend hits.



# Dynamic programming

Once a method for scoring alignments is selected, an algorithm to find the best alignment between two sequences can be aligned

The most obvious one is exhaustive searching and it is not possible.

E.g. two sequences with 100 and 95 residues, all possible alignments are ~55 million.

It is necessary to develop a smart algorithm

Method of breaking a problem apart into reasonably sized sub problems and using these partial results to compute the final answer: Dynamic programming

```
                 First position      Score  Remaining sequence
                        C              +1          ACGA
                        C                          GA
                 ----------------------------------------
CACGA                   C              -1          ACGA
CGA                     -                          CGA
                 ----------------------------------------
                        -              -1          CACGA
                        C                          GA
                 ----------------------------------------
```

# Gene and Protein Sequence Alignment

***Example:***

Sequence a:  ATTCTTGC

Sequence b: ATCCTATTCTAGC

```
Best Alignment:          ATTCTTGC
                     ATCCTATTCTAGC
                        /|\
                        gap

Bad Alignment:    AT    TCTT     GC
                  ATCCTATTCTAGC
                /|\         /|\
                   gap       gap
```

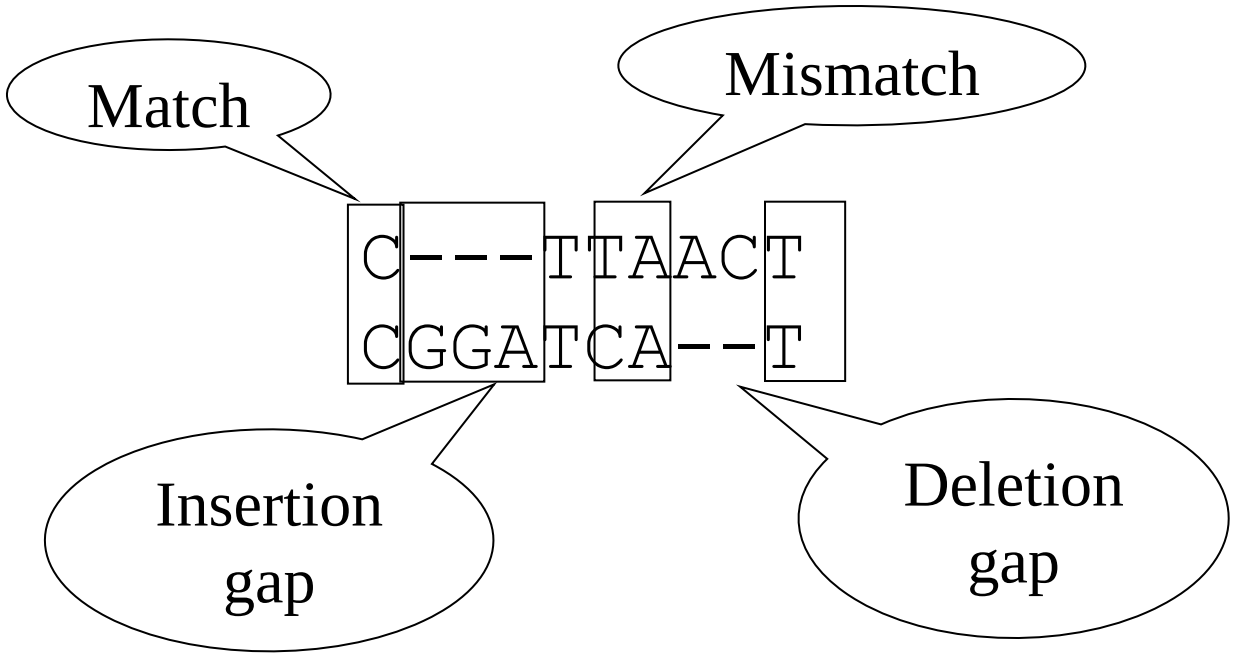
What is a good alignment?

# Pairwise Alignment

Sequence a: CTTAACT

Sequence b: CGGATCAT

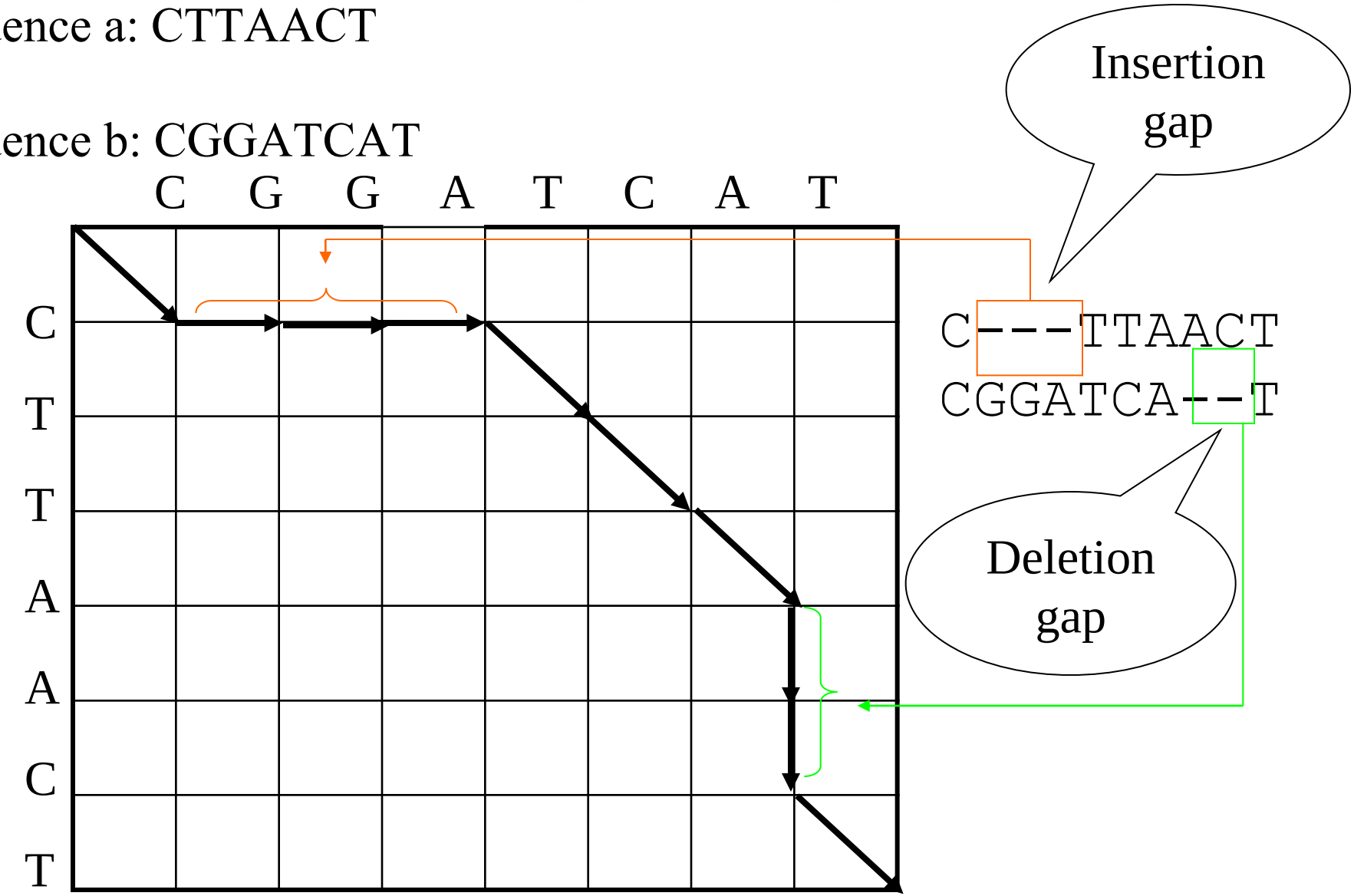
An alignment of a and b:



# Alignment Graph

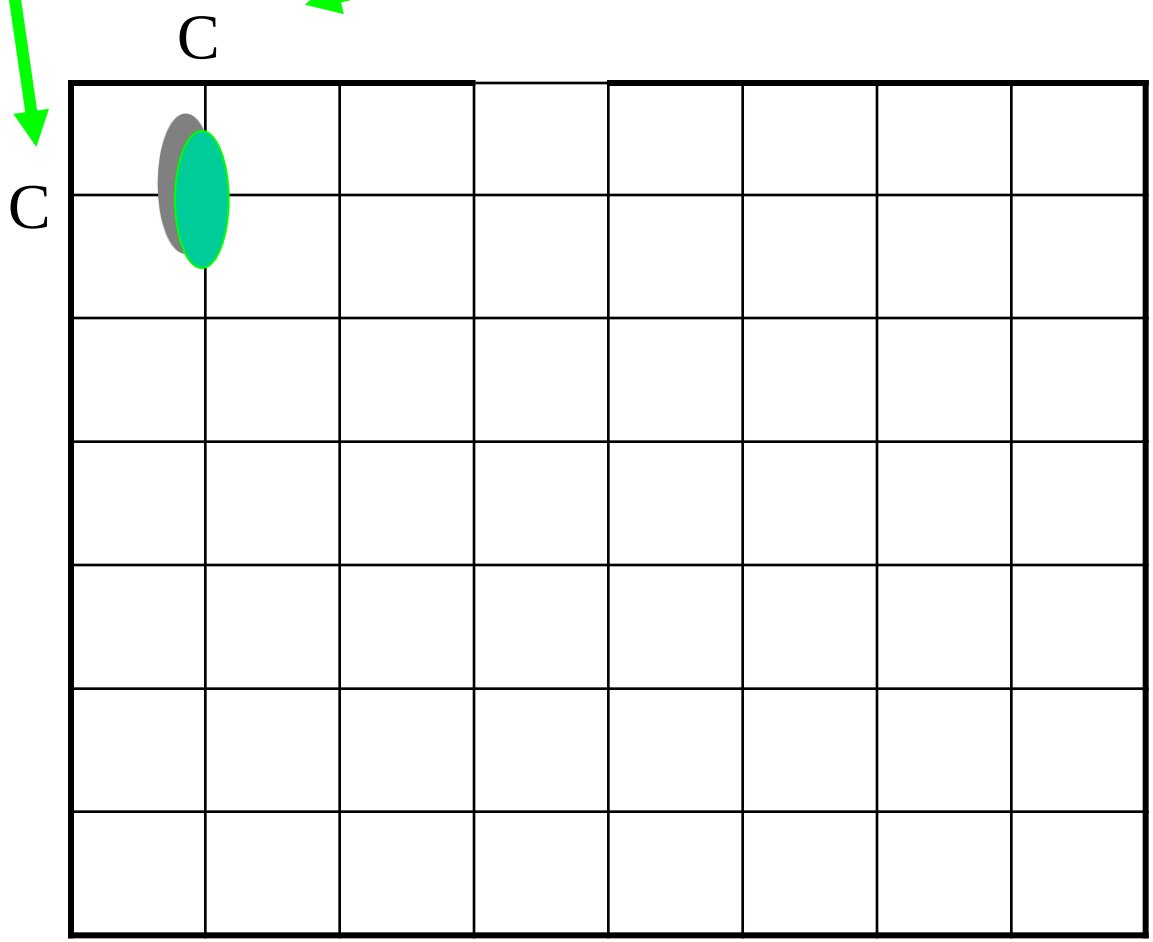
Sequence a: CTTAACT

Sequence b: CGGATCAT



# Graphic representation of an alignment

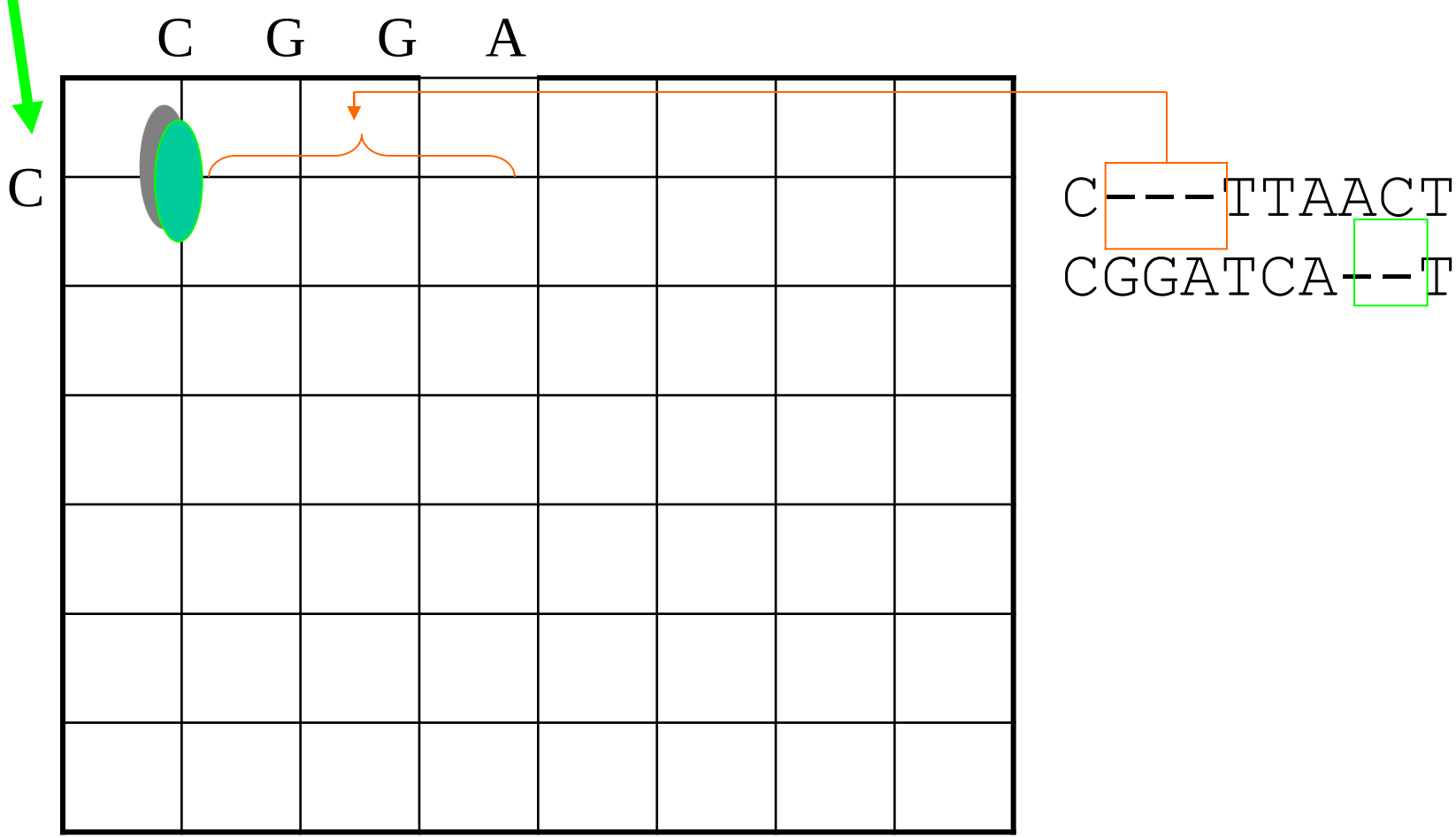
Sequence a: CTTAACT  Sequence b: CGGATCAT




```
C---TTAACT
CGGATCA--T
```

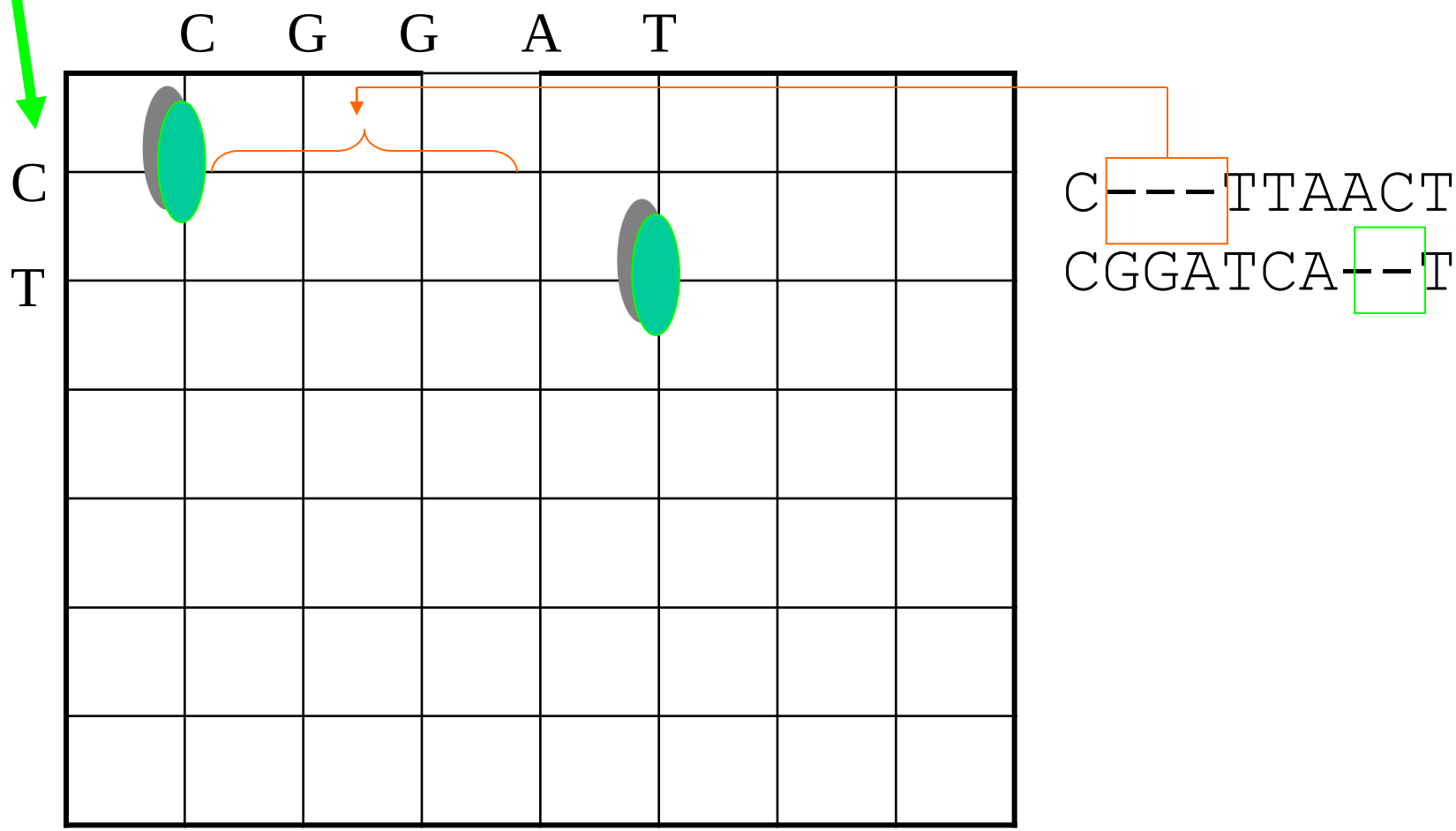
# Graphic representation of an alignment

Sequence a: CTTAACT  Sequence b: CGGATCAT



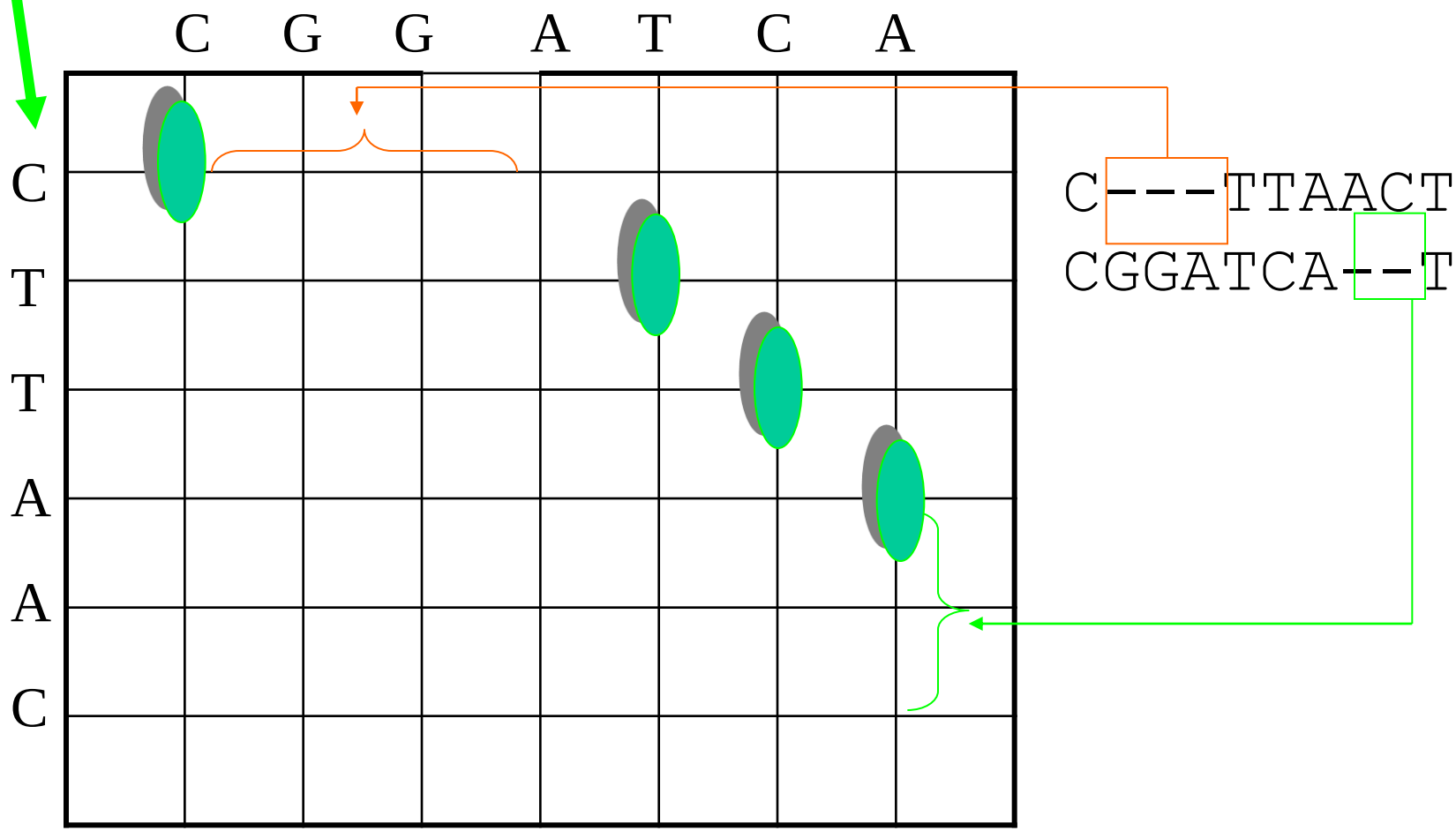
# Graphic representation of an alignment

Sequence a: CTTAACT  Sequence b: CGGATCAT



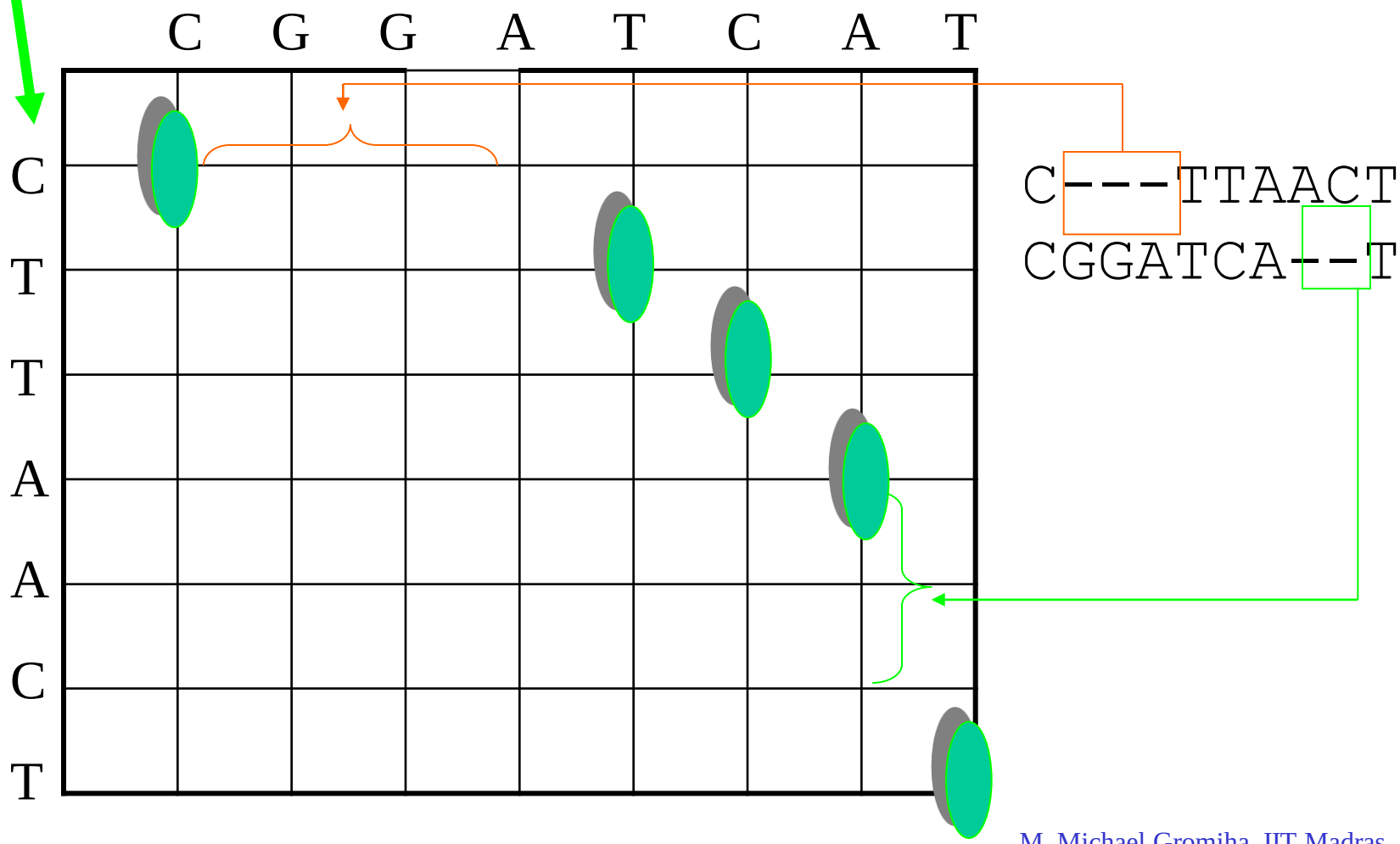
# Graphic representation of an alignment

Sequence a: CTTAACT  Sequence b: CGGATCAT



# Graphic representation of an alignment

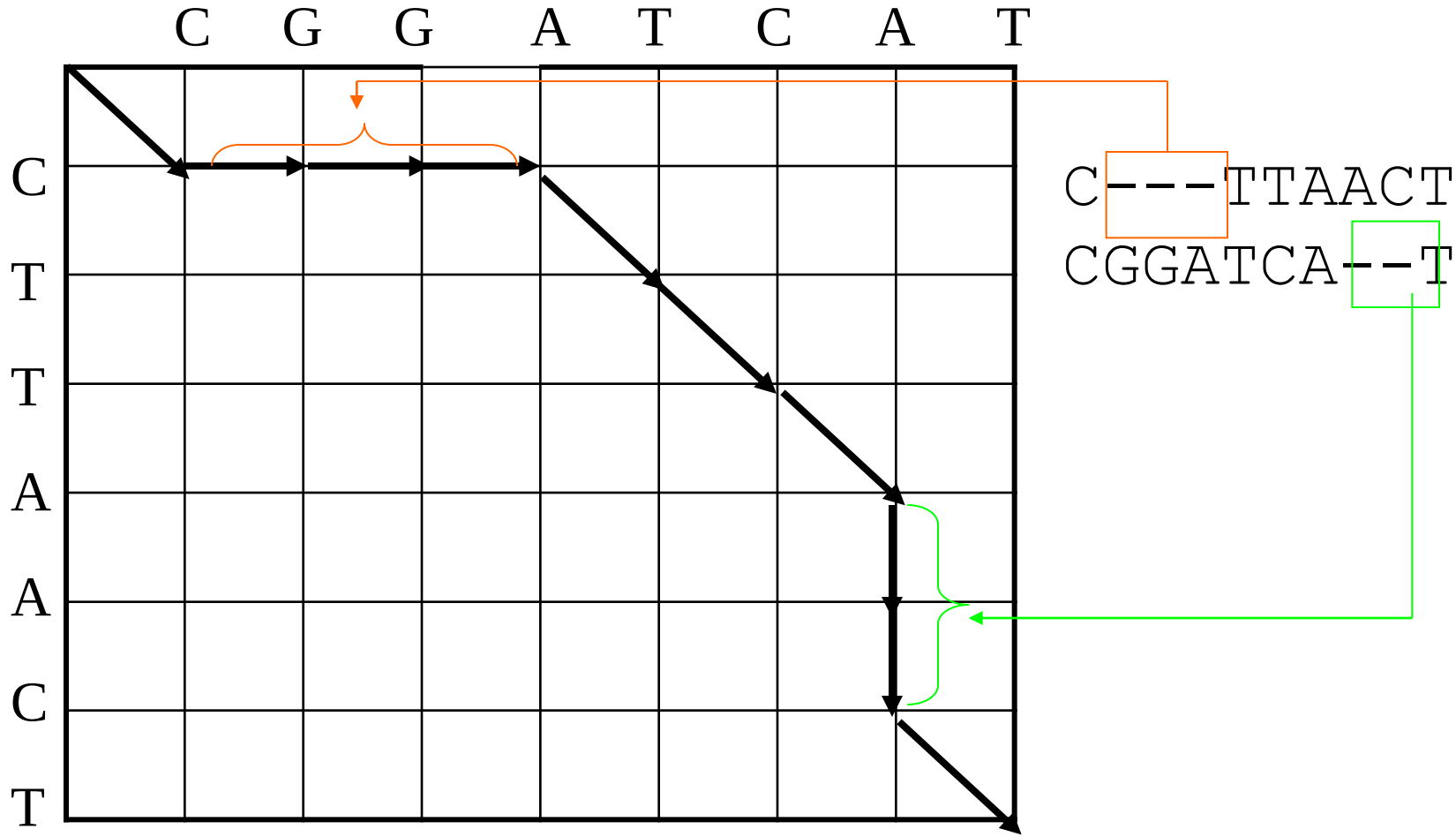
Sequence a: CTTAACT  Sequence b: CGGATCAT



# Pathway of an alignment

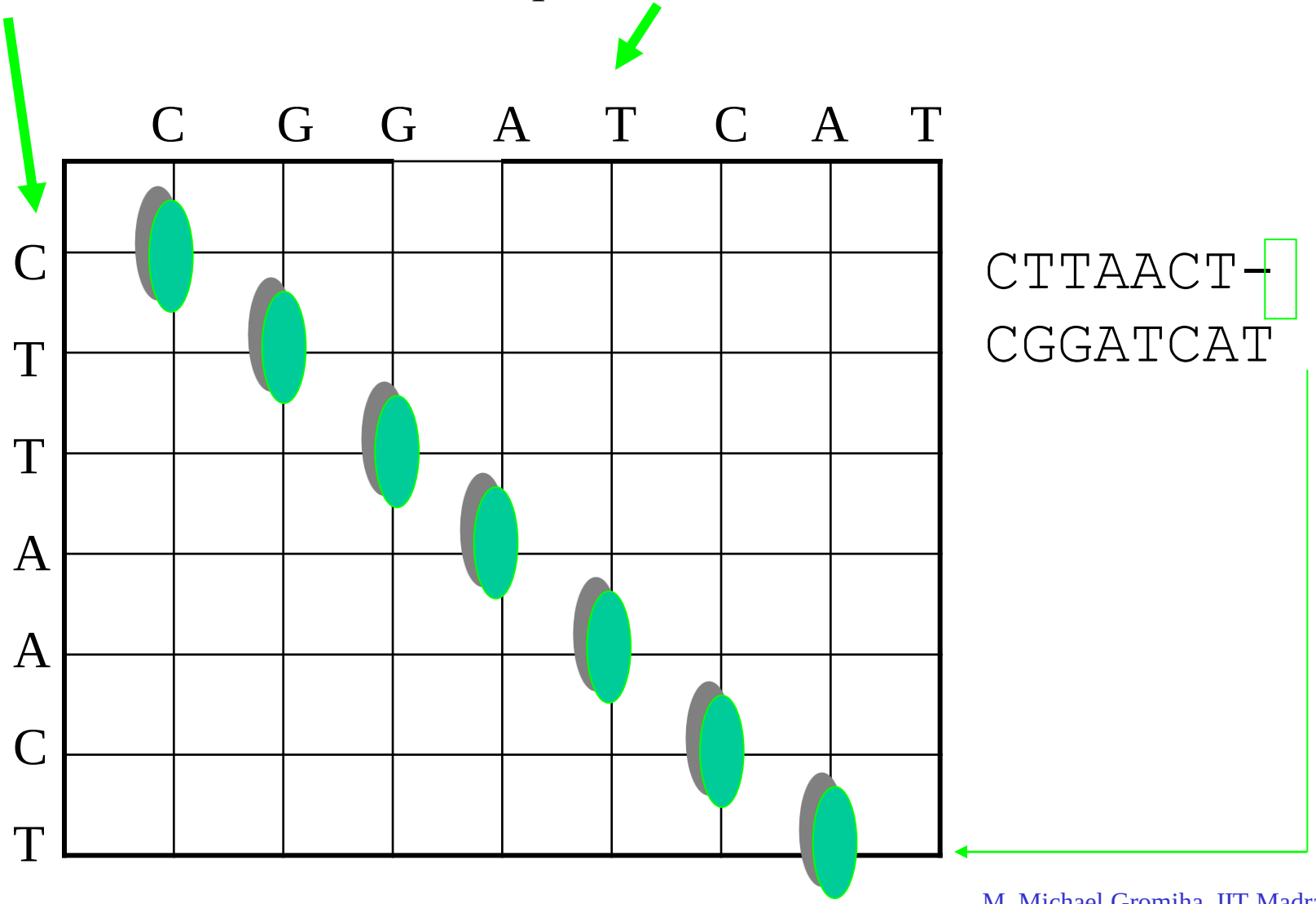
Sequence a: CTTAACT

Sequence b: CGGATCAT



# Graphic representation of an alignment

Sequence a: CTTAACT  Sequence b: CGGATCAT

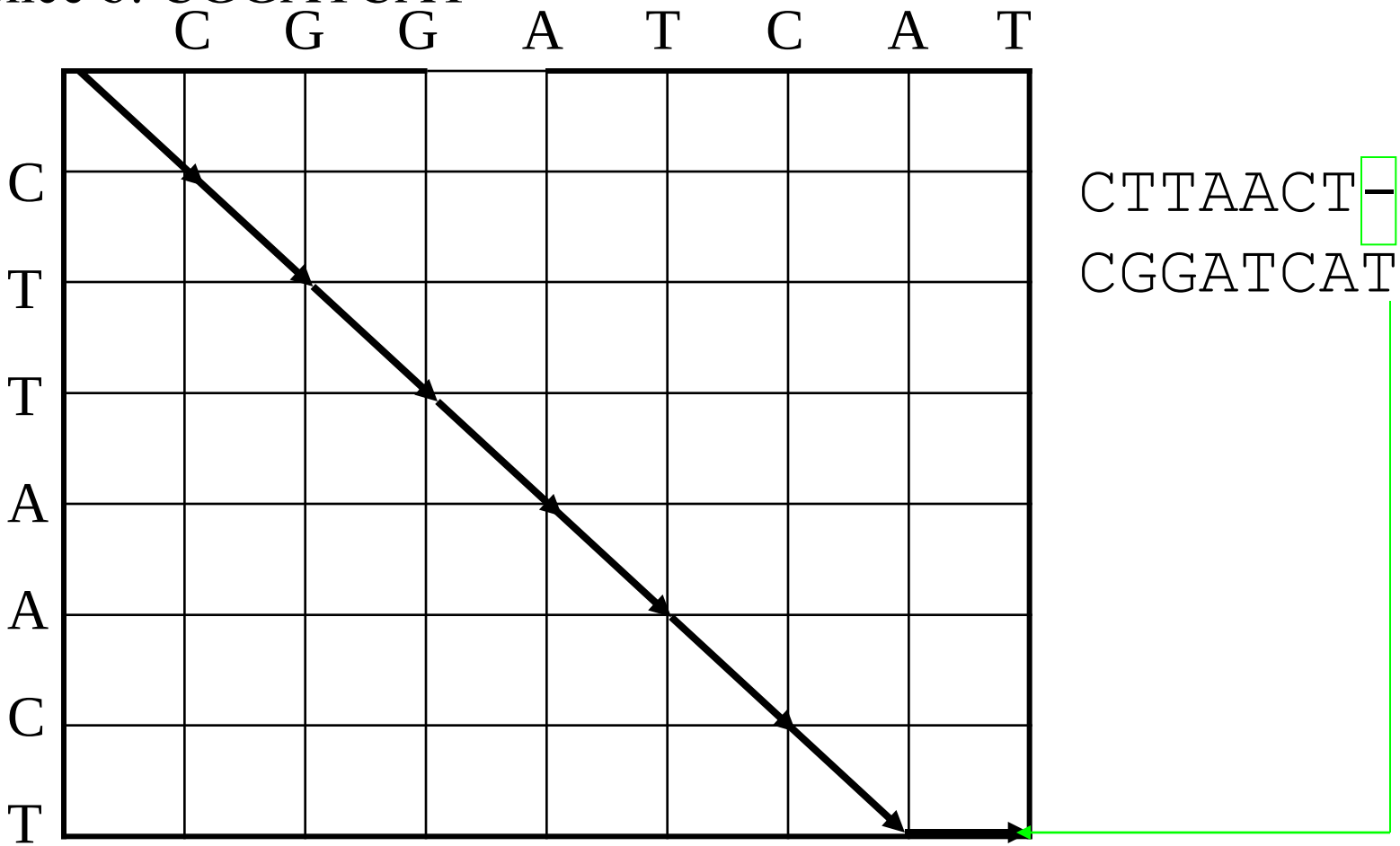



```
CTTAACT-
CGGATCAT
```

# Pathway of an alignment

Sequence a: CTTAACT

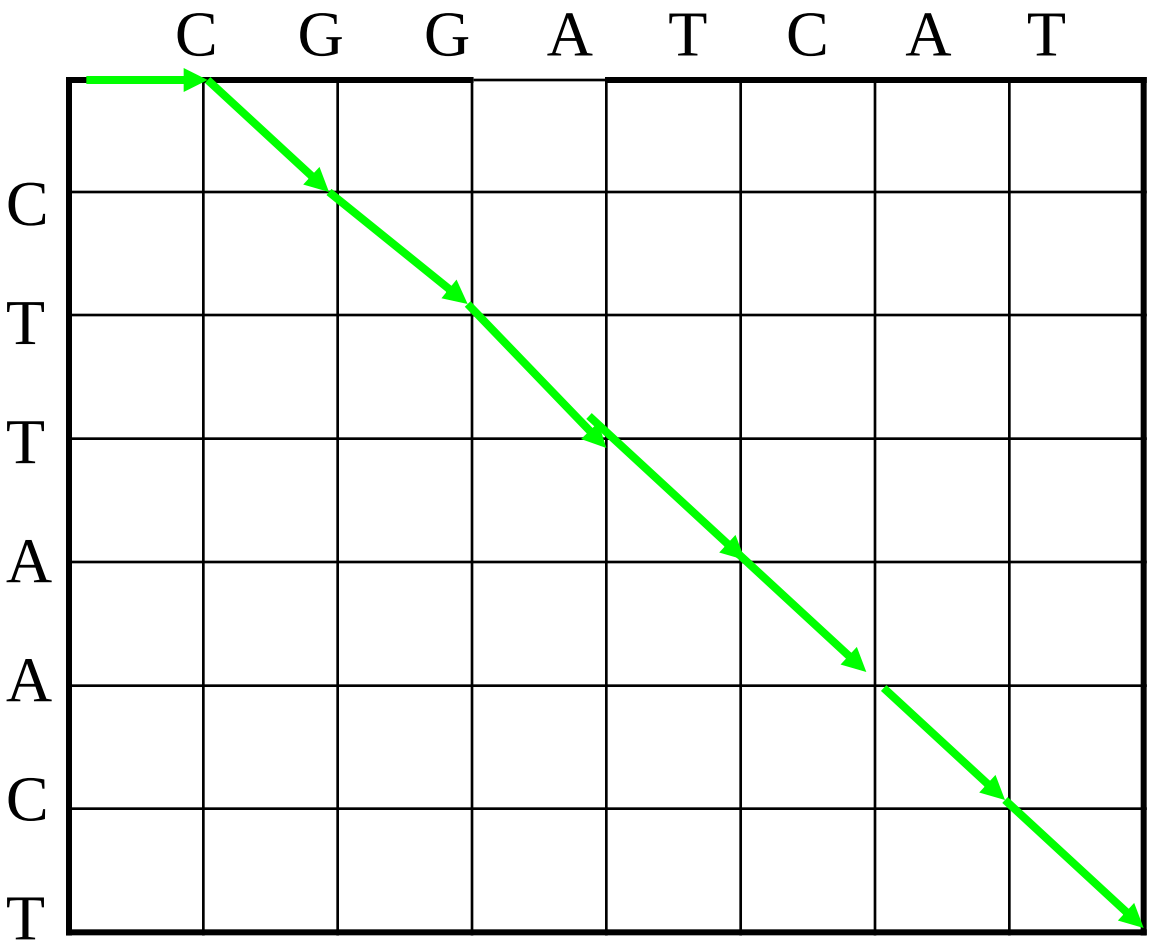
Sequence b: CGGATCAT



# Use of graph to generate alignments

Sequence a: CTTAACT

Sequence b: CGGATCAT

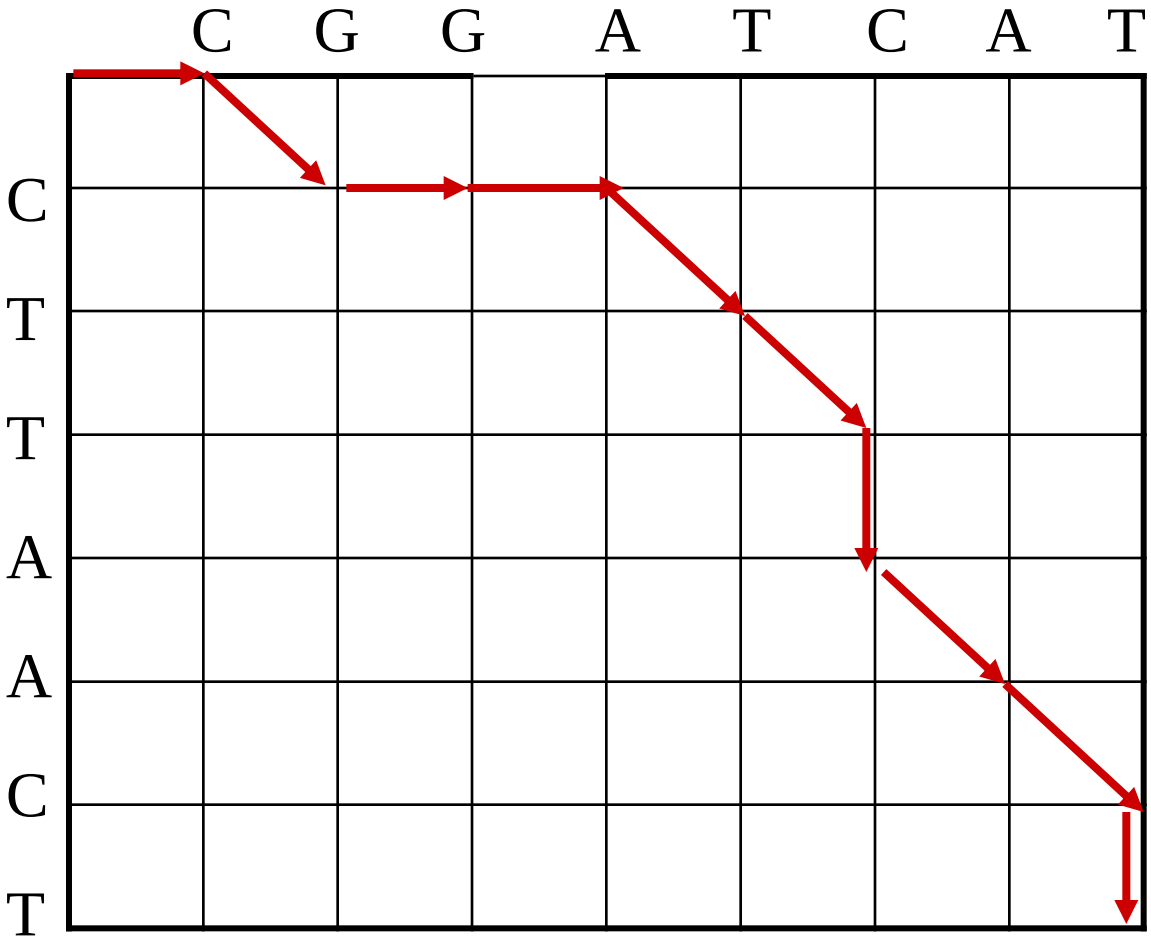



**- CTTAACT**
**CGGATCAT**

# Use of graph to generate alignments

Sequence a: CTTAACT

Sequence b: CGGATCAT

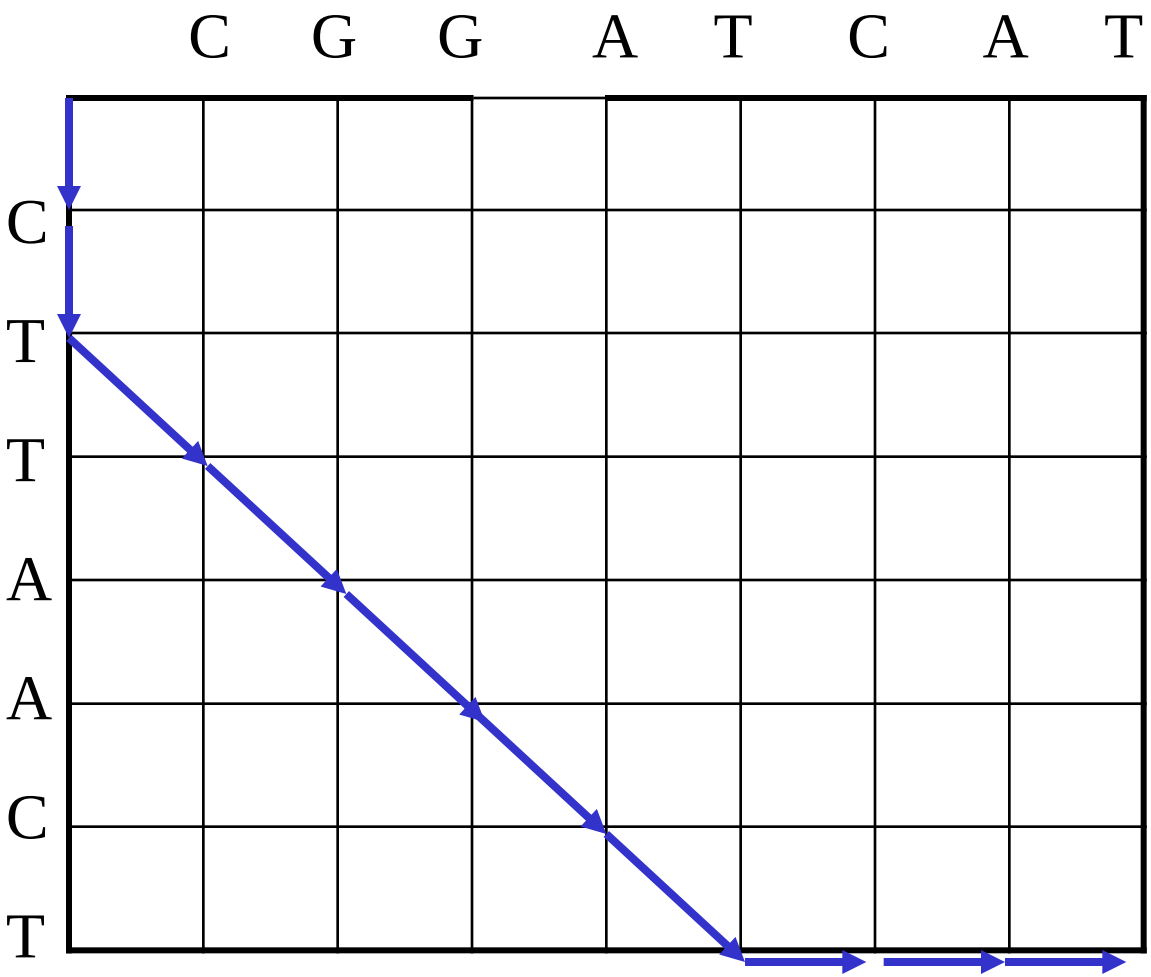



**- C - - TTAACT**
**CGGATC - AT -**

# Use of graph to generate alignments

Sequence a: CTTAACT

Sequence b: CGGATCAT



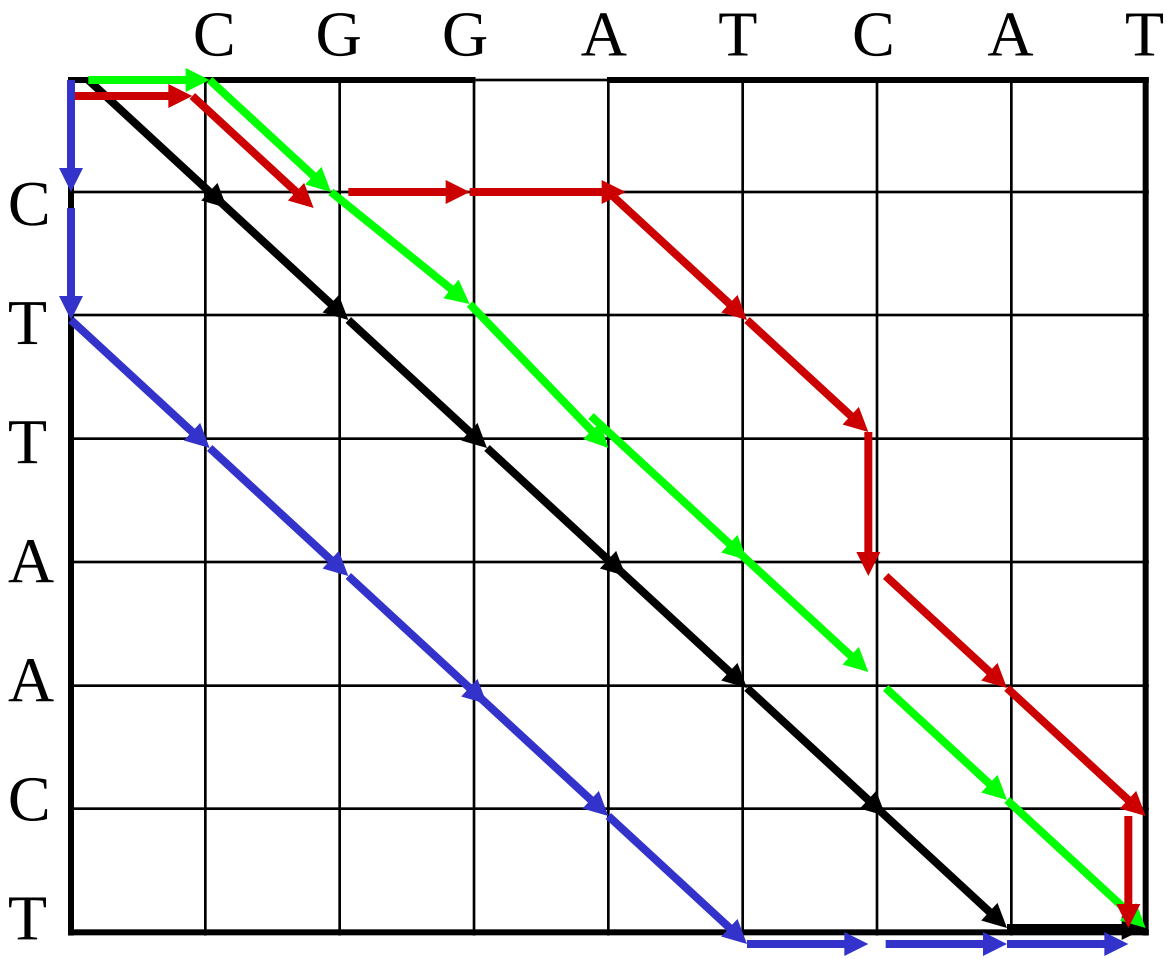

**CTTAACT - - -**

**- - CGGATCAT**

# Which pathway is better?

Sequence a: CTTAACT

Sequence b: CGGATCAT




**Multiple pathways**

**Each with a unique scoring function**

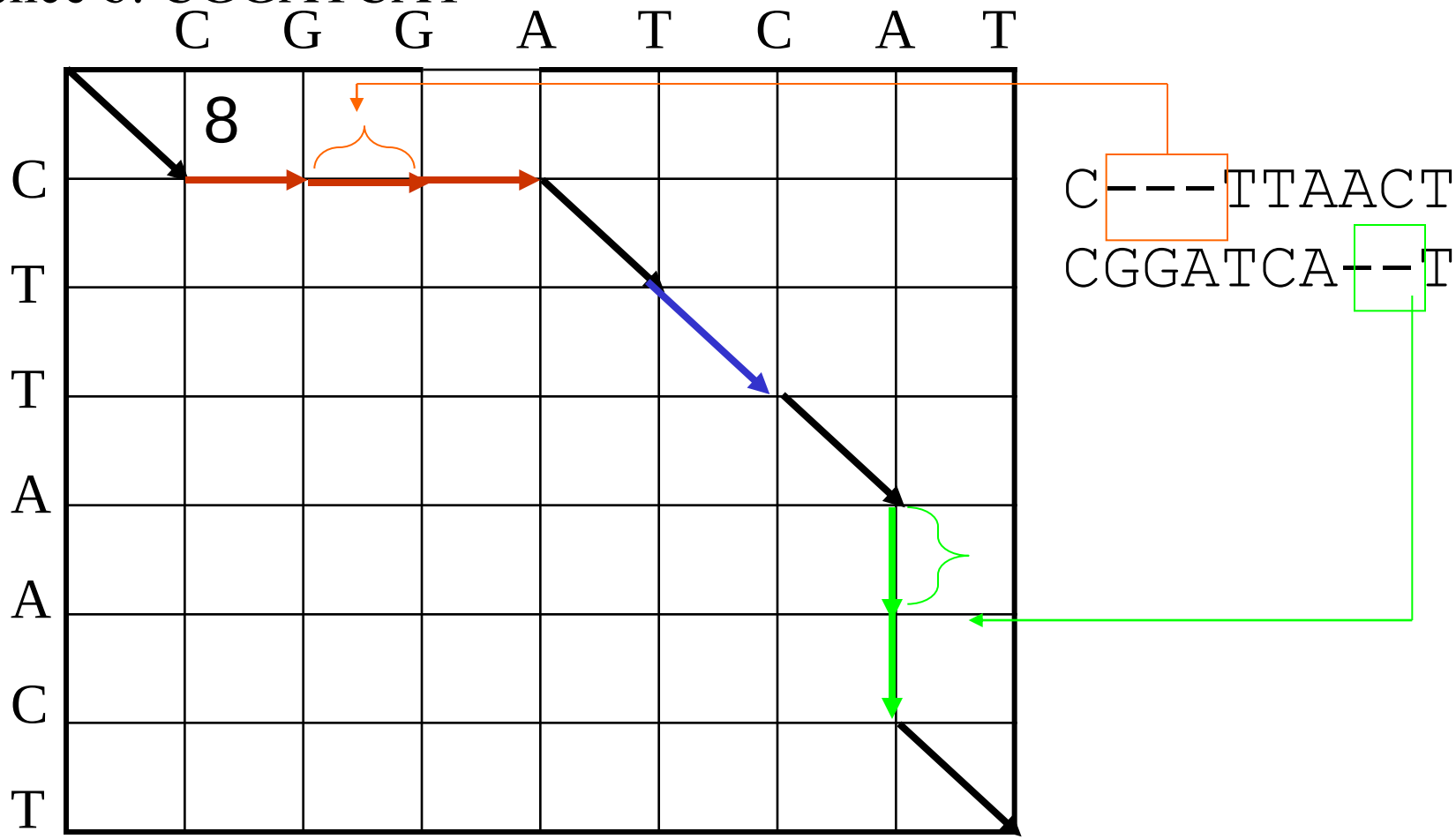
# How to rate an alignment?

- Match: +8 ($w(x, y) = 8$, if $x = y$)
- Mismatch: -5 ($w(x, y) = -5$, if $x \neq y$)
- Each gap symbol: -3 ($w(-,x)=w(x,-)=-3$)

# Alignment Score

Sequence a: CTTAACT

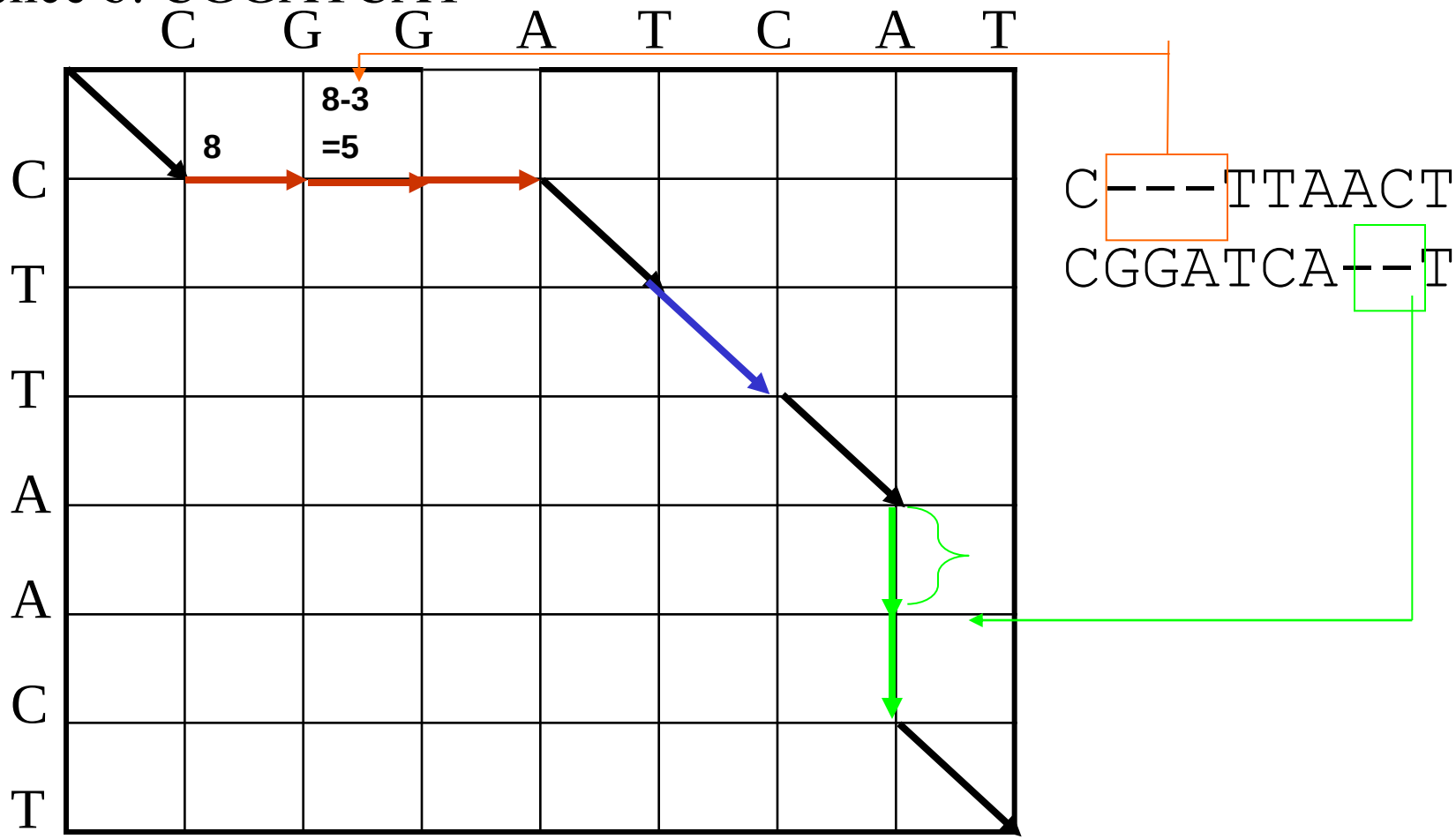
Sequence b: CGGATCAT



# Alignment Score

Sequence a: CTTAACT

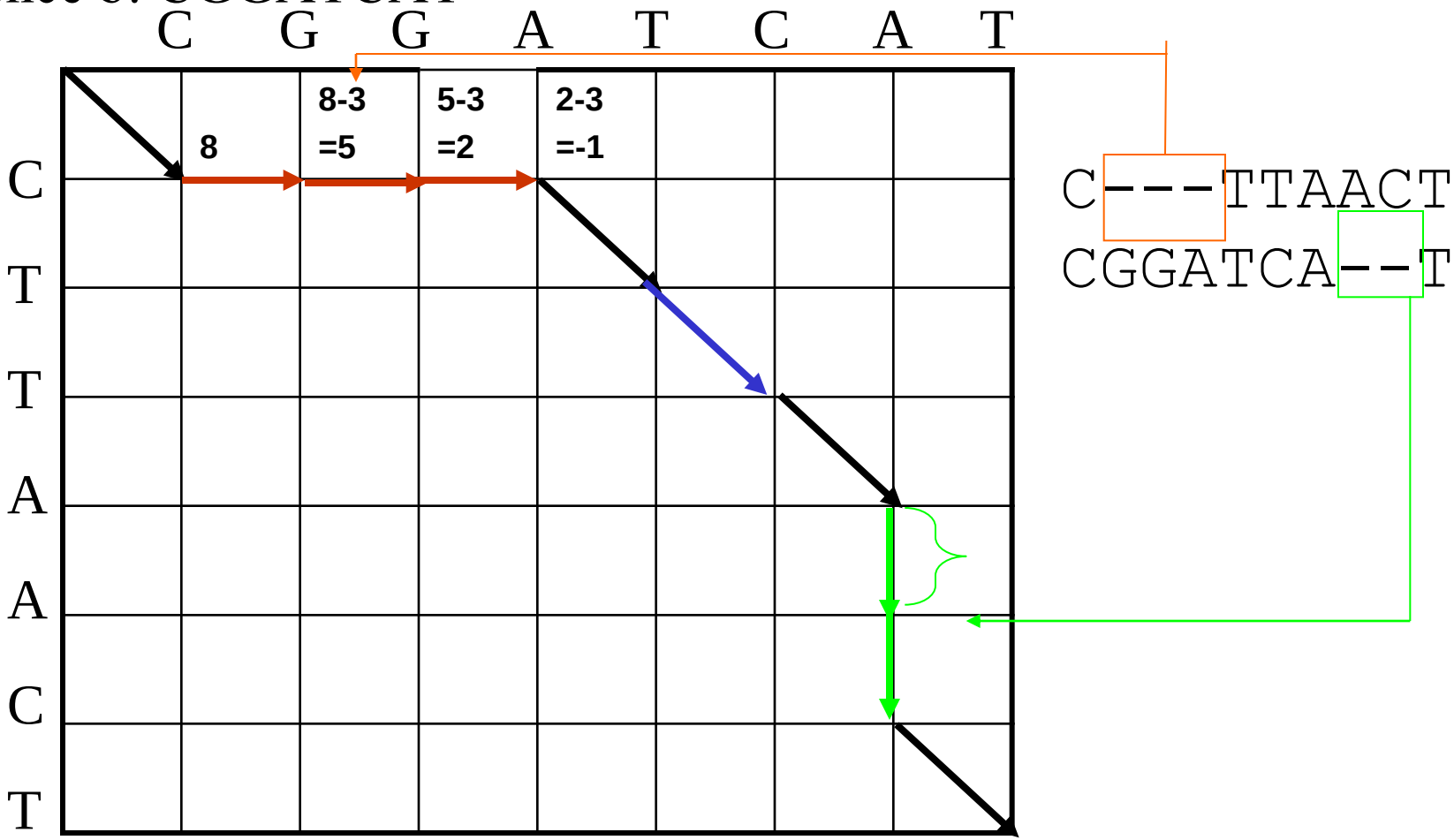
Sequence b: CGGATCAT



# Alignment Score

Sequence a: CTTAACT

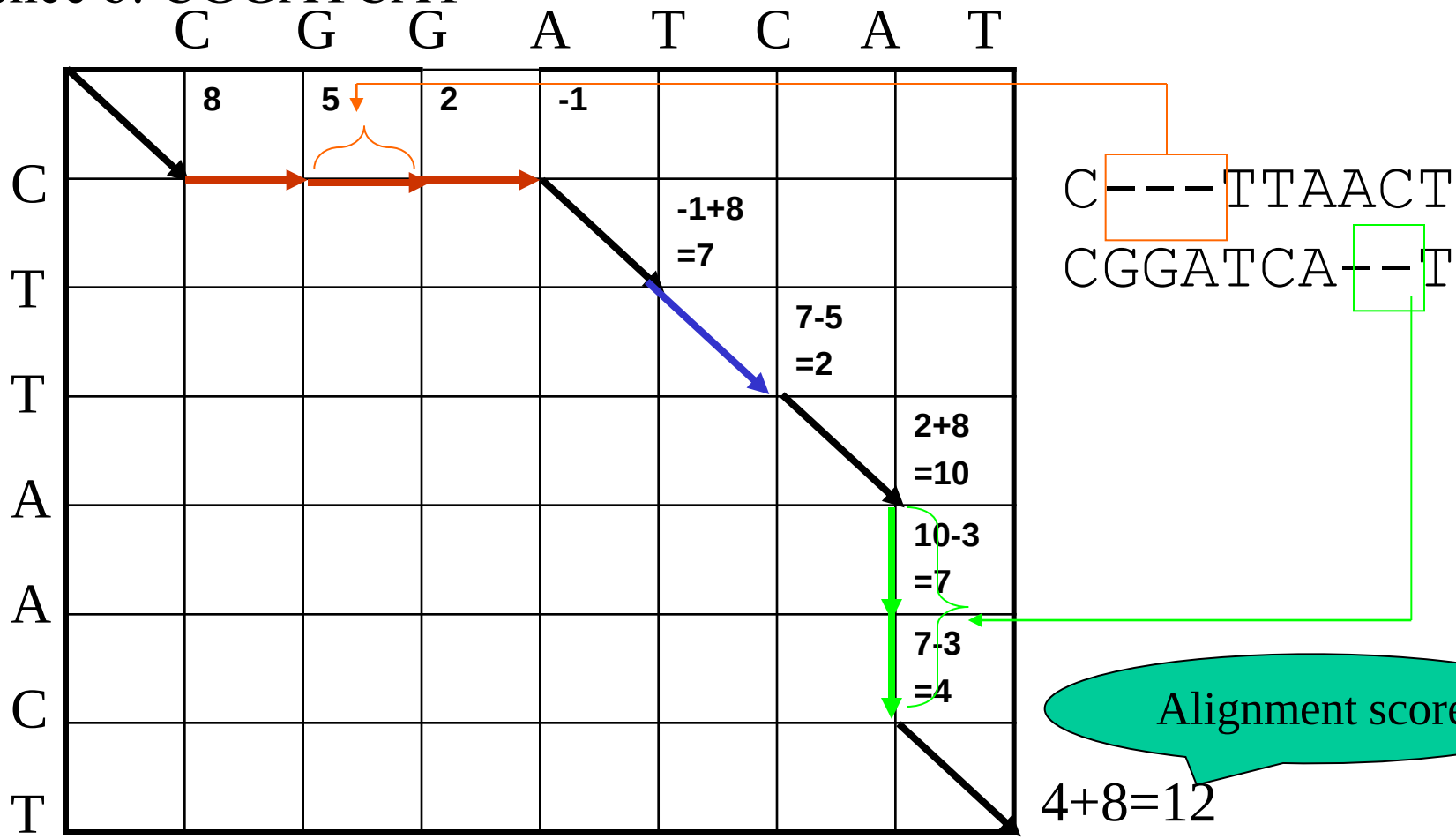
Sequence b: CGGATCAT



# Alignment Score

Sequence a: CTTAACT

Sequence b: CGGATCAT




Alignment score

4+8=12

# An optimal alignment
## -- the alignment of maximum score
# Needleman and Wunsch Algorithm

Let A=$a_1a_2\ldots a_m$ and $B$=$b_1b_2\ldots b_n$ .

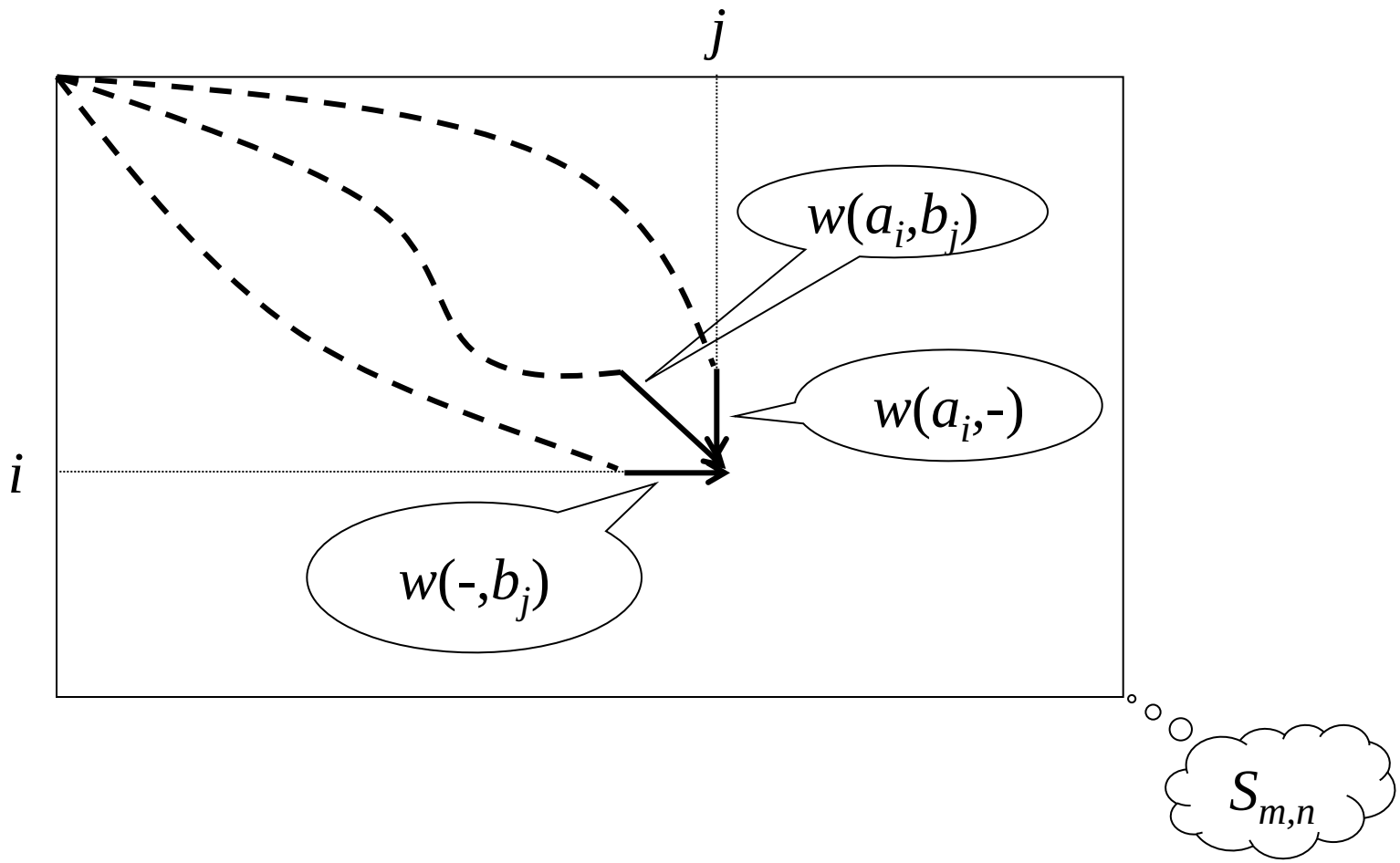
$S_{i,j}$: the score of an optimal alignment between $a_1a_2\ldots a_i$ and $b_1b_2\ldots b_j$

With proper initializations, $S_{i,j}$ can be computed as follows.

$$s_{i,j} = \max\begin{cases} s_{i-1,j} + w(a_i,-) \\ s_{i,j-1} + w(-,b_j) \\ s_{i-1,j-1} + w(a_i,b_j) \end{cases}$$

# Computing $S_{i,j}$

**Take value from left and add gap penalty along the left axis (i)**
**Take value from above and gap penalty along the top axis (j)**
**Take value from diagonal element and add match bonus/mismatch penalty (i,j)**



Gap symbol: -3

# Initializations

| | | C | G | G | A | T | C | A | T |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | -3 | -6 | -9 | -12 | -15 | -18 | -21 | -24 |
| C | -3 | | | | | | | | |
| T | -6 | | | | | | | | |
| T | -9 | | | | | | | | |
| A | -12 | | | | | | | | |
| A | -15 | | | | | | | | |
| C | -18 | | | | | | | | |
| T | -21 | | | | | | | | |

$S_{0,0}= 0$

$S_{0,1}=-3$, $S_{0,2}=-6$,

$S_{0,3}=-9$, $S_{0,4}=-12$,

$S_{0,5}=-15$, $S_{0,6}=-18$,

$S_{0,7}=-21$, $S_{0,8}=-24$

$S_{1,0}=-3$, $S_{2,0}=-6$,

$S_{3,0}=-9$, $S_{4,0}=-12$,

$S_{5,0}=-15$, $S_{6,0}=-18$,

$S_{7,0}=-21$

**Match: 8**

**Mismatch: -5**

**Gap symbol: -3**

$$S_{1,1} = ?$$

| | | C | G | G | A | T | C | A | T |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | -3 | -6 | -9 | -12 | -15 | -18 | -21 | -24 |
| C | -3 | ? | | | | | | | |
| T | -6 | | | | | | | | |
| T | -9 | | | | | | | | |
| A | -12 | | | | | | | | |
| A | -15 | | | | | | | | |
| C | -18 | | | | | | | | |
| T | -21 | | | | | | | | |

**Option 1:**

$S_{1,1} = S_{0,0} + w(a_1, b_1)$

$= 0 + 8 = 8$

**Option 2:**

$S_{1,1} = S_{0,1} + w(a_1, -)$

$= -3 - 3 = -6$

**Option 3:**

$S_{1,1} = S_{1,0} + w(-, b_1)$

$= -3 - 3 = -6$

**Optimal:**

$S_{1,1} = 8$

**Match: 8**

**Mismatch: -5**

**Gap symbol: -3**

$$S_{1,2} = ?$$

| | | C | G | G | A | T | C | A | T |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | -3 | -6 | -9 | -12 | -15 | -18 | -21 | -24 |
| C | -3 | 8 | ? | | | | | | |
| T | -6 | | | | | | | | |
| T | -9 | | | | | | | | |
| A | -12 | | | | | | | | |
| A | -15 | | | | | | | | |
| C | -18 | | | | | | | | |
| T | -21 | | | | | | | | |

**Option 1:**

$S_{1,2} = S_{0,1} + w(a_1, b_2)$

$= -3 - 5 = -8$

**Option 2:**

$S_{1,2} = S_{0,2} + w(a_1, -)$

$= -6 - 3 = -9$

**Option 3:**

$S_{1,2} = S_{1,1} + w(-, b_2)$

$= 8 - 3 = 5$

**Optimal:**

$S_{1,2} = 5$

Match: 8

Mismatch: -5

Gap symbol: -3

$$S_{2,1} = ?$$

| | | C | G | G | A | T | C | A | T |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | -3 | -6 | -9 | -12 | -15 | -18 | -21 | -24 |
| C | -3 | 8 | 5 | | | | | | |
| T | -6 | ? | | | | | | | |
| T | -9 | | | | | | | | |
| A | -12 | | | | | | | | |
| A | -15 | | | | | | | | |
| C | -18 | | | | | | | | |
| T | -21 | | | | | | | | |

**Option 1:**

$S_{2,1} = S_{1,0} + w(a_2, b_1)$

$= -3 - 5 = -8$

**Option 2:**

$S_{2,1} = S_{1,1} + w(a_2, -)$

$= 8 - 3 = 5$

**Option 3:**

$S_{2,1} = S_{2,0} + w(-, b_1)$

$= -6 - 3 = -9$

**Optimal:**

$S_{2,1} = 5$

Match: 8

Mismatch: -5

Gap symbol: -3

$$S_{2,2} = ?$$

| | | C | G | G | A | T | C | A | T |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | -3 | -6 | -9 | -12 | -15 | -18 | -21 | -24 |
| C | -3 | 8 | 5 | | | | | | |
| T | -6 | 5 | ? | | | | | | |
| T | -9 | | | | | | | | |
| A | -12 | | | | | | | | |
| A | -15 | | | | | | | | |
| C | -18 | | | | | | | | |
| T | -21 | | | | | | | | |

**Option 1:**

$S_{2,2} = S_{1,1} + w(a_2, b_2)$

$= 8 - 5 = 3$

**Option 2:**

$S_{2,2} = S_{1,2} + w(a_2, -)$

$= 5 - 3 = 2$

**Option 3:**

$S_{2,2} = S_{2,1} + w(-, b_2)$

$= 5 - 3 = 2$

**Optimal:**

$S_{2,2} = 3$

# $S_{3,5} = ?$

| | | C | G | G | A | T | C | A | T |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | -3 | -6 | -9 | -12 | -15 | -18 | -21 | -24 |
| C | -3 | 8 | 5 | 2 | -1 | -4 | -7 | -10 | -13 |
| T | -6 | 5 | 3 | 0 | -3 | 7 | 4 | 1 | -2 |
| T | -9 | 2 | 0 | -2 | -5 | ? | | | |
| A | -12 | | | | | | | | |
| A | -15 | | | | | | | | |
| C | -18 | | | | | | | | |
| T | -21 | | | | | | | | |

# $S_{3,5} = ?$

| | | C | G | G | A | T | C | A | T |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | -3 | -6 | -9 | -12 | -15 | -18 | -21 | -24 |
| C | -3 | 8 | 5 | 2 | -1 | -4 | -7 | -10 | -13 |
| T | -6 | 5 | 3 | 0 | -3 | 7 | 4 | 1 | -2 |
| T | -9 | 2 | 0 | -2 | -5 | 5 | -1 | -4 | 9 |
| A | -12 | -1 | -3 | -5 | 6 | 3 | 0 | 7 | 6 |
| A | -15 | -4 | -6 | -8 | 3 | 1 | -2 | 8 | 5 |
| C | -18 | -7 | -9 | -11 | 0 | -2 | 9 | 6 | 3 |
| T | -21 | -10 | -12 | -14 | -3 | 8 | 6 | 4 | 14 |

optimal score

```
C T T A A C – T
C G G A T C A T
```

8 – 5 –5 +8 -5 +8 -3 +8 = 14

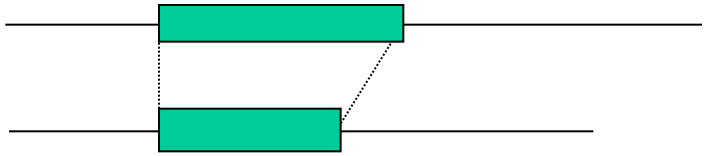
| | | C | G | G | A | T | C | A | T |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | -3 | -6 | -9 | -12 | -15 | -18 | -21 | -24 |
| C | -3 | 8 | 5 | 2 | -1 | -4 | -7 | -10 | -13 |
| T | -6 | 5 | 3 | 0 | -3 | 7 | 4 | 1 | -2 |
| T | -9 | 2 | 0 | -2 | -5 | 5 | -1 | -4 | 9 |
| A | -12 | -1 | -3 | -5 | 6 | 3 | 0 | 7 | 6 |
| A | -15 | -4 | -6 | -8 | 3 | 1 | -2 | 8 | 5 |
| C | -18 | -7 | -9 | -11 | 0 | -2 | 9 | 6 | 3 |
| T | -21 | -10 | -12 | -14 | -3 | 8 | 6 | 4 | 14 |

# Global Alignment vs. Local Alignment

- global alignment:

All sections are counted

- local alignment:



Only local sections (normally separated by gaps) are counted

# Local vs. Global Sequence Alignment:

***Example:***

DNA sequence a: ATTCTTGC

DNA sequence b: ATCCTATTCTAGC

```
Local Alignment:          ATTCTTGC    Gaps ignored in local alignments
                     ATCCTATTCTAGC
                        /|\
                       gap

Global Alignment:  AT     TCTT      GC
                     ATCCTATTCTAGC
                 /|\          /|\
                      gap         gap     Gaps counted in global alignments
```

# Semi-global alignment

AAACACGTGTCT

ACGT

```
AAACACGTGTCT
----ACGT----
```

**Shorter sequence appears entirely within the longer sequence.**

**Terminal gaps are usually the result of incomplete data acquisition and do not have biological significance.**

**It is appropriate to treat them separately than internal gaps.**

**This approach is referred as semi-global alignment.**

# Local alignment

**Semi-global alignment do not afford flexibility needed in a sequence search.**

**Eg. Find any sub-sequences that are similar to any part of the yeast genome.**

**All the non-matching residues will be penalized in semi-global alignment.**

AACCTATAGCT

GCGATATA

```
AAC-CTATAGCT
-GCGATATA---
```

**One sense: fairly a bad alignment;**

**This alignment reveals the matching region TATA.**

**The approach to find the best matching sub sequences within the two search sequences is called local alignment.**

**F. Smith and M. Waterman**

**Fourth option for alignment with the minimum value "Zero"**

# An optimal local alignment

**$S_{i,j}$: the score of an optimal local alignment ending at $a_i$ and $b_j$**

**With proper initializations, $S_{i,j}$ can be computed as follows.**

**$S_{ij}$ is the maximum among the four values**

**(i) $S_{i-1,j-1} + w(a_i,b_j)$**

**(ii) $S_{i-1,j} + w(a_i,-)$**

**(iii) $S_{i,j-1} + w(-,b_j)$**

**(iv) 0**

Match: 8

Mismatch: -5

Gap symbol: -3

# Initializations

| | | C | G | G | A | T | C | A | T |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 0 | | | | | | | | |
| T | 0 | | | | | | | | |
| T | 0 | | | | | | | | |
| A | 0 | | | | | | | | |
| A | 0 | | | | | | | | |
| C | 0 | | | | | | | | |
| T | 0 | | | | | | | | |

Match: 8

Mismatch: -5

Gap symbol: -3

$$S_{1,1} = ?$$

| | | C | G | G | A | T | C | A | T |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 0 | ? | | | | | | | |
| T | 0 | | | | | | | | |
| T | 0 | | | | | | | | |
| A | 0 | | | | | | | | |
| A | 0 | | | | | | | | |
| C | 0 | | | | | | | | |
| T | 0 | | | | | | | | |

**Option 1:**

$S_{1,1} = S_{0,0} + w(a_1, b_1)$

$= 0 + 8 = 8$

**Option 2:**

$S_{1,1} = S_{0,1} + w(a_1, -)$

$= 0 - 3 = -3$

**Option 3:**

$S_{1,1} = S_{1,0} + w( - , b_1)$

$= 0 - 3 = -3$

**Option 4:**

$S_{1,1} = 0$

**Optimal:**

$S_{1,1} = 8$

# local alignment

Match: 8

Mismatch: -5

Gap symbol: -3

| | | C | G | G | A | T | C | A | T |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 0 | 8 | 5 | 2 | 0 | 0 | 8 | 5 | 2 |
| T | 0 | 5 | 3 | 0 | 0 | 8 | 5 | 3 | 13 |
| T | 0 | 2 | 0 | 0 | 0 | 8 | 5 | 2 | 11 |
| A | 0 | 0 | 0 | 0 | 8 | 5 | 3 | ? | |
| A | 0 | | | | | | | | |
| C | 0 | | | | | | | | |
| T | 0 | | | | | | | | |

```
A - C - T
A T C A T
8-3+8-3+8 = 18
```

# local alignment

| | | C | G | G | A | T | C | A | T |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 0 | 8 | 5 | 2 | 0 | 0 | 8 | 5 | 2 |
| T | 0 | 5 | 3 | 0 | 0 | 8 | 5 | 3 | 13 |
| T | 0 | 2 | 0 | 0 | 0 | 8 | 5 | 2 | 11 |
| A | 0 | 0 | 0 | 0 | 8 | 5 | 3 | 13 | 10 |
| A | 0 | 0 | 0 | 0 | 8 | 5 | 2 | 11 | 8 |
| C | 0 | 8 | 5 | 2 | 5 | 3 | 13 | 10 | 7 |
| T | 0 | 5 | 3 | 0 | 2 | 13 | 10 | 8 | 18 |

The best score

# Local sequence alignment: Example 2

AACCTATAGCT

GCGATATA

# Initializations

Match: 1;

Mismatch: -1;

Gap symbol: -1

| | | A | A | C | C | T | A | T | A | G | C | T |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G | 0 | | | | | | | | | | | |
| C | 0 | | | | | | | | | | | |
| G | 0 | | | | | | | | | | | |
| A | 0 | | | | | | | | | | | |
| T | 0 | | | | | | | | | | | |
| A | 0 | | | | | | | | | | | |
| T | 0 | | | | | | | | | | | |
| A | 0 | | | | | | | | | | | |

# Alignment

Match: 1;

Mismatch: -1;

Gap symbol: -1

| | | A | A | C | C | T | A | T | A | G | C | T |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| C | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 1 |
| G | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| A | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| T | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 1 | 0 | 0 | 1 |
| A | 0 | 1 | 1 | 0 | 0 | 0 | 2 | 0 | 3 | 2 | 1 | 0 |
| T | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 3 | 2 | 2 | 1 | 2 |
| A | 0 | 1 | 1 | 0 | 0 | 0 | 2 | 2 | 4 | 3 | 2 | 1 |

**TATA**
**TATA**

# Questions

1. Using the Needleman and Wunsch dynamic programming method, construct the partial alignment score table for the following two sequences, using the scoring parameters: match score: +1; mismatch score: 0 and gap penalty: -1

   ACAGTCGAACG and ACCGTCCG

2. Using the Smith-Waterman method, construct the partial alignment scoring table for a local alignment of the following two sequences:

   ACGTATCGCGTATA and GATGCTCTCGGAAA

   scoring parameters: match score: +1; mismatch score: 0 and gap penalty: -1

# Questions

1. Using the Needleman and Wunsch dynamic programming method, construct the partial alignment score table for the following two sequences, using the scoring parameters: match score: +1; mismatch score: 0 and gap penalty: -1

   ACAGTCGAACG and ACCGTCCG

2. Using the Smith-Waterman method, construct the partial alignment scoring table for a local alignment of the following two sequences:

   ACGTATCGCGTATA and GATGCTCTCGGAAA

   scoring parameters: match score: +1; mismatch score: 0 and gap penalty: -1

# Database searches

**The common use for sequence alignments is to search through a database of many sequences to retrieve similar sequences to the query sequence.**

**E.g. we have a region of human genome with unidentified function**

**Search with millions of other sequences in Genbank of NCBI (or EBI or DDBJ)**

**Get the regions that align well with the query sequence**

**Compare with their functional role**

**Information on regulation/expression and its relationship with other genes.**

**Major points:**

**1. Size of the query sequence**

**2. Number of sequences in the database**

**Proper methods are necessary to identify correct sequence matches.**

# BLAST

**BLAST: Basic Local Alignment Search Tool**

**BLAST finds sub-sequences from a sequence database for any query sequence.**

| Program name | Query sequence | Database type |
|---|---|---|
| Blastp | Protein | Protein |
| Blastn | Nucleic acid | Nucleic acid |
| Tblastn | Protein | Nucleic acid (translated) |
| Tblastx | Nucleic acid (translated) | Nucleic acid (translated) |

**Blastp: searches for protein sequence matches using PAM or BLOSSUM matrices to score the ungapped alignments.**

# BLAST: Example

**Blastp first breaks down the query sequence into words or subsequences of fixed length**

**All possible pairs are calculated using sliding windows**

**E.g. AILVPTVI -> AILV, ILVP, LVPT, VPTV and PTVI**

**Search for word matches (also called High Scoring Pairs or HSPs):**

**MVQGTIPKLHAILVGTVIAML …**

**AILVPTVI**

**Extend the match until the local alignment score falls below a fixed threshold**

**It also allows gaps in the extended length.**

# FASTA

**FASTA is another program for sequence similarity search and sequence alignment.**

**FASTA breaks the words into 4-6 nucleotides or 1-2 amino acids**

**Eg. Query sequence: FAMLGFIKYLPGCM**

| Word | A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Position | 2 | 13 | | | 1 | 5 | | 7 | 8 | 4 | 3 | | 11 | | | | | | | 9 |
| | | | | | 6 | 12 | | | | 10 | 14 | | | | | | | | | |

**Target sequence: TGFIKYLPGACT**

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| T | G | F | I | K | Y | L | P | G | A | C | T |
| | 3 | -2 | 3 | 3 | 3 | -3 | 3 | -4 | -8 | 2 | |
| | 10 | 3 | | | | 3 | | 3 | | | |

**Large number of sequences have the number 3;**

**Offsetting with 3 gives a reasonable alignment**

```
FAMLGFIKYLPGCM
  ||||||||
 TGFIKYLPGACT
```

# Alignment score and statistical significance

**The primary indicator of how similar the search results are to a query sequence is the alignment score (S).**

**Score is given with P or E value.**

**E-value is the expected number of sequences of score S that would be found by random choice**

**P-value is the probability that one or more sequences of score S would have been found randomly.**

**Low values of E and P indicate that the search result was unlikely to have been obtained by random chance, and thus is likely to bear an evolutionary relationship to the query sequence.**

**E values of less than $10^{-3}$ are often considered indicative of statistically significant results and search algorithms produce matches with E values on the order of $10^{-50}$.**

# BLAST: Features

**(i) identifying protein sequences similar to the query,**

**(ii) finding members of a protein family or build a custom position-specific scoring matrix,**

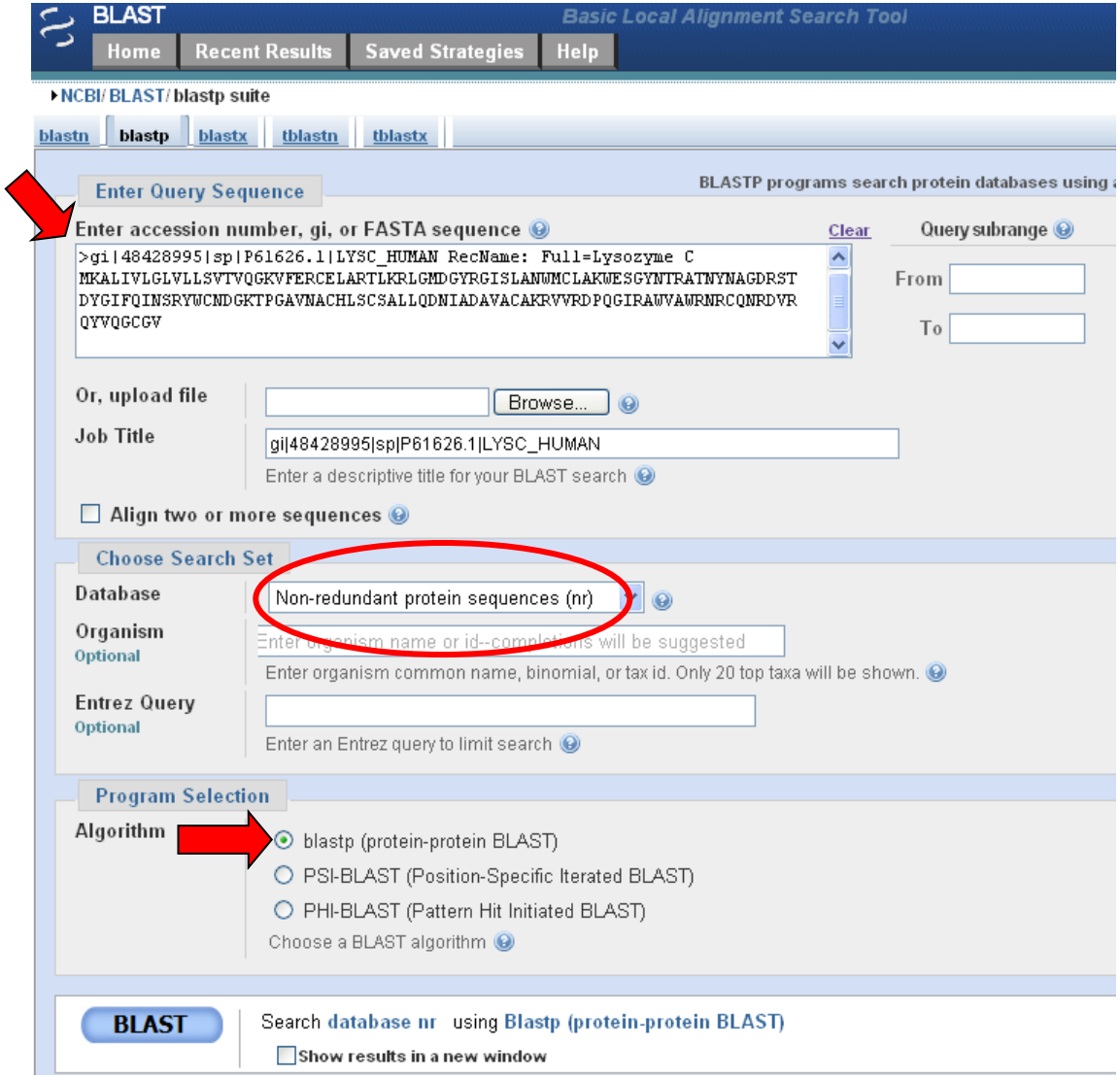
**(iii) finding proteins similar to the query around a given pattern,**

**(iv) finding conserved domains in the query and**

**(v) searching for peptide motifs.**

**BLAST is available at http://www.ncbi.nlm.nih.gov/BLAST/.**

# BLAST: Search



# BLAST: Options

**Accepts gi number or FASTA format**

**gi : bar separated NCBI sequence identifier (e.g., gi|48428995).**

**Accession number : number allotted in Uniprot for each sequence (e.g. P61626)**

**FASTA format**

**begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">") symbol at the beginning.**

**Example sequence in FASTA format:**

**> LYSC_HUMAN RecName: Full=Lysozyme**
**CMKALIVLGLVLLSVTVQGKVFERCELARTLKRLGMDGYRGISLANWMCLAKWESG**
**YNTRATNYNAGDRSTDYGIFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVAC**
**AKRVVRDPQGIRAWVAWRNRCQNRDVRQYVQGCGV**

# File formats

**1. FASTA format**

```
> LYSC_HUMAN RecName: Full=Lysozyme
    CMKALIVLGLVLLSVTVQGKVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTRATNYNAGDRSTDYG
    IFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVRDPQGIRAWVAWRNRCQNRDVRQYVQGCG
    V
```

**Files in FASTA format have the extension ".fasta"**

**2. NBRF/PIR (National Biomedical Research Foundation/Protein Information Resource)**

**First line begins with >P1 for protein sequence or >N1 for nucleic acid sequence.**

```
>P1; LYSC_HUMAN
    CMKALIVLGLVLLSVTVQGKVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTRATNYNAGDRSTDYG
    IFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVRDPQGIRAWVAWRNRCQNRDVRQYVQGCG
    V

*
```

**Files in NBRF/PIR format have the extension ".seq" or ".pir"**

**3. GDE format, (essentially the same as FASTA, the difference is starts with %). The file format is ".gde**

**All three file formats ignore spaces and carriage returns.**

# Searchable databases

## Peptide Sequence Databases

**nr**

All non-redundant GenBank CDS translations + RefSeq Proteins + PDB + SwissProt + PIR + PRF

**refseq**

RefSeq protein sequences from NCBI's Reference Sequence Project.

**Uniprot**

Last major release of the Uniprot protein sequence database.

**pat**

Proteins from the Patent division of GenPept.

**pdb**

Sequences derived from the 3-dimensional structure from Protein Data Bank

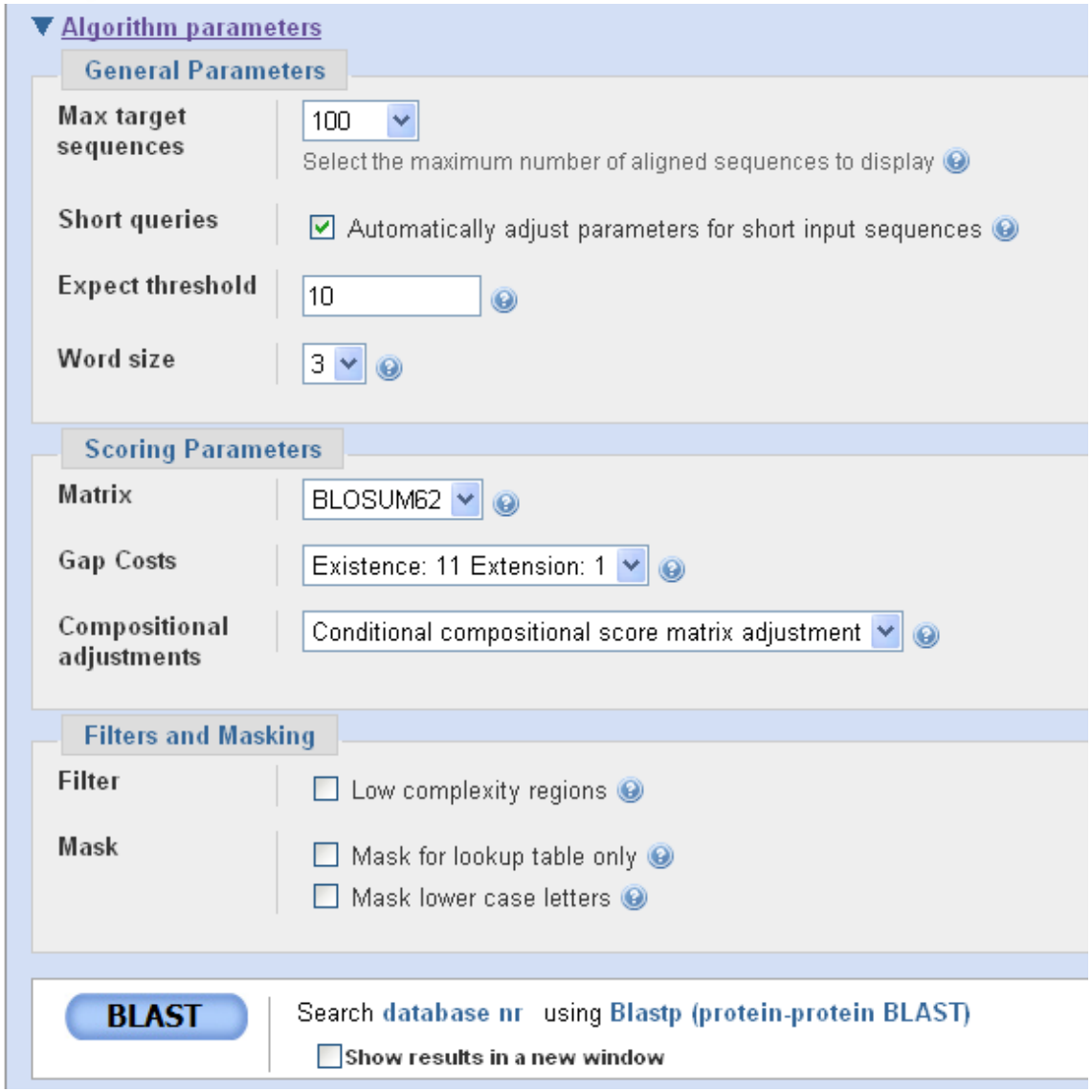
**month**

All new or revised GenBank CDS translation+PDB+SwissProt+PIR+PRF released in the last 30 days.

**env_nr**

Metagenomic Protein sequences.

# Algorithm parameters



# Algorithm parameters

**Displaying maximum number of aligned sequences**

**Expect threshold (e-value): expected number of chance matches in a random model and it is set to 10 as default value.**

**Word size: length of the seed that initiates an alignment.**

**scoring parameters can be selected for matrix, gap cost and compositional adjustments.**

**Substitution matrix: It is a key element in evaluating the quality of a pairwise sequence alignment, which assigns a score for aligning any possible pair of residues.**

**Generally BLOSUM62 is used as the substitution matrix, which is a 20x20 matrix obtained for all possible substitutions of 20 amino acid residues. It is based on a likelihood method by estimating the occurrence of each possible pairwise substitution using the biochemical character of amino acid residues (aliphatic, aromatic, positive charged, negative charged; polar, sulfur containing).**

**The gap cost is a cost to create and extend a gap in an alignment.**

**Further, options are available to filter the low complexity regions and mask query and lower case letters in the sequence.**

# BLAST: Output

**gi|48428995|sp|P61626.1|LYSC_HUMAN RecName:...**

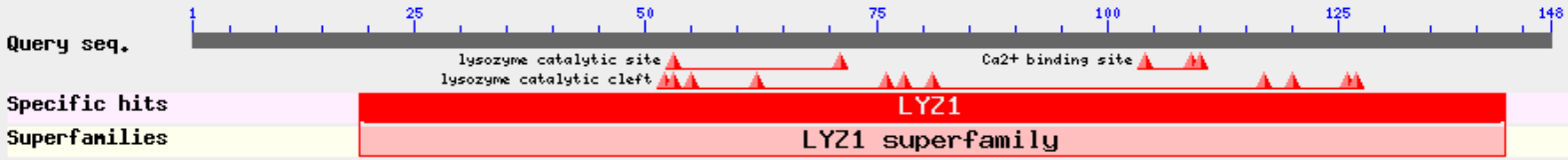
| | | | |
|---|---|---|---|
| **Query ID** | lcl\|55010 | **Database Name** | nr |
| **Description** | gi\|48428995\|sp\|P61626.1\|LYSC_HUMAN RecName: Full=Lysozyme C | **Description** | All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects |
| **Molecule type** | amino acid | | |
| **Query Length** | 148 | **Program** | BLASTP 2.2.25+ ▸Citation |

Other reports: ▸Search Summary [Taxonomy reports] [Distance tree of results] [Related Structures] [Multiple alignment]

▼ Graphic Summary

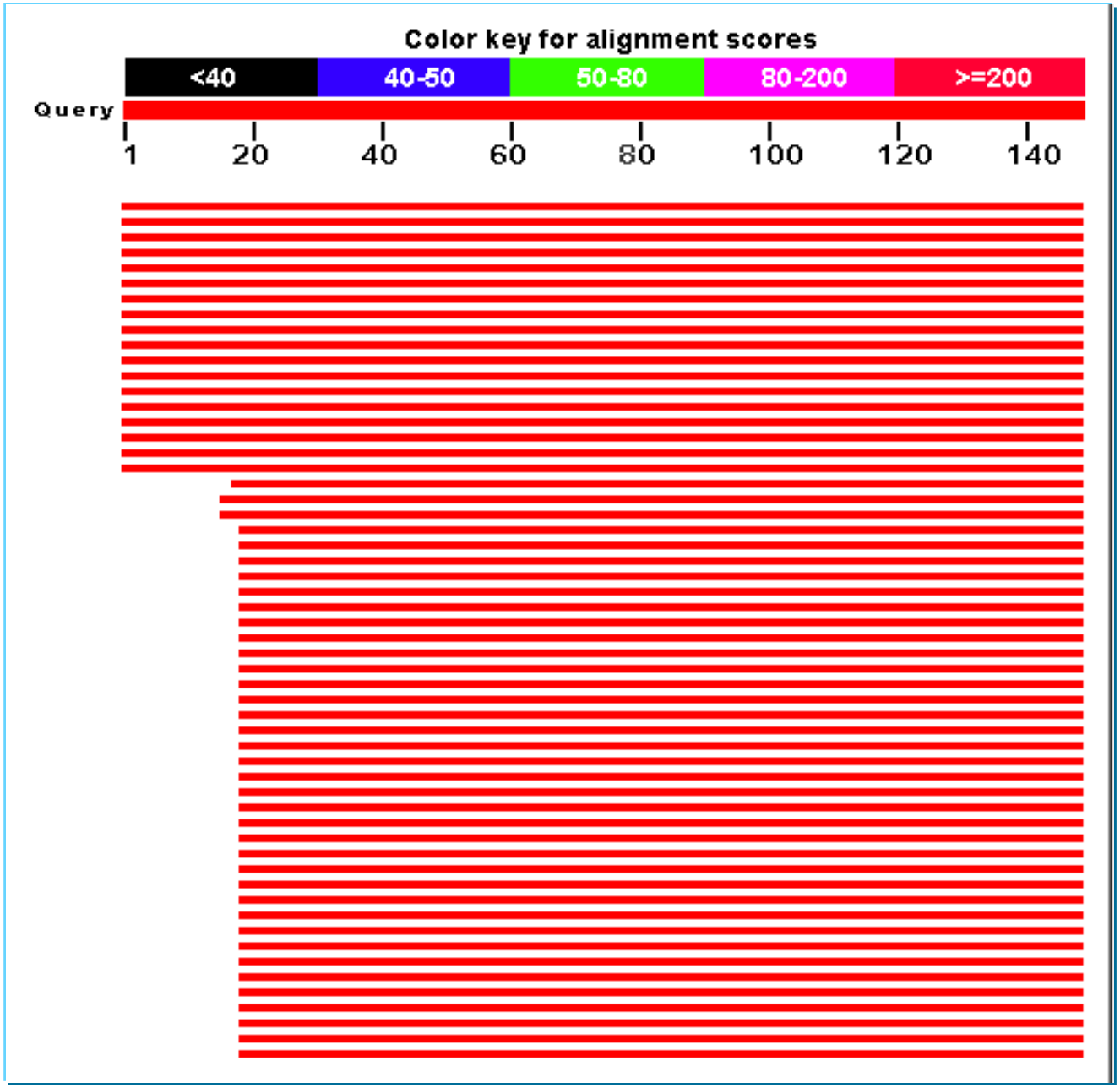
▼ Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.




**Lysozyme catalytic site**

**$Ca^{2+}$ binding site**


Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
20
40
60
80
100
120
140

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay

**Sequences producing significant alignments:**

| Accession | Description | Max score | Total score | Query coverage | E value | Links |
|---|---|---|---|---|---|---|
| NP_000230.1 | lysozyme C precursor [Homo sapiens] >ref\|NP_001009073.1\| ly | 303 | 303 | 100% | 4e-81 | U G M |
| AAA36188.1 | lysozyme precursor (EC 3.2.1.17) [Homo sapiens] | 303 | 303 | 100% | 6e-81 | G M |
| BAG73364.1 | lysozyme [synthetic construct] | 301 | 301 | 100% | 2e-80 | |
| P79179.1 | RecName: Full=Lysozyme C; AltName: Full=1,4-beta-N-acetylm | 301 | 301 | 100% | 2e-80 | |
| XP_002823550.1 | PREDICTED: lysozyme C-like [Pongo abelii] >sp\|P79239.1\|LYSC | 300 | 300 | 100% | 5e-80 | G M |
| XP_003259554.1 | PREDICTED: lysozyme C-like [Nomascus leucogenys] | 298 | 298 | 100% | 1e-79 | G M |
| AAC63078.1 | lysozyme precursor [Homo sapiens] | 297 | 297 | 100% | 4e-79 | G M |
| P79180.1 | RecName: Full=Lysozyme C; AltName: Full=1,4-beta-N-acetylm | 295 | 295 | 100% | 2e-78 | |
| P61633.1 | RecName: Full=Lysozyme C; AltName: Full=1,4-beta-N-acetylm | 283 | 283 | 100% | 5e-75 | |
| NP_001095203.1 | lysozyme C precursor [Macaca mulatta] >sp\|P30201.1\|LYSC_M | 280 | 280 | 100% | 6e-74 | U G M |
| P79811.1 | RecName: Full=Lysozyme C; AltName: Full=1,4-beta-N-acetylm | 280 | 280 | 100% | 6e-74 | |
| NP_001106112.1 | lysozyme C precursor [Papio anubis] >sp\|P61629.1\|LYSC_PAPA | 279 | 279 | 100% | 8e-74 | U G |
| P79806.1 | RecName: Full=Lysozyme C; AltName: Full=1,4-beta-N-acetylm | 279 | 279 | 100% | 8e-74 | |
| P79847.1 | RecName: Full=Lysozyme C; AltName: Full=1,4-beta-N-acetylm | 278 | 278 | 100% | 1e-73 | |
| P67979.1 | RecName: Full=Lysozyme C; AltName: Full=1,4-beta-N-acetylm | 278 | 278 | 100% | 3e-73 | |
| P67977.1 | RecName: Full=Lysozyme C; AltName: Full=1,4-beta-N-acetylm | 278 | 278 | 100% | 2e-73 | |
| P79687.1 | RecName: Full=Lysozyme C; AltName: Full=1,4-beta-N-acetylm | 277 | 277 | 100% | 4e-73 | |
| P61631.1 | RecName: Full=Lysozyme C; AltName: Full=1,4-beta-N-acetylm | 276 | 276 | 100% | 5e-73 | |
| 1C46_A | Chain A, Mutant Human Lysozyme With Foreign N-Terminal Res | 272 | 272 | 88% | 1e-71 | S |
| 1C7P_A | Chain A, Crystal Structure Of Mutant Human Lysozyme With Fc | 272 | 272 | 89% | 1e-71 | S |
| CAA53144.1 | lysozyme [synthetic construct] >gb\|AAQ72808.1\| lysozyme [s | 271 | 271 | 87% | 3e-71 | |
| 1IOC_A | Chain A, Crystal Structure Of Mutant Human Lysozyme, Eaea-I | 271 | 271 | 89% | 3e-71 | S |
| 1LZS_A | Chain A, Structural Changes Of The Active Site Cleft And Diffe | 270 | 270 | 87% | 4e-71 | S |
| 1GBW_A | Chain A, Crystal Structure Of Mutant Human Lysozyme Substit | 270 | 270 | 87% | 5e-71 | S |
| 1GB6_A | Chain A, Crystal Structure Of Mutant Human Lysozyme Substit | 270 | 270 | 87% | 5e-71 | S |

# Alignment score

**Maxscore: Bit score of high scoring pairs (HSPs), similar to Expect Value**

**Total score: sum of the scores of all HSPs from the same database sequence.**

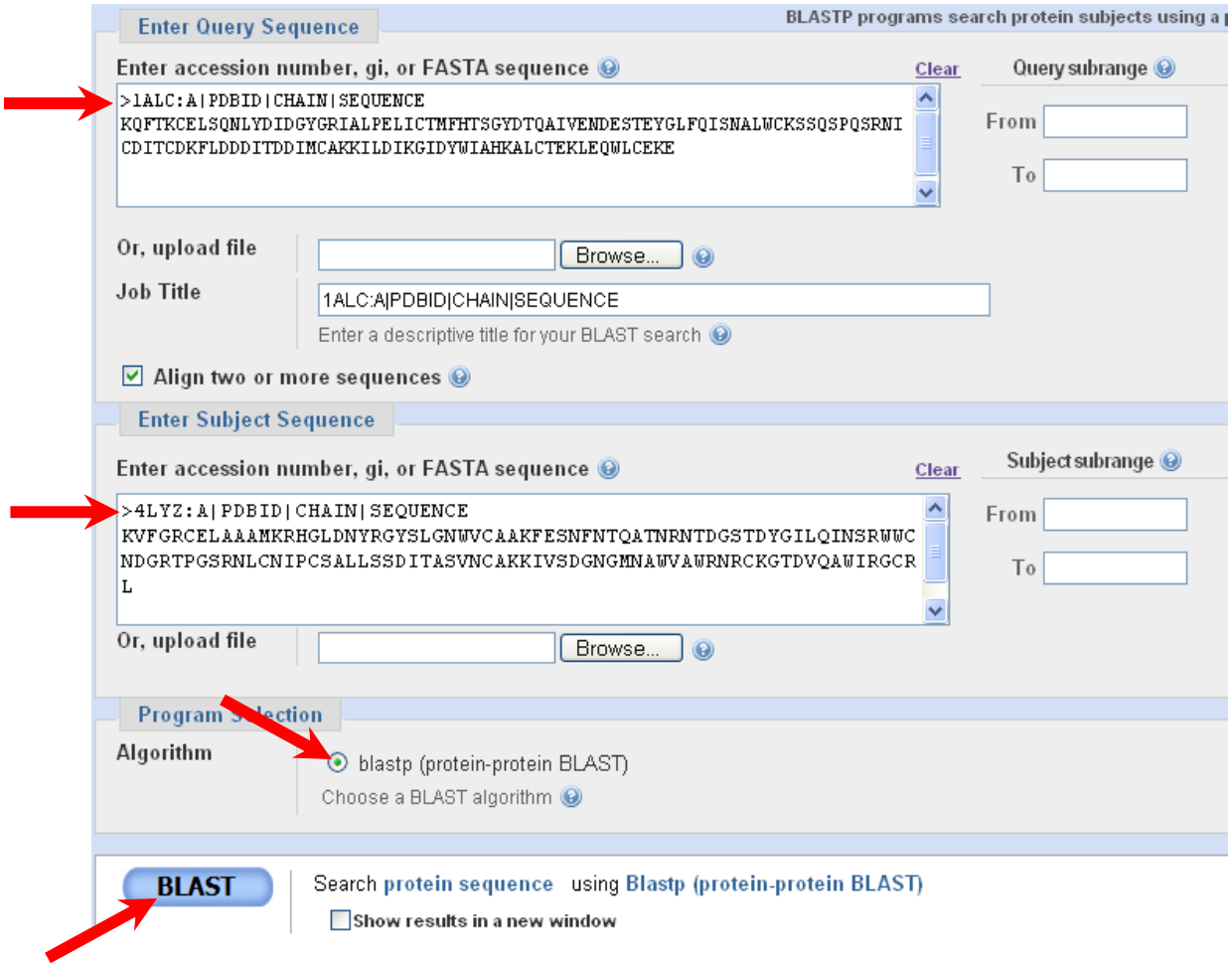
**Query Coverage: By the percent of length coverge for the query**

**E-value: expected number of chance matches in a random model**

# Alignment scores

```
>gb|AAA36188.1| lysozyme precursor (EC 3.2.1.17)
Length=148

 GENE ID: 4069 LYZ | lysozyme (renal amyloidosis) [Homo sapiens]
(Over 10 PubMed links)

 Score =  303 bits (775),  Expect = 4e-81, Method: Compositional matrix adjust.
 Identities = 147/148 (99%), Positives = 148/148 (100%), Gaps = 0/148 (0%)

Query  1    MKALIVLGLVLLSVTVQGKVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTRA  60
            MKALIVLGLVLLSVTVQGKVFERCELARTLKRLGMDGYRG+SLANWMCLAKWESGYNTRA
Sbjct  1    MKALIVLGLVLLSVTVQGKVFERCELARTLKRLGMDGYRGMSLANWMCLAKWESGYNTRA  60

Query  61   TNYNAGDRSTDYGIFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVRD  120
            TNYNAGDRSTDYGIFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVRD
Sbjct  61   TNYNAGDRSTDYGIFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVRD  120

Query  121  PQGIRAWVAWRNRCQNRDVRQYVQGCGV  148
            PQGIRAWVAWRNRCQNRDVRQYVQGCGV
Sbjct  121  PQGIRAWVAWRNRCQNRDVRQYVQGCGV  148
```

# Sequence similarity



# Sequence similarity

Blast 2 sequences

## 1ALC:A|PDBID|CHAIN|SEQUENCE

| | | | |
|---|---|---|---|
| Query ID | lcl\|62163 | Subject ID | 62165 |
| Description | 1ALC:A\|PDBID\|CHAIN\|SEQUENCE | Description | 4LYZ:A\|PDBID\|CHAIN\|SEQUENCE |
| Molecule type | amino acid | Molecule type | amino acid |
| Query Length | 123 | Subject Length | 129 |
| | | Program | BLASTP 2.2.21+ ▸Citation |

Other reports: ▸Search Summary [Taxonomy reports]

▸ **Graphic Summary**

▸ **Dot Matrix View**

▿ **Descriptions**

```
                                                          Score     E
Sequences producing significant alignments:              (Bits)  Value

lcl|62165  4LYZ:A|PDBID|CHAIN|SEQUENCE                    89.4    2e-23
```

▿ **Alignments** Select All Get selected sequences NEW

```
>lcl|62165 4LYZ:A|PDBID|CHAIN|SEQUENCE
Length=129

 Score = 89.4 bits (220),  Expect = 2e-23, Method: Compositional matrix adjust.
 Identities = 44/115 (38%), Positives = 67/115 (58%), Gaps = 4/115 (3%)

Query  1    KQFTKCELSQNL--YDIDGYGRIALPELICTMFHTSGYDTQAIVEN-DESTEYGLFQISN  57
            K F +CEL+  +  + +D Y   +L   +C     S ++TQA   N D ST+YG+ QI++
Sbjct  1    KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINS  60

Query  58   ALWCKSSQSPQSRNICDITCDKFLDDDITDDIMCAKKIL-DIKGIDYWIAHKALC  111
              WC   ++P SRN+C+I C   L  DIT  + CAKKI+ D  G++ W+A +  C
Sbjct  61   RWWCNDGRTPGSRNLCNIPCSALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRC  115
```

# How far hemoglobins in human and chicken are similar?

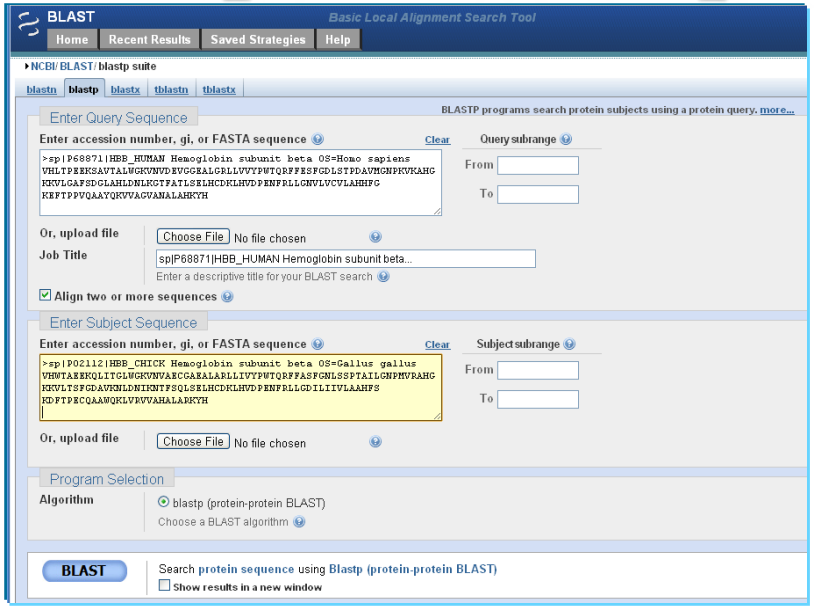
Get the sequences and compare the amino acids one by one

>sp|P68871|HBB_HUMAN Hemoglobin subunit beta OS=Homo sapiens

VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPD
AVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFR
LLGNVLVCVLAHHFG KEFTPPVQAAYQKVVAGVANALAHKYH

>sp|P02112|HBB_CHICK Hemoglobin subunit beta OS=Gallus gallus
VHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSP
TAILGNPMVRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENF
RLLGDILIIVLAAHFS KDFTPECQAAWQKLVRVVAHALARKYH

# Comparing Human and Chicken protein sequences




Bioinformatics program BLAST

Give sequences

1. Human
2. Chicken

```
 Identities = 102/147 (69%), Positives = 121/147 (82%), Gaps = 0/147 (0%)

Query  1    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
            MVH T EEK  +T LWGKVNV E G EAL RLL+VYPWTQRFF SFG+LS+P A++GNP
Sbjct  1    MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPM  60

Query  61   VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG  120
            V+AHGKKVL +F D + +LDN+K TF+ LSELHCDKLHVDPENFRLLG++L+ VLA HF
Sbjct  61   VRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFS  120

Query  121  KEFTPPVQAAYQKVVAGVANALAHKYH  147
            K+FTP  QAA+QK+V  VA+ALA KYH
Sbjct  121  KDFTPECQAAWQKLVRVVAHALARKYH  147
```

EMBL-EBI | EB-eye Search | All Databases | Enter Text Here | Go | Reset | Advanced Search | Give us feedback

Databases | Tools | EBI Groups | Training | Industry | About Us | Help | Site Index

- Help Index
- General Help
- Formats
- Gaps
- Matrix
- References
- FASTA Help
- MView Help
- VisualFASTA Help

- View all FASTA's at EBI
- FASTA Programmatic Access

- Database Information
  - UniProt
  - UniParc

- Similar Applications
  - FASTA
  - BLAST
  - GGSEARCH
  - GLSEARCH
  - MPsrch
  - ScanPS
  - SSEARCH

**FASTA Related Literature**

Search for FASTA related literature in Medline...
more

EBI > Tools > Similarity & Homology

## FASTA/SSEARCH/GGSEARCH/GLSEARCH - Protein Similarity Search

Provides sequence similarity searching against protein databases using the FASTA and SSEARCH programs. **SSEARCH** does a rigorous Smith-Waterman search for similarity between a query sequence and a database. **GGSEARCH** compares a protein or DNA sequence to a sequence database producing global-global alignment (Needleman-Wunsch). **GLSEARCH** compares a protein or DNA sequence to a sequence database. **FASTA** can be very specific when identifying long regions of low similarity especially for highly diverged sequences. You can also conduct sequence similarity searching against nucleotide databases or complete proteome/genome databases using the FASTA programs.

**Download Software**

| PROGRAM | DATABASES | RESULTS | SEARCH TITLE | YOUR EMAIL |
|---|---|---|---|---|
| FASTA | Protein (UniProt Knowledgebas; UniProtKB/Swiss-Prot; UniProt Clusters 100%) | interacti | Sequence | |

| MATRIX | GAP OPEN | GAP EXTEND | KTUP | EXPECTATION UPPER VALUE | EXPECTATION LOWER VALUE |
|---|---|---|---|---|---|
| BLOSUM | -10 | -2 | 2 | 10.0 | default |

| DNA STRAND | HISTOGRAM | MOLECULE TYPE |
|---|---|---|
| none | no | Protein |

| SCORES | ALIGNMENTS | SEQUENCE RANGE | DATABASE RANGE | FILTER | STATISTICAL ESTIMATES |
|---|---|---|---|---|---|
| 50 | 50 | START-END | START-END | none | Regress |

Enter or Paste a PROTEI Sequence in any format: Help

```
>2LZM:A|PDBID|CHAIN|SEQUENCE
MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRNC
NGVITKDEAEKLFNQDVDAAVRGILR
NAKLKPVYDSLDAVRRCALINMVFQMGETGVAGFTNSLRMLQQKRWDEAAVNLA
KSRWYNQTPNRAKRVITTFRTGTWDA
YKNL
```

Upload a file: Browse... Run Reset

## FASTA Results

| SUBMISSION PARAMETERS | | | |
|---|---|---|---|
| Title | Sequence | Database | uniprot |
| Sequence length | 164 | Sequence type | p |
| Program | fasta | Version | 35.04 July. 4, 2009 |
| Expectation upper value | 10.0 | Matrix | BL50 |
| Sequence range | 1- | Number of scores | 50 |
| Number of alignments | 50 | Word size | 2 |
| Open gap penalty | -10 | Gap extension penalty | -2 |
| Histogram | false | | |

Show Annotation | FASTA Result | MView | VisualFasta | XML | SUBMIT ANOTHER JOB

Clear all | Check all | Invert selection | Reset | Show Alignments | Download | fasta

| Alignment | DB:ID | Source | Length | Identity% | Similar% | Overlap | E() |
|---|---|---|---|---|---|---|---|
| 1 | UNIPROT:LYS_BPT4 | Lysozyme OS=Enterobacteria pha | 164 | 100.0 | 100.0 | 164 | 2.3e-67 |
| 2 | UNIPROT:C3V1I9_9CAUD | Soluble lysozyme OS=Entero | 164 | 98.8 | 100.0 | 164 | 1.2e-66 |
| 3 | UNIPROT:Q7M2A4_BPT2 | Lysozyme OS=Enterobacteria | 164 | 98.2 | 100.0 | 164 | 3.1e-66 |
| 4 | UNIPROT:Q06EK2_BPR32 | Lysozyme OS=Enterobacteria | 164 | 98.2 | 100.0 | 164 | 3.1e-66 |
| 5 | UNIPROT:C3V2B5_BPR51 | Soluble lysozyme OS=Entero | 164 | 98.2 | 100.0 | 164 | 3.1e-66 |
| 6 | UNIPROT:Q7Y2B5_BPR69 | Lysozyme OS=Enterobacteria | 157 | 82.7 | 93.6 | 156 | 2.8e-53 |
| 7 | UNIPROT:A8R9C2_9CAUD | Lysozyme OS=Enterobacteria | 162 | 75.3 | 91.4 | 162 | 9.6e-50 |
| 8 | UNIPROT:C4MZK9_9CAUD | E Lysozyme murein hydrolas | 162 | 75.3 | 91.4 | 162 | 9.6e-50 |
| 9 | UNIPROT:LYST1_DICDI | Probable T4-type lysozyme 1 | 170 | 46.7 | 76.6 | 167 | 1.1e-26 |
| 10 | UNIPROT:LYST2_DICDI | Probable T4-type lysozyme 2 | 170 | 45.1 | 76.8 | 164 | 7.8e-25 |
| 11 | UNIPROT:Q56EM5_9CAUD | Lysozyme OS=Aeromonas phag | 600 | 42.9 | 70.6 | 163 | 2e-21 |
| 12 | UNIPROT:Q6U9G4_9CAUD | Lysozyme OS=Aeromonas phag | 600 | 42.9 | 71.2 | 163 | 2.4e-21 |
| 13 | UNIPROT:Q19CF2_9CAUD | Lysozyme OS=Aeromonas phag | 164 | 42.7 | 68.3 | 164 | 2.5e-21 |
| 14 | UNIPROT:A7XF92_9CAUD | Lysozyme OS=Enterobacteria | 599 | 43.2 | 69.8 | 162 | 1e-20 |
| 15 | UNIPROT:C4MYV6_9CAUD | Baseplate hub subunit and | 599 | 43.2 | 69.8 | 162 | 1e-20 |

# Multiple sequence alignment

A **multiple sequence alignment** (MSA) is a sequence alignment of **three or more** biological sequences, generally **protein, DNA, or RNA**.

The input set of query sequences are assumed to have an **evolutionary relationship** by which they share a lineage and are descended from a **common ancestor**.

Multiple sequence alignment is a 2D table in which the rows represent individual sequences and the columns the residue positions.

Sequences are laid on the grid in such a way that (i) the relative positioning of residues within any sequence is preserved and (ii) similar residues in all sequences are brought into vertical register.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| I | Y | D | G | G | A | V | - | E | A | L |
| II | Y | D | G | G | - | - | - | E | A | L |
| III | F | E | G | G | I | L | V | E | A | L |
| IV | F | D | - | G | I | L | V | Q | A | V |
| V | Y | E | G | G | A | V | V | Q | A | L |
| **Consensus** | y | d | G | G | A/I | V/L | V | e | A | l |

# Computational complexity

**Pairwise alignment techniques generally use processing time and memory space related to the products of the lengths of the sequences being compared [($O(m_1m_2)$, where O is the order of the time taken and m: sequence lengths).**

**By extending the pairwise comparison to three dimensions, we have the complexity of $O(m_1m_2m_3)$**

**For n sequences it will be $O(m^n)$, n is the number of sequences and m is the length of the sequences.**

**Time taken to compute an alignment rises exponentially with the number of mn sequences that are to be aligned.**

# Clustal

The Clustal approach exploits the fact that similar sequences are likely to be evolutionarily related.

Clustal aligns sequences in pairs, following the branching order of a family tree.

Similar sequences are aligned first, and more distantly related sequences are added later.

Once pairwise alignment scores for each sequences relative to all others have been calculated, they are used to cluster the sequences into groups, which are then aligned against each other to generate the final multiple alignment.

ClustalW uses the positioning of gaps in closely related sequences to guide the insertion of gaps into those that are more distant.

Information compiled during the alignment process about the variability of the most similar sequences is used to help vary the gap penalties on a residue and position specific basis.

## ClustalW2

ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms.

New users, please read the FAQ.

**>> Download Software**

| YOUR EMAIL | ALIGNMENT TITLE | RESULTS | ALIGNMENT | |
|---|---|---|---|---|
| | Sequence | interactive | full | |
| **KTUP (WORD SIZE)** | **WINDOW LENGTH** | **SCORE TYPE** | **TOPDIAG** | **PAIRGAP** |
| def | def | percent | def | def |
| **MATRIX** | **GAP OPEN** | **NO END GAPS** | **GAP EXTENSION** | **GAP DISTANCES** |
| def | def | yes | def | def |
| | **ITERATION** | | **NUMITER** | |
| | none | | 1 | |

| OUTPUT | | PHYLOGENETIC TREE | | | |
|---|---|---|---|---|---|
| **OUTPUT FORMAT** | **OUTPUT ORDER** | **TREE TYPE** | **CORRECT DIST.** | **IGNORE GAPS** | **CLUSTERING** |
| aln w/numbers | aligned | none | off | off | NJ |

Enter or paste a set of sequences in any supported format: Help

```
>479227|Genbank|Outer membrane integral membrane
protein|outD protein
MLLLSGSVLLMASSLAWSAEFSASFKGTDIQEFINTVSKNLNKTVIIDPSVSGT
ITVRSY
DMMNEEQYYQFFLSVLDVYGFTVIPMDNNVLKIIRSKDAKSTSMPLATDRQPGI
GDEVVT
RVVPVNNVAARDFGRSSRVERQRVAWDVWRLRTCERRRDDWPRGVIHAVMTIVE
RVDQTG
DRNVTTIPLSYASSTEVVKMVNELNKMDEKSALPGMLTANVVADERTNSAAGFG
EPNSRQ
```

Upload a file: Browse... Run Reset

| Sequence | Alignment | $R_p$ |
|---|---|---|
| 479227\|Genbank\|Outer | --FKGTDIQEFINTVS-KNLNKTVIIDPS-VSGTITVRSYDMMNEEQ--- | 67 |
| P31701\|SwissProt\|Outer | --FKGTDIQEFINTVS-KNLNKTVIIDPS-VSGTITVRSYDMMNEEQ--- | 67 |
| P31700\|SwissProt\|Outer | --FKGTDIQEFINTVS-KNLNKTVIIDPT-VRGTISVRSYDMMDEGQ--- | 76 |
| Q01565\|SwissProt\|Outer | --FKGTDIQEFINTVS-KNLNKTVIIDPT-VRGTISVRSYDMMNEGQ--- | 76 |
| 7466966\|Genbank\|Outer | --FKDTDIQEFINTVS-KNLHKTVIINPD-VQGTITVRSYDMLNEEQ--- | 63 |
| P15644\|SwissProt\|Outer | --FKGTDIQEFINTVS-KNLNKTVIIDPS-VRGTITVRSYDMLNEEQ--- | 76 |
| P31780\|SwissProt\|Outer | --FKNADIEEFINTVG-KNLSKTIIIEPS-VRGKINVRSYDLLNEEQ--- | 74 |
| P45778\|SwissProt\|Outer | --FKNADIEEFINTVG-KNLSKTIIIEPS-VRGKINVRSYDLLNEEQ--- | 74 |
| P45779\|SwissProt\|Outer | --FKGTDIQEFINIVG-RNLEKTIIVDPS-VRGKVDVRSFDTLNEEQ--- | 73 |
| 4139236\|Genbank\|Outer | --FVNADIDQVAKAIG-AATGKTIIVDPR-VKGQLNLVAERPVPEDQ--- | 69 |
| 11559475\|Genbank\|Outer | --FVNADIDQVAKAIG-AATGKTIIVDPR-VKGQLNLVAERPVPEDQ--- | 76 |
| 3978475\|Genbank\|Outer | INMKDADIRDFIDQVA-QISGETFVVDPR-VKGQVTVISKTPLGLEE--- | 91 |
| P35818\|SwissProt\|Outer | INLKDADIREFIDQIS-EITGETFVVDPR-VKGQVSVVSKAQLSLSE--- | 99 |
| 11352555\|Genbank\|Outer | INMKDAEIGDFIEQVS-SISGQTFVVDPR-VKGRVTVVSQARLSLAE--- | 92 |
| 15597065\|Genbank\|Outer | INMKDAEIGDFIEQVS-SISGQTFVVDPR-VKGRVTVVSQARLSLAE--- | 92 |
| 2120685\|Genbank\|Outer | VNFVDTELGEFIDSVS-RITGTTFIVDPR-VKGKVTVRTVDLHDADA--- | 83 |
| 13421292\|Genbank\|Outer | LNVQDADIRVFIQDVA-KSTGTTFIIDPR-VKGTVTVASNGPLNRRE--- | 81 |
| 7469078\|Genbank\|Outer | --FEDISILELLQFVS-KISGTNFVFDSNDLQFNVTIVSHDPTSVDD--- | 312 |
| 11362809\|Genbank\|Outer | --FEDISILELLQFVS-KISGTNFVFDSNDLQFNVTIVSHDPTSVDD--- | 63 |
| 7468594\|Genbank\|Outer | --FEDISILELLQFVS-KISGTNFVFDSNDLQFNVTIVSHDPTSVDD--- | 310 |
| 11360974\|Genbank\|Outer | FNFEGESLQAVVKAILGDMLGQNYFIASG-VQGTVTLSTPKPVSSAQ--- | 153 |
| P29041\|SwissProt\|Outer | FNFEGESVQAVVKAILGDMLGQNYVIAPG-VQGTVTLATPNPVSPAQ--- | 141 |
| 13475694\|Genbank\|Outer | LNLVNAPIADAAKAVLGDALHLNYIVDPR-VQGTVTLQTSQPVSQDA--- | 139 |
| 12721580\|Genbank\|Outer | ----------MWRAFRKIS-LVYFLCGVAYVGS--- | 22 |
| P31772\|SwissProt\|Outer | ----------MKKYFLKCGYFLVCFCLPLIVFA--- | 23 |
| 11354911\|Genbank\|Outer | -ALCASSMVFSAESATANQLENIDFRVNKEKAAVLIVELASPSAVVD--- | 73 |
| P34749\|SwissProt\|Outer | ----------MKQWIAALLLMLIPGVQAAKP--- | 21 |
| 11360960\|Genbank\|Outer | -AENKQAIPVPKVANAPLSVSKIDFKRGDDGSGRLILKFDGQGATPD--- | 93 |
| P34750\|SwissProt\|Outer | -ASYAQPIKPKPYVPAGRAIRNIDFQRGEKGEGNVVIDLSDPTLSPD--- | 190 |
| 4027986\|Genbank\|Outer | SAKQQAAAPAKQQAAAPAKQTNIDFRKDGKNAGIIELAALGFAGQPD--- | 260 |
| 11353851\|Genbank\|Outer | SAKQQTAAPAKQQAATPAKQTNIDFRKDGKNAGIIELAALGFAGQPD--- | 244 |
| 4768955\|Genbank\|Outer | ---------- | |
| 12721161\|Genbank\|Outer | ---------- | |
| 7208425\|Genbank\|Outer | VNLPSVKASMSASRRLLTASVAALLALTSTAPVFADGPIGGSHTYRP--- | 51 |
| 13474660\|Genbank\|Outer | ---------- | |
| 12620550\|Genbank\|Outer | --MKVLNNAAGATQPPAITNPQRSAALNRLCYLLCG---------- | 34 |
| 13475294\|Genbank\|Outer | ---------- | |
| 7468922\|Genbank\|Outer | ALWNHPETTIYNLVSDYGDEQSIYVIPQNVGAMRITAMSKLVVPKEG--- | 164 |
| 11360973\|Genbank\|Outer | ALWNHPETTIYNLVSDYGDEQSIYLIPQNVGAMRITAMSKLVVPKEG--- | 164 |
| 7468239\|Genbank\|Outer | ALWNHPETTIYNLVTDYGTEDSIYLIPQEIGAIKIATLSKFVVPKES--- | 165 |

# Exercise

**Get the multiple alignment for hemoglobin A chain from different organisms.**

# Steps

1. Search for hemoglobin A chain sequences
2. Select relevant entries
3. Get the sequences in FASTA format (view or save).
4. Give the sequences as input for ClustalW

# Other software

**MAFFT**

**http://mafft.cbrc.jp/alignment/software/**

**MUSCLE**

**http://www.ebi.ac.uk/Tools/msa/muscle/**

**PROMOLS**

**http://prodata.swmed.edu/promals/promals.php**

# PSI-BLAST

**PSI-BLAST (Position-Specific Iterative BLAST) is a program that searches database of sequences similar to query sequence.**

**PSI-BLAST begins with search results obtained with BLAST and derives pattern information from a multiple sequence alignment of the initial hits.**

**It repeats the process and fine-tuning the pattern in successive cycles.**

# PSI-BLAST

## PSI-BLAST - Protein Similarity Search

PSI-BLAST is similar to NCBI BLAST except that it uses position-specific scoring matrices derived during the search, this tool is used to detect distant evolutionary relationships.

**Use this tool**

**STEP 1 - Select your database**

PROTEIN DATABASES

UniProt Knowledgebase

**STEP 2 - Enter your input sequence**

Enter or paste a PROTEIN sequence in any supported format:

```
>sp|P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapiens GN=HBA1
PE=1 SV=2
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
```

Upload a file: Choose File No file chosen

**STEP 3 - Set your parameters**

PSI-BLAST THRESHOLD

1.0e-3

*The default settings will fulfill the needs of most users and, for that reason, are not visible.*

More options... *(Click here, if you want to view or change the default settings.)*

**STEP 4 - Submit your job**

☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

## PSI-BLAST Results

Summary Table | Tool Output | Visual Output | Functional Predictions | Submission Details | Submit Another Job

### PSI-BLAST Iteration

**Threshold:** 1.0e-3 Run next iteration

### Alignments

**Selection:** Show Annotations | Hide Annotations | Show Alignments | Hide Alignments

Download in fasta format

Clear Selection | Select All | Invert Selection

| Align. | DB:ID | Source | Length | Score | Identities | Positives | E() |
|---|---|---|---|---|---|---|---|
| 1 New | TR:Q5R9M5_PONAB | Putative uncharacterized protein DKFZp468J1717 OS=Pongo abelii GN=DKFZp468J1717 PE=2 SV=1<br>*Cross-references and related information in:*<br>▶Nucleotide Sequences ▶Genomes ▶Ontologies ▶Protein Families ▶Protein Sequences | 142 | 733 | 100.0 | 100.0 | 5.0E-76 |
| 2 New | TR:D1MGQ2_HUMAN | Alpha-2 globin chain OS=Homo sapiens GN=HBA2 PE=3 SV=1<br>*Cross-references and related information in:*<br>▶Nucleotide Sequences ▶Genomes ▶Ontologies ▶Protein Families ▶Literature ▶Protein Sequences | 142 | 733 | 100.0 | 100.0 | 5.0E-76 |
| 3 New | SP:HBA_PANTR | Hemoglobin subunit alpha OS=Pan troglodytes GN=HBA1 PE=1 SV=2<br>*Cross-references and related information in:*<br>▶Nucleotide Sequences ▶Ontologies ▶Protein Families ▶Literature ▶Protein Sequences | 142 | 733 | 100.0 | 100.0 | 5.0E-76 |
| 4 New | SP:HBA_PANPA | Hemoglobin subunit alpha OS=Pan paniscus GN=HBA1 PE=1 SV=2<br>*Cross-references and related information in:*<br>▶Ontologies ▶Protein Families ▶Literature ▶Protein Sequences | 142 | 733 | 100.0 | 100.0 | 5.0E-76 |

# Conservation score

In biology, **conserved sequences** are similar or identical sequences that occur within nucleic acid or protein sequences.

The conservation score at a site corresponds to the site's evolutionary rate.

The rate of evolution is not constant among amino acid sites: some positions evolve slowly and are commonly referred to as "conserved", while others evolve rapidly and are referred to as "variable".

The conservation score for all the residues in a protein can be obtained by comparing the sequence of a PDB chain with the proteins deposited in Swiss-Prot/Uniprot and finds the ones that are homologous to the PDB sequence.

# Algorithm development

## Multiple sequence alignment

```
sp|P69905|     MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG 60
sp|P69907|     MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG 60
sp|P06635|     MVLSPADKTNVKTAWGKVGAHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKDHG 60
sp|P01958|     MVLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG 60
sp|P01959|     MVLSAADKTNVKAAWSKVGGNAGEFGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG 60
sp|P01965|     -VLSAADKANVKAAWGKVGGQAGAHGAEALERMFLGFPTTKTYFPHFNLSHGSDQVKAHG 59
sp|P01966|     MVLSAADKGNVKAAWGKVGGHAAEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG 60
sp|P60529|     -VLSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHFDLSPGSAQVKAHG 59
sp|P01942|     MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHG 60
sp|P01946|     MVLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFSHIDVSPGSAQVKAHG 60
```

# Steps

**1. Position specific amino acid frequencies**

**(a) Unweighted amino acid frequencies**

**(b) Weighted amino acid frequencies**

**(c) Estimated independent counts**

**2. Calculation of conservation scores**

**(a) Entropy based measure**

**(b) Variance based measure**

**(c) Sum of pairs measure**

**Bioinformatics 17, 700-711 (2001)**

# Unweighted amino acid frequencies

$f_a^{\mathbf{u}}(i) = n_a(i)/n(i)$

$n_a(i)$:number of sequences in which position i is occupied by amino acid a

n(i): total number of aligned sequences

$= \quad n_a(i), i=1,20$

For specific group of sequences

# Weighted amino acid frequencies

$f_a^w(i) = \delta(a,k,i)\, w_k / \; w_k,\ k=1,n(i)$

$w_k$: weight of a sequence k

$\delta(a,k,i) = 1$, if amino acid a is in sequence k at position i

$\delta(a,k,i) = 0$, otherwise

Correct for unequal distances between different sequence pairs

# Estimated independent counts

$f_a^{ic}(i) = n_a^{ic}(i)/n^{ic}(i)$

Probable independent observation with respect to random distribution; E.g. probability of a at position i (F)

20 amino acid residues with random distribution,

$F = 20\,(1-0.95^N)$

$N_{eff} = \ln(1-F/20)/\ln 0.95$,

Any F, amino acid a at position i in a single sequence, $n_a^{ic}(i) = 1$

More sequences: $n_a^{ic}(i) = 1$ (identical)

$$n_a^{ic}(i) = n_a(i)$$

Correlation between aligned sequences

$F = 20\,(1-0.95^N)$

If, N= 1, F = 1; True

Assume N = n,

f(n,i) probability of i different amino acids to occur at the position

$F = f(n,i)i = 20(1-0.95^n)$

If N=n+1, number of amino acid is the same with the probability of i/20 (adding additional one with existing one) or i+1 with probability (20-i)/20.

$F = [i(i/20)+(i+1)(20-i)/20]\, f(n,i)$

$= (1+0.95i)\, f(n,i)$

$= 1+0.95*20(1-0.95^n)$

$= 20\,(1-0.95^{n+1})$

# Conservation index

**Entropy based measure**

$$C^e(i) = f_a(i).\ln f_a(i),\ a = 1,20$$

**Order of a system can be measured with entropy**

**Measure for sequence variability**

**Maximal if all 20 amino acids have equal frequencies**

**Not biased with amino acid composition or similarities among amino acids**

# Conservation index

**Variance based measure**

$$C^v(i) = \{ \quad [f_a(i) - f_a]^2\}^{0.5}, \; a = 1,20$$

**$f_a$: overall frequency for amino acid a in the alignment**

**$= \quad n_a(i)/ \; n(i), \; i=1,l$ , l: total number of aligned positions (unweighted)**

**$= \quad \delta(a,k,i) \; w_k/ \quad w_k, \; i=1,l$ , $k=1,n(i)$(weighted)**

**Advantages: overall amino acid frequencies, which differ for different protein families.**

# Conservation index

**Sum of pairs measure**

$$C^p(i) = \{ \quad [f_a(i)\, f_b(i)\ \ S_{ab}]^2\}^{0.5},\ a = 1,20;\ b=1,20$$

**$S_{ab}$: amino acid scoring matrix**

**Conservation score will be higher for the positions occupied by simiar amino acids**

**Depends on amino acid type**

# Normalization

$C_n(i) = (C(i) - C')/ \ _c$

$C' = \ C(i)/l, \ i=1,l$

$_C = [ \ (C(i)-C')^2/(l-1)]^{0.5}$

**Bioinformatics 17, 700-711 (2001)**

# Example

## Unweighted frequency

**Position 3**

$f_a(L) = 10/10 = 1$

**Position 6**

$f_a(A) = 8/10 = 0.8$

$f_a(D) = 1/10 = 0.1$

$f_a(E) = 1/10 = 0.1$

## Entropy based measure

**Position 3**

$c(i) = 1 \ln 1$

$= 0$

**Position 6**

$c(i) = 0.8 \ln 0.8 + 0.1 \ln 0.1 + 0.1 \ln 0.1$

$= 0.8 * -0.223 + 0.1 * -2.302 + 0.1 * -2.302$

$= -0.1784 - 0.23 - 0.23$

$= -0.638$

## AL2CO sequence conservation analysis server

**The AL2CO program calculates positional conservation for a multiple sequence alignment.**
**[Documentation]**

### DATA INPUT

**Enter protein sequence alignment in CLUSTAL format:** Clear sequences

```
sp|P69905|    MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG 60
sp|P69907|    MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG 60
sp|P06635|    MVLSPADKTNVKTAWGKVGAHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKDHG 60
sp|P01958|    MVLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG 60
sp|P01959|    MVLSAADKTNVKAAWSKVGGNAGEFGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG 60
sp|P01965|    -VLSAADKANVKAAWGKVGGQAGAHGAEALERMFLGFPTTKTYFPHFNLSHGSDQVKAHG 59
sp|P01966|    MVLSAADKGNVKAAWGKVGGHAAEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG 60
sp|P60529|    -VLSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHFDLSPGSAQVKAHG 59
sp|P01942|    MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHG 60
sp|P01946|    MVLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFSHIDVSPGSAQVKAHG 60
```

Or upload a file Choose File No file chosen

### DATA SUBMIT

Enter **email** to receive the result (optional):

Enter a **job name** (optional): Submit Reset

### PARAMETERS

- sequence weighting scheme: ○ henikoff-henikoff ○ independent count ◉ unweighted
- conservation calculation method: ◉ entropy ○ variance ○ sum-of-pairs
- For sum-of-pairs method only:
  scoring matrix:
  ◉ BLOSUM62 matrix ○ identity matrix
  scoring matrix transformation :
  ◉ no transformation ○ normalization ○ adjustment
- normalize conservation values: ◉ True ○ False

**http://prodata.swmed.edu/al2co/al2co.php**

## AL2CO alignment conservation server

---

**RESULTS**:

- The list of positional conservation values is here.
- The alignment with integer conservation indices is here.

**INPUTS:**

- Input alignment is here.
- Input pdb file: none.

| | | |
|---|---|---|
| 1 | M | 0.000 |
| 2 | V | 0.000 |
| 3 | L | 0.000 |
| 4 | S | 0.000 |
| 5 | P | -0.943 |
| 6 | A | -0.639 |
| 7 | D | 0.000 |
| 8 | K | 0.000 |
| 9 | T | -0.940 |
| 10 | N | 0.000 |
| 11 | V | -0.611 |
| 12 | K | 0.000 |
| 13 | A | -0.940 |
| 14 | A | -0.639 |
| 15 | W | 0.000 |
| 16 | G | -0.802 |
| 17 | K | 0.000 |
| 18 | V | -0.611 |
| 19 | G | 0.000 |
| 20 | A | -0.611 |
| 21 | H | -0.639 |
| 22 | A | -0.500 |
| 23 | G | -0.500 |
| 24 | E | -0.802 |
| 25 | Y | -0.639 |
| 26 | G | 0.000 |
| 27 | A | -0.639 |
| 28 | E | 0.000 |
| 29 | A | 0.000 |
| 30 | L | 0.000 |

| | | |
|---|---|---|
| 1 | M | 0.796 |
| 2 | V | 0.796 |
| 3 | L | 0.796 |
| 4 | S | 0.796 |
| 5 | P | -1.958 |
| 6 | A | -1.070 |
| 7 | D | 0.796 |
| 8 | K | 0.796 |
| 9 | T | -1.950 |
| 10 | N | 0.796 |
| 11 | V | -0.987 |
| 12 | K | 0.796 |
| 13 | A | -1.950 |
| 14 | A | -1.070 |
| 15 | W | 0.796 |
| 16 | G | -1.545 |
| 17 | K | 0.796 |
| 18 | V | -0.987 |
| 19 | G | 0.796 |
| 20 | A | -0.987 |
| 21 | H | -1.070 |
| 22 | A | -0.665 |
| 23 | G | -0.665 |
| 24 | E | -1.545 |
| 25 | Y | -1.070 |
| 26 | G | 0.796 |
| 27 | A | -1.070 |
| 28 | E | 0.796 |
| 29 | A | 0.796 |
| 30 | L | 0.796 |

**(x-X')/**

Weighted

| | | |
|---|---|---|
| 1 | M | 0.000 |
| 2 | V | 0.000 |
| 3 | L | 0.000 |
| 4 | S | 0.000 |
| 5 | P | -0.956 |
| 6 | A | -0.815 |
| 7 | D | 0.000 |
| 8 | K | 0.000 |
| 9 | T | -0.966 |
| 10 | N | 0.000 |
| 11 | V | -0.689 |
| 12 | K | 0.000 |
| 13 | A | -1.121 |
| 14 | A | -0.875 |
| 15 | W | 0.000 |
| 16 | G | -0.815 |
| 17 | K | 0.000 |
| 18 | V | -0.689 |
| 19 | G | 0.000 |
| 20 | A | -0.498 |
| 21 | H | -0.663 |
| 22 | A | -0.613 |
| 23 | G | -0.482 |
| 24 | E | -0.897 |
| 25 | Y | -0.663 |
| 26 | G | 0.000 |
| 27 | A | -0.875 |
| 28 | E | 0.000 |
| 29 | A | 0.000 |
| 30 | L | 0.000 |

```
Conservation:        9999039909390391939334413939993969109999999969664949969990 99
sp|P69905|           MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
sp|P69907|           MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
sp|P06635|           MVLSPADKTNVKTAWGKVGAHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKDHG
sp|P01958|           MVLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
sp|P01959|           MVLSAADKTNVKAAWSKVGGNAGEFGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
sp|P01965|           -VLSAADKANVKAAWGKVGGQAGAHGAEALERMFLGFPTTKTYFPHFNLSHGSDQVKAHG
sp|P01966|           MVLSAADKGNVKAAWGKVGGHAAEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
sp|P60529|           -VLSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHFDLSPGSAQVKAHG
sp|P01942|           MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHG
sp|P01946|           MVLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFSHIDVSPGSAQVKAHG
```

**9: conserved**

**0: variable**

# The ConSurf Server

Method of calculating the amino acid conservation scores (obligatory)

Method: Max. Likelihood (ML)

Protein Structure (obligatory)

Enter the PDB ID: 4lyz

Or

Enter your own PDB file

Browse...

Chain Identifier: A

Please enter your e-mail address: michael-gromiha@aist.go.jp

We will use this address in case your job has exceeded its maximal running time.

Send a link to the results by e-mail

Providing your e-mail address ensures that you don't lose the URL to your results.

Submit Clear

http://consurf.tau.ac.il/.

>2LZM:A|PDBID|CHAIN|SEQUENCE
MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRNCNGVITKDEAEKLFN
QDVDAAVRGILRNAKLKPVYDSLDAVRRCALINMVFQMGETGVAGFTNSLRMLQQKRWDEAAVN
LAKSRWYNQTPNRAKRVITTFRTGTWDA YKNL

## ConSurf Color-Coded MSA

```
Input_pdb_SEQRES_A          M N I F E M L R I D E G L R L K I Y K D T E G Y Y T I G I G H L L T K S - P S L N A A K - - S E L D
UniRef90_Q7Y2B5_1_156       - - - - - M L R N D E G L R L T L Y K D T E G F W T I G I G H L V T K N - P S L A V A K - - A E L D
UniRef90_A8R9C2_1_162       M D I F D M L R Q D E G L D L N L Y K D T E G Y W T I G I G Q L V T K N - P S K D V A R - - A E L D
UniRef90_Q556F2_4_170       - S I K D M L K Y D E G E K L E M Y K D T E G Y Y T I G I G H L I T R I - K E R N A A I - - L S L E
UniRef90_Q86AA1_7_170       - - - - D M L K Y D E G E K L E M Y K D T E G N Y T I G I G H L I T K N - K D K N E A I - - K I L E
UniRef90_Q56EM5_184_346     V T I E D M L R Y D E G I R W V V Y W D S E G Y P T V G I G H L I I R E - K T K N M S R I N S L L S
UniRef90_Q19CF2_1_164       M T L E D M L I Y D E G R V L K V Y W D H L G Y P T I G I G H L I I P Q - K T T D M A L I N H T L S
UniRef90_Q8SDG3_181_342     - - I E K M L R G D E G Y R E K W Y L D S E G Y P T I G I G H L I I Y K - K T S D L G I I N N E L S
UniRef90_D5JFK5_180_341     - - I E K M L K Q D E G I R T R W Y T D S E G Y P T I G I G H L L I R E - K T R D T A K I N A A I S
UniRef90_Q56EE1_1_164       M T L E D M L V Y D E G R K L K V Y W D H L G Y P T V G I G H L I V L R - E T K D M G V I N H M L G
UniRef90_P16009_174_339     M S M A E M L R R D E G L R L K V Y W D T E G Y P T I G I G H L I M K Q - P V R D M A Q I N K V L S
UniRef90_Q76YA6_4_163       - - - - Q M L K Q D E G Y K E S V Y W D T E G Y P T I G I G H L I L R K - K T K D M G E I N R E L S
UniRef90_Q19CN7_158_320     V T I E D M L R Y D E G I R W S V Y W D S E G Y P T V G I G H L I V H E - K T R N M T R I N Q L L S
UniRef90_Q56BJ4_180_343     - - I E K M I R G D E G I R L T W Y Y D V K G Y - T I G I G H F F L T A P Q G T D P A V V N A A L S
UniRef90_C4MZP0_176_339     - - I E N M L H R D E G L R L K V Y W D T E G Y P T I G I G H L I T P Q - P I R D M N Q I N K I L S
UniRef90_Q7Y4Y4_176_339     - - I T E M L R R D E G L R D K V Y W D H L G Y P T V G I G H L I V M E - K T R D M S R I N K L L S
UniRef90_Q76YN5_188_349     - - I E K M L V Q D E G V R T K W Y L D S E G Y P T I G I G H L I I R E - R T S N L V T I N S I L S
UniRef90_A4SN28_23_148      - - - - - - - - F D N G M F L R - Y K D S L G Y W T I G Y G H L I K P N - E S Y - - - - - - - - - -
UniRef90_A9I970_8_140       - - - - - L L R G D E G E V L H A Y R D H L G Y L T I G V G R L I D K R - K G - - - - - - - - - - -
```

POS: The position of the AA in the SEQRES derived sequence.

SEQ: The SEQRES derived sequence in one letter code.

3LATOM: The ATOM derived sequence in three letter code, including the AA's positions as they appear in the PDB file and the chain identifier.

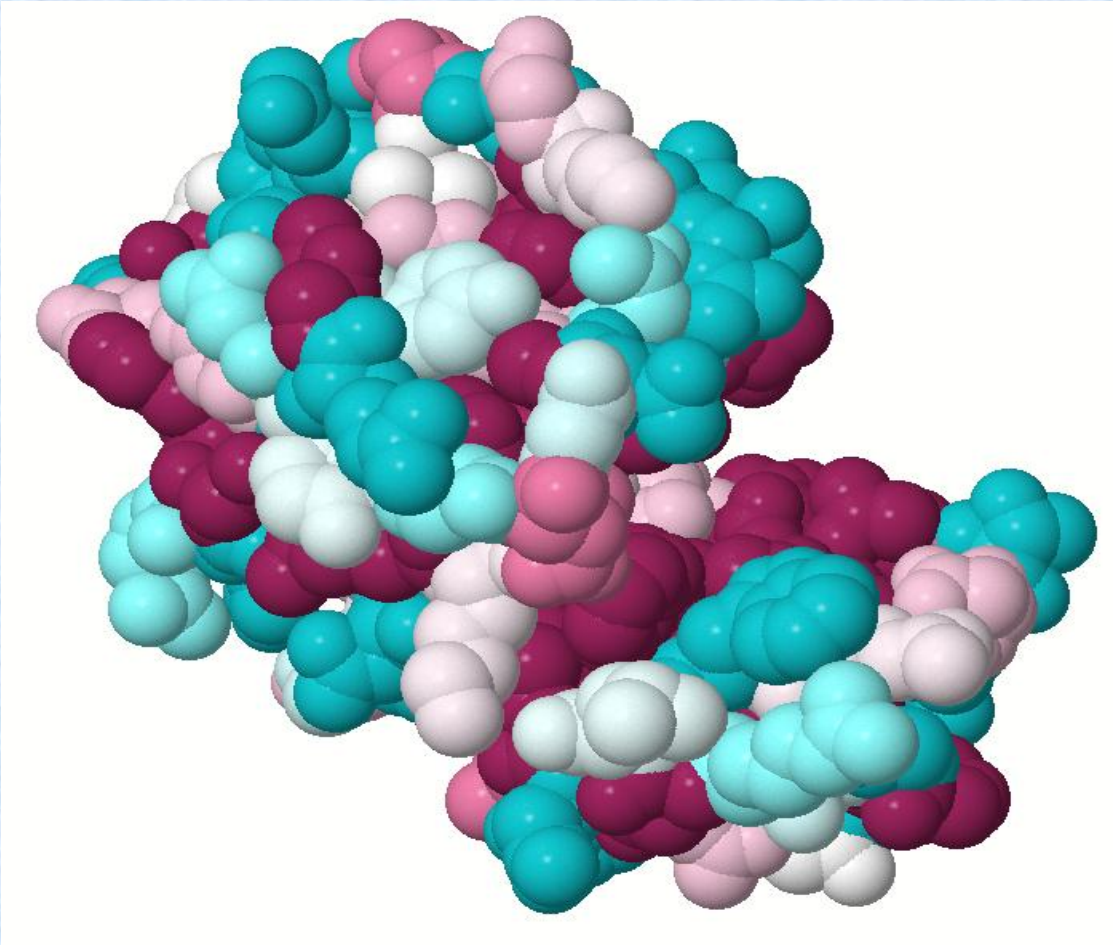
SCORE: The normalized conservation scores.

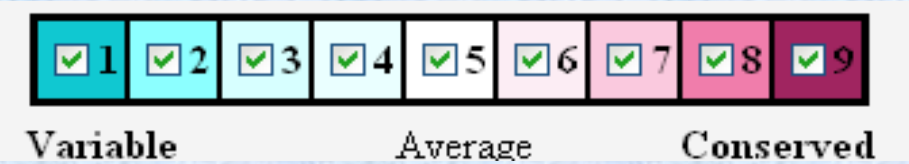
COLOR: The color scale representing the conservation scores (9 - conserved, 1 - variable).

MSA DATA: The number of aligned sequences having an amino acid (non-gapped) from the overall number of sequences at each position. –

RESIDUE VARIETY: The residues variety at each position of the multiple sequence alignment.

| POS | SEQ | 3LATOM | SCORE (normalized) | COLOR | MSA DATA | RESIDUE VARIETY |
|---|---|---|---|---|---|---|
| 1 | K | LYS1:A | -0.877 | 9 | 49/50 | K |
| 2 | V | VAL2:A | 0.436 | 3 | 49/50 | I,K,T,V |
| 3 | F | PHE3:A | 0.763 | 1 | 50/50 | F,Y |
| 4 | G | GLY4:A | 1.212 | 1 | 50/50 | D,E,G,K,Q,S,T |
| 5 | R | ARG5:A | -0.673 | 8 | 50/50 | Q,R |
| 6 | C | CYS6:A | -0.877 | 9 | 50/50 | C |
| 7 | E | GLU7:A | -0.877 | 9 | 50/50 | E |
| 8 | L | LEU8:A | 0.174 | 4 | 50/50 | A,F,L,W |
| 9 | A | ALA9:A | -0.877 | 9 | 50/50 | A |
| 10 | A | ALA10:A | -0.477 | 7 | 50/50 | A,K,R |




1
2
3
4
5
6
7
8
9
Variable
Average
Conserved

# awk programming

**awk is a pattern-matching program for processing files, especially when they are databases (when each line has a simple field oriented layout).**

**The awk utility is a data extraction and reporting tool that uses a data-driven scripting language consisting of a set of actions to be taken against textual data for the purpose of producing formatted reports.**

**It is a very powerful program for handling large amount of data (especially parsing data files in bioinformatics).**

**awk is one of the earliest tools to appear in Unix and gained popularity as a way to add computational features to a Unix pipeline.**

**awk was created at Bell Labs in the 1970s, and its name is derived from the family names of its authors — Alfred Aho, Peter Weinberger, and Brian Kernighan.**

# What can we do with awk?

- Handle a text file as made up of records and fields in a textual database.
- Perform arithmetic and string operations.
- Use programming constructs such as loops and conditionals.
- Produce formatted reports.

AWK - the original from AT&T
NAWK - A newer, improved version from AT&T
GAWK - The Free Software foundation's version

- With nawk, you can also:
- Execute Unix commands from a script.
- Process the results of Unix commands.
- Process command-line arguments more gracefully.
- Work more easily with multiple input streams.
- Perform more powerful string substitutions (gawk)

# awk one liners

**Syntax**

**awk [options] 'script' var=value file(s)**

**Pattern-action statement**

**awk '<u>pattern</u> {*action*}' file name**

**E.g. awk '<u>NF>1</u> {*print $1*}' abc.dat**

**awk [options] –f scriptfile var=value file(s)**

**-F *fs:* Use *fs* for the input field separator (the value of the FS predefined variable).**

**-f *program-file:* Read the awk program source from the file *program-file*, instead of from the first command line argument.**

**-v *var=val:* Assign the variable *var* the value *val* before program execution begins.**

**--: Signal the end of options.**

# Sample input files

## test1.dat

```
% Summary reports on aggregation-prone regions
% Three types of aggregation-prone regions:
% Type1: TANGO score >= 10%
% Type2: PAGE Zscore >= 1.96
% Type3: TANGO score >= 5% and PAGE Zscore >= 1

# 1dbuacaMS1       3 APRs
  41     45            SFK--TLLVA--ENG      2
  53     61        QKK--LACFVLATA--NLN      2
  89     93            KST--GYLVG--GIS      1

# 1prxacaMS2       5 APRs
  29     34           GDS--WGILFS--HPR      3
  63     68           NVK--LIALSI--DSV      2
 105    112         NRE--LAILLGML--DPA      4
 128    133           TAR--VVFVFG--PDK      4
 158    165           ILR--VVISLQLT--A      1

# 1m5sacaMS3       6 APRs
  23     27            IAR--VLITA--ATK      2
  32     38          TKR--WALVAAT--EAT      3
  43     48           ATG--FATSVI--MCP      1
  72     79         RPG--VYVQICTF--KYE      4
 121    127          GFK--LKFFADG--MES      2
 164    171         IAG--GNFFIFGD--SQM      0

# 1iq6acaMS4       2 APRs
  16     22          AAE--VAAFAAL--SED      3
  49     59      VHG--MLLASLFSGLL--GQQ      0

# 1spvacaMS5       3 APRs
   9     18       QGD--ITKLAVDVIV--NAA      1
  61     66           TGH--AVITLA--GDL      1
 149    155          LPE--QVYFVCY--DEE      5
```

## test2.dat

```
HEADER    SIGNALING PROTEIN                       17-JUL-00   1FC3
TITLE     THE CRYSTAL STRUCTURE OF TRANS-ACTIVATION DOMAIN OF THE
TITLE    2 SPORULATION RESPONSE REGULATOR, SPO0A
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: SPO0A;
COMPND   3 CHAIN: A, B, C;
SOURCE   2 ORGANISM_SCIENTIFIC: GEOBACILLUS STEAROTHERMOPHILUS;
SOURCE   3 ORGANISM_TAXID: 1422;
SOURCE   4 EXPRESSION_SYSTEM: ESCHERICHIA COLI BL21;
SOURCE   5 EXPRESSION_SYSTEM_TAXID: 511693;
SEQRES   1 A  120  ASN LYS PRO LYS ASN LEU ASP ALA SER ILE THR SER ILE
SEQRES   2 A  120  ILE HIS GLU ILE GLY VAL PRO ALA HIS ILE LYS GLY TYR
SEQRES   3 A  120  LEU TYR LEU ARG GLU ALA ILE ALA MET VAL TYR HIS ASP
SEQRES   4 A  120  ILE GLU LEU LEU GLY SER ILE THR LYS VAL LEU TYR PRO
SEQRES   5 A  120  ASP ILE ALA LYS LYS TYR ASN THR THR ALA SER ARG VAL
SEQRES   6 A  120  GLU ARG ALA ILE ARG HIS ALA ILE GLU VAL ALA TRP SER
SEQRES   7 A  120  ARG GLY ASN LEU GLU SER ILE SER SER LEU PHE GLY TYR
SEQRES   8 A  120  THR VAL SER VAL SER LYS ALA LYS PRO THR ASN SER GLU
SEQRES   9 A  120  PHE ILE ALA MET VAL ALA ASP LYS LEU ARG LEU GLU HIS
SEQRES  10 A  120  LYS ALA SER
HELIX    1   1 ASN A  140  GLY A  157  1                                18
HELIX    2   2 ILE A  162  ASP A  178  1                                17
HELIX    3   3 ILE A  179  ILE A  185  5                                 7
HELIX    4   4 VAL A  188  ASN A  198  1                                11
HELIX    5   5 THR A  200  ARG A  218  1                                19
HELIX    6   6 ILE A  224  GLY A  229  1                                 6
HELIX    7   7 GLY A  229  VAL A  234  1                                 6
HELIX    8   8 THR A  240  GLU A  255  1                                16
HELIX    9   9 LYS B  141  GLY B  157  1                                17
HELIX   10  10 ILE B  162  ASP B  178  1                                17
HELIX   11  11 ILE B  179  ILE B  185  5                                 7
HELIX   12  12 VAL B  188  TYR B  197  1                                10
HELIX   13  13 THR B  200  GLY B  219  1                                20
HELIX   14  14 THR B  240  LEU B  254  1                                15
ATOM      8  ND2 ASN A 140     -14.365  43.311  15.200  1.00 41.86      N
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17      N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38      C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33      C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98      O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89      C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74      C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23      C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17      C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63      N
ATOM   1497  ND1 HIS B 210      12.713  31.145  12.352  1.00 24.71      N
ATOM   1498  CD2 HIS B 210      13.114  33.174  13.046  1.00 24.85      C
ATOM   1499  CE1 HIS B 210      13.280  31.795  11.352  1.00 23.41      C
ATOM   1500  NE2 HIS B 210      13.532  33.027  11.746  1.00 23.59      N
ATOM   1501  N   ALA B 211      14.746  32.053  16.235  1.00 26.10      N
ATOM   1502  CA  ALA B 211      16.084  32.634  16.246  1.00 26.32      C
ATOM   1503  C   ALA B 211      17.083  31.563  16.650  1.00 26.17      C
ATOM   1504  O   ALA B 211      18.156  31.442  16.068  1.00 26.46      O
```

# Sample input files

## test3.dat

```
==== Secondary Structure Definition by the program DSSP, CMBI version by M.L. Hekkelman/2010-10-21 ==== DATE=2011-08-08        .
REFERENCE W. KABSCH AND C.SANDER, BIOPOLYMERS 22 (1983) 2577-2637                                                          .
   0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-2), SAME NUMBER PER 100 RESIDUES                           .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-1), SAME NUMBER PER 100 RESIDUES                           .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+0), SAME NUMBER PER 100 RESIDUES                           .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+1), SAME NUMBER PER 100 RESIDUES                           .
   13 13.3   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+2), SAME NUMBER PER 100 RESIDUES                           .
    9  9.2   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+3), SAME NUMBER PER 100 RESIDUES                           .
    2  2.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+4), SAME NUMBER PER 100 RESIDUES                           .
    1  1.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+5), SAME NUMBER PER 100 RESIDUES                           .
  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30     *** HISTOGRAMS OF ***       .
  0  3  0  0  1  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    ANTIPARALLEL BRIDGES PER LADDER .
  1  0  2  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    LADDERS PER SHEET               .
  #  RESIDUE AA STRUCTURE BP1 BP2  ACC     N-H-->O    O-->H-N    N-H-->O    O-->H-N    TCO  KAPPA ALPHA  PHI   PSI    X-CA   Y-CA   Z-CA
    1    0 A A              0   0   87     0, 0.0     2,-0.4     0, 0.0    27,-0.2   0.000 360.0 360.0 360.0 114.9   19.4   18.8   44.7
    2    1 A A  E     -a   28   0A  17    25,-2.4    27,-3.5    18,-0.0     2,-0.4  -0.751 360.0-150.5 -94.6 137.3   15.8   20.2   44.7
    3    2 A I  E     -a   29   0A  91    -2,-0.4     2,-0.4    25,-0.2    27,-0.2  -0.908  13.2-179.4-111.1 130.4   13.2   19.0   47.2
    4    3 A V  E     -a   30   0A   0    25,-2.7    27,-2.6    -2,-0.4     2,-0.3  -0.998  17.5-145.0-130.0 125.1    9.5   19.1   46.4
    5    4 A K  E     -aB  31  16A  73    11,-2.8    11,-2.4    -2,-0.4     2,-0.8  -0.713   9.7-136.8 -90.9 142.9    6.9   17.9   49.0
    6    5 A L  E    S+aB  32  15A   0    25,-2.6    27,-2.2    -2,-0.3    28,-2.2  -0.887  78.8   7.0-101.0 102.9    3.8   16.1   47.9
    7    6 A G  S    S-     0   0    0     7,-2.8     2,-0.2    -2,-0.8     6,-0.2   0.167  89.3 -96.3  97.5 142.4    0.9   17.5   49.9
    8    7 A G    >   -     0   0    5     4,-2.4     3,-2.4    25,-0.1    -2,-0.0  -0.571  29.1-113.0 -88.3 155.2    1.1   20.5   52.2
    9    8 A D  T 3  S+     0   0  136     1,-0.3    -1,-0.1    -2,-0.2     0, 0.0   0.732 123.4  54.9 -56.8 -28.6    1.6   20.2   56.0
   10    9 A D  T 3  S-     0   0  142    24,-0.1    -1,-0.3     2,-0.0    -3,-0.0   0.396 127.7-103.9 -87.2  -0.5   -2.0   21.6   56.2
   11   10 A G    <   +     0   0   38    -3,-2.4    -2,-0.2     1,-0.3     2,-0.1   0.495  65.9 158.0  88.4  11.5   -3.1   18.8   53.9
   12   11 A S        -     0   0   48     1,-0.1    -4,-2.4    -5,-0.0     2,-1.8  -0.414  50.3-121.1 -66.1 143.7   -3.5   20.9   50.8
   13   12 A L  S    S+     0   0   60    -6,-0.2     2,-0.3    76,-0.1    -1,-0.1  -0.588  72.4 119.3 -87.0  72.1   -3.3   18.8   47.5
   14   13 A A  S    S-     0   0   29    -2,-1.8    -7,-2.8    77,-0.1     2,-0.4  -0.965  70.1-117.0-139.6 156.3   -0.3   20.7   46.2
   15   14 A F  E     -B    6   0A   4    77,-0.4    -9,-0.2    -2,-0.3    77,-0.1  -0.777  45.1-120.7 -89.1 129.8    3.3   20.4   45.2
   16   15 A V  E    S+B    5   0A  63   -11,-2.4   -11,-2.8    -2,-0.4     2,-0.1  -0.943 101.5  16.0-123.7 106.5    5.4   22.5   47.6
   17   16 A P  S    S-     0   0   60     0, 0.0    -1,-0.3     0, 0.0     3,-0.1   0.677  79.3-165.0 -81.7 170.6    6.8   24.5   46.0
   18   17 A N  S    S+     0   0   97    -2,-0.1    76,-1.9     1,-0.1     2,-0.3   0.238  77.9  50.9-103.7   8.9    4.9   24.4   42.7
   19   18 A N  E     +d   94   0B  81    74,-0.2     2,-0.3     2,-0.0    76,-0.2  -0.967  67.6 161.4-144.4 126.4    7.5   26.2   40.8
   20   19 A I  E     -d   95   0B  19    74,-1.6    76,-2.7    -2,-0.3     2,-0.4  -0.982  26.0-148.5-149.4 151.7   11.2   25.3   40.9
```

# Sample input files

## test4.dat

```
1prxacaMS2  79    WSK   DINAYN      SEE   32.1852     6
1m5sacaMS3  20    IKI   ARVLIT      AAT   11.4172     6
1m5sacaMS3  164   AGG   NFFIFG      DSQ   96.2088     6
1m5sacaMS3  242   DVN   AVYEIV      ING   19.8543     6
1iq6acaMS4  50    GML   LASLFS      GLL   10.7893     6
1spvacaMS5  147   ALP   EQVYFV      CYD   96.1525     6
1x7dacaMS7  62    KSR   YAFKYV      NGH   58.0513     6
1x7dacaMS7  149   GIE   EIVAYD      TDP   47.4865     6
1x7dacaMS7  236   NAR   VFVEYE      PQT   74.9182     6
1lzlacaMS8  130   DCY   AALLYI      HAH   48.2344     6
1lzlacaMS8  228   EDP   DVSIYA      APS   25.9341     6
1ul1acaMS10 35    MSI   YQFLIA      VRQ   95.5041     6
1ul1ccaMS11 61    HKE   AHQLFL      EPE   10.7389     6
1sxjacaMS12 22    DCV   QLVNFQ      CKE   81.4939     6
1sxjccaMS13 26    LSD   SINIIT      KET   42.2134     6
1sxjccaMS13 111   KSG   FLQFFL      APK   14.4927     6
1rypacaMS14 15    GRN   FQVEYA      VKA   87.3043     6
1yx3acaMS15 45    EHW   DIINFL      REY   10.2064     6
1ljlacaMS16 0     X     TIYFIC      TGN   15.1024     6
1p1jacaMS17 61    IAS   NDILYN      DKL   93.6116     6
1p1jacaMS17 92    KVA   MDEYYS      ELM   44.0619     6
1m3sacaMS18 76    EEG   DLVIIG      SGS   54.9764     6
```

## test6.dat

```
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17          N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38          C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33          C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98          O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89          C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74          C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23          C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17          C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63          N
ATOM   1497  ND1 HIS B 210      12.713  31.145  12.352  1.00 24.71          N
ATOM   1498  CD2 HIS B 210      13.114  33.174  13.046  1.00 24.85          C
ATOM   1499  CE1 HIS B 210      13.280  31.795  11.352  1.00 23.41          C
ATOM   1500  NE2 HIS B 210      13.532  33.027  11.746  1.00 23.59          N
ATOM   1501  N   ALA B 211      14.746  32.053  16.235  1.00 26.10          N
ATOM   1502  CA  ALA B 211      16.084  32.634  16.246  1.00 26.32          C
ATOM   1503  C   ALA B 211      17.083  31.563  16.650  1.00 26.17          C
ATOM   1504  O   ALA B 211      18.156  31.442  16.068  1.00 26.46          O
```

## test5.dat

```
  16  11   1   8   5  11   2  10  19  16   0   6   6   5   3   7   9  11   0   3
  17  12   0   9   6  12   0  11  20  17   0   7   7   6   4   8  10  12   1   4
 10.74  7.38   .67  5.37  3.36  7.38  1.34  6.71 12.75 10.74   .00  4.03  4.03  3.36  2.01  4.70  6.04  7.38   .00  2.01
 11.49  8.11   .00  6.08  4.05  8.11   .00  7.43 13.51 11.49   .00  4.73  4.73  4.05  2.70  5.41  6.76  8.11   .68  2.70
  14  13   0  11  10   9   3  11  10  21   2   7  11   1  10   8  10  11   2   2
  15  14   1  12  11  10   4  12  11  22   0   0  12   2  11   9  11  12   3   3
  8.43  7.83   .00  6.63  6.02  5.42  1.81  6.63  6.02 12.65  1.20  4.22  6.63   .60  6.02  4.82  6.02  6.63  1.20  1.20
  9.55  8.92   .64  7.64  7.01  6.37  2.55  7.64  7.01 14.01   .00   .00  7.64  1.27  7.01  5.73  7.01  7.64  1.91  1.91
  41  10   4  29  16  28   1  26  22  17   6  11  13   7   7  11  20  19   1   8
  42  11   5  30  17  29   2  27  23  18   7  12  14   8   8  12  21  20   2   9
 13.80  3.37  1.35  9.76  5.39  9.43   .34  8.75  7.41  5.72  2.02  3.70  4.38  2.36  2.36  3.70  6.73  6.40   .34  2.69
 14.14  3.70  1.68 10.10  5.72  9.76   .67  9.09  7.74  6.06  2.36  4.04  4.71  2.69  2.69  4.04  7.07  6.73   .67  3.03
  18   4   0   8   9  12   2   4   6  17   1   1   6   5   5   7   9  11   0   1
  19   5   0   9  10  13   3   5   7  18   0   2   7   6   6   8  10  12   1   2
 14.29  3.17   .00  6.35  7.14  9.52  1.59  3.17  4.76 13.49   .79   .79  4.76  3.97  3.97  5.56  7.14  8.73   .00   .79
 14.18  3.73   .00  6.72  7.46  9.70  2.24  3.73  5.22 13.43   .00  1.49  5.22  4.48  4.48  5.97  7.46  8.96   .75  1.49
```

# Sample input files

## test8.dat

```
139 1a1sa.dssp1
139 1a1sb.dssp1
219 1a2za.dssp1
219 1a2zb.dssp1
350 1a59a.dssp1
350 1a59b.dssp1
385 1a5aa.dssp1
385 1a5ab.dssp1
100 1a6fa.dssp1
100 1a6fb.dssp1
103 1a81a.dssp1
103 1a81b.dssp1
 93 1adja.dssp1
 93 1adjb.dssp1
153 1ahja.dssp1
153 1ahjb.dssp1
100 1aipa.dssp1
100 1aipb.dssp1
 93 1aipc.dssp1
 93 1aipd.dssp1
341 1aj8a.dssp1
341 1aj8b.dssp1
164 1amua.dssp1
164 1amub.dssp1
 80 1amuc.dssp1
 80 1amud.dssp1
 83 1amue.dssp1
 83 1amuf.dssp1
```

## test9.dat

```
     9  N        -9.552    1.00 39.17
    10  CA        -8.497    1.00 38.38
    11  C        -8.194    1.00 37.33
   12  O          -7.793    1.00 36.98
13  CB         -7.165    1.00 38.89
  14  CG        -6.899    1.00 39.74
  15  CD        -6.400    1.00 41.23
```

## test10.dat

```
9 N -9.552  39.17
10 CA -8.497  38.38
11 C -8.194  37.33
12 O -7.793  36.98
13 CB -7.165  38.89
14 CG -6.899  39.74
15 CD -6.400  41.23
```

## 1. Print column 6

```
awk '{print $6}' test4.dat
```

```
32.1852
11.4172
96.2088
19.8543
10.7893
96.1525
58.0513
47.4865
74.9182
48.2344
25.9341
95.5041
10.7389
81.4939
42.2134
```

```
1prxacaMS2   79   WSK   DINAYN   SEE   32.1852   6
1m5sacaMS3   20   IKI   ARVLIT   AAT   11.4172   6
1m5sacaMS3   164  AGG   NFFIFG   DSQ   96.2088   6
1m5sacaMS3   242  DVN   AVYEIV   ING   19.8543   6
1iq6acaMS4   50   GML   LASLFS   GLL   10.7893   6
1spvacaMS5   147  ALP   EQVYFV   CYD   96.1525   6
1x7dacaMS7   62   KSR   YAFKYV   NGH   58.0513   6
1x7dacaMS7   149  GIE   EIVAYD   TDP   47.4865   6
1x7dacaMS7   236  NAR   VFVEYE   PQT   74.9182   6
1lzlacaMS8   130  DCY   AALLYI   HAH   48.2344   6
1lzlacaMS8   228  EDP   DVSIYA   APS   25.9341   6
1ul1acaMS10  35   MSI   YQFLIA   VRQ   95.5041   6
1ul1ccaMS11  61   HKE   AHQLFL   EPE   10.7389   6
1sxjacaMS12  22   DCV   QLVNFQ   CKE   81.4939   6
1sxjccaMS13  26   LSD   SINIIT   KET   42.2134   6
1sxjccaMS13  111  KSG   FLQFFL   APK   14.4927   6
1rypacaMS14  15   GRN   FQVEYA   VKA   87.3043   6
1yx3acaMS15  45   EHW   DIINFL   REY   10.2064   6
1ljlacaMS16  0    X     TIYFIC   TGN   15.1024   6
1p1jacaMS17  61   IAS   NDILYN   DKL   93.6116   6
1p1jacaMS17  92   KVA   MDEYYS   ELM   44.0619   6
1m3sacaMS18  76   EEG   DLVIIG   SGS   54.9764   6
```

## 2. Print columns 1 and 6

```
awk '{print $1,$6}' test4.dat
```

```
1prxacaMS2 32.1852
1m5sacaMS3 11.4172
1m5sacaMS3 96.2088
1m5sacaMS3 19.8543
1iq6acaMS4 10.7893
1spvacaMS5 96.1525
1x7dacaMS7 58.0513
1x7dacaMS7 47.4865
1x7dacaMS7 74.9182
1lzlacaMS8 48.2344
1lzlacaMS8 25.9341
1ul1acaMS10 95.5041
1ul1ccaMS11 10.7389
1sxjacaMS12 81.4939
1sxjccaMS13 42.2134
```

## 3. Print in reverse order (columns 6 and 1)

```
awk '{print $6, $1}' test4.dat
```

## 4. Write the results in a file

```
awk '{print $1" "$6}' test4.dat > test1.result
```

## 5. Delete empty lines

NF: number of fields

```
awk ' NF>1 {print}' test1.dat
```

```
% Summary reports on aggregation-prone regions
% Three types of aggregation-prone regions:
% Type1: TANGO score >= 10%
% Type2: PAGE Zscore >= 1.96
% Type3: TANGO score >= 5% and PAGE Zscore >= 1
# 1dbuacaMS1      3 APRs
  41    45          SFK--TLLVA--ENG     2
  53    61       QKK--LACFVLATA--NLN    2
  89    93          KST--GYLVG--GIS     1
# 1prxacaMS2      5 APRs
  29    34         GDS--WGILFS--HPR     3
  63    68         NVK--LIALSI--DSV     2
 105   112       NRE--LAILLGML--DPA     4
 128   133         TAR--VVFVFG--PDK     4
 158   165         ILR--VVISLQLT--A     1
# 1m5sacaMS3      6 APRs
  23    27          IAR--VLITA--ATK     2
  32    38        TKR--WALVAAT--EAT     3
  43    48         ATG--FATSVI--MCP     1
  72    79       RPG--VYVQICTF--KYE     4
 121   127        GFK--LKFFADG--MES     2
 164   171       IAG--GNFFIFGD--SQM     0
```

## 6. Number each line

**FNR: file line number**

**awk '{print FNR $0}' test4.dat**

```
11prxacaMS2 79    WSK   DINAYN      SEE   32.1852     6
21m5sacaMS3 20    IKI   ARVLIT      AAT   11.4172     6
31m5sacaMS3 164   AGG   NFFIFG      DSQ   96.2088     6
41m5sacaMS3 242   DVN   AVYEIV      ING   19.8543     6
51iq6acaMS4 50    GML   LASLFS      GLL   10.7893     6
61spvacaMS5 147   ALP   EQVYFV      CYD   96.1525     6
71x7dacaMS7 62    KSR   YAFKYV      NGH   58.0513     6
81x7dacaMS7 149   GIE   EIVAYD      TDP   47.4865     6
91x7dacaMS7 236   NAR   VFVEYE      PQT   74.9182     6
1011zlacaMS8      130   DCY   AALLYI      HAH   48.2344
```

## 7. Number each line with tab

**awk '{print FNR "\t" $0}' test4.dat**

**$0 denotes all fields**

```
1     1prxacaMS2  79    WSK   DINAYN      SEE   32.1852     6
2     1m5sacaMS3  20    IKI   ARVLIT      AAT   11.4172     6
3     1m5sacaMS3  164   AGG   NFFIFG      DSQ   96.2088     6
4     1m5sacaMS3  242   DVN   AVYEIV      ING   19.8543     6
5     1iq6acaMS4  50    GML   LASLFS      GLL   10.7893     6
6     1spvacaMS5  147   ALP   EQVYFV      CYD   96.1525     6
7     1x7dacaMS7  62    KSR   YAFKYV      NGH   58.0513     6
8     1x7dacaMS7  149   GIE   EIVAYD      TDP   47.4865     6
9     1x7dacaMS7  236   NAR   VFVEYE      PQT   74.9182     6
10    11zlacaMS8  130   DCY   AALLYI      HAH   48.2344     6
```

## 8. Count lines (similar to wc -l)

**awk 'END {print NR}' test4.dat**

**NR: line number**

**22**

```
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17           N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38           C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33           C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98           O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89           C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74           C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23           C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17           C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63           N
ATOM   1497  ND1 HIS B 210      12.713  31.145  12.352  1.00 24.71           N
ATOM   1498  CD2 HIS B 210      13.114  33.174  13.046  1.00 24.85           C
ATOM   1499  CE1 HIS B 210      13.280  31.795  11.352  1.00 23.41           C
ATOM   1500  NE2 HIS B 210      13.532  33.027  11.746  1.00 23.59           N
ATOM   1501  N   ALA B 211      14.746  32.053  16.235  1.00 26.10           N
ATOM   1502  CA  ALA B 211      16.084  32.634  16.246  1.00 26.32           C
ATOM   1503  C   ALA B 211      17.083  31.563  16.650  1.00 26.17           C
ATOM   1504  O   ALA B 211      18.156  31.442  16.068  1.00 26.46           O
```

```
N
C
C
O
C
C
C
C
N
N
C
C
N
N
C
C
O
```

## 9. Print the last field of each line

**awk '{print $NF}' test6.dat**

## 10. Print the last field of last line

**awk ' END {print $NF}' test6.dat**

**O**

## 11. Print every line, where the value of the 6th field is more than 50

**awk '$6 > 50 {print}' test4.dat**

```
1m5sacaMS3  164   AGG   NFFIFG     DSQ   96.2088    6
1spvacaMS5  147   ALP   EQVYFV     CYD   96.1525    6
1x7dacaMS7  62    KSR   YAFKYV     NGH   58.0513    6
1x7dacaMS7  236   NAR   VFVEYE     PQT   74.9182    6
1ul1acaMS10 35    MSI   YQFLIA     VRQ   95.5041    6
1sxjacaMS12 22    DCV   QLVNFQ     CKE   81.4939    6
1rypacaMS14 15    GRN   FQVEYA     VKA   87.3043    6
1p1jacaMS17 61    IAS   NDILYN     DKL   93.6116    6
1m3sacaMS18 76    EEG   DLVIIG     SGS   54.9764    6
```

## 12. print the lines starting from 15

**awk ' NR > 14 {print}' test3.dat**

```
 3    2 A I  E      -a   29   0A  91     -2,-0.4     2,-0.4    25,-0.2    27,-0.2  -0.908  13.2-179.4-111.1 130.4   13.2   19.0   47.2
 4    3 A V  E      -a   30   0A   0     25,-2.7    27,-2.6    -2,-0.4     2,-0.3  -0.998  17.5-145.0-130.0 125.1    9.5   19.1   46.4
 5    4 A K  E      -aB  31  16A  73     11,-2.8    11,-2.4    -2,-0.4     2,-0.8  -0.713   9.7-136.8 -90.9 142.9    6.9   17.9   49.0
 6    5 A L  E     S+aB  32  15A   0     25,-2.6    27,-2.2    -2,-0.3    28,-2.2  -0.887  78.8   7.0-101.0 102.9    3.8   16.1   47.9
 7    6 A G  S     S-     0   0   0      7,-2.8     2,-0.2    -2,-0.8     6,-0.2   0.167  89.3 -96.3  97.5 142.4    0.9   17.5   49.9
 8    7 A G    >   -      0   0   5      4,-2.4     3,-2.4    25,-0.1    -2,-0.0  -0.571  29.1-113.0 -88.3 155.2    1.1   20.5   52.2
 9    8 A D  T 3  S+      0   0 136      1,-0.3    -1,-0.1    -2,-0.2     0, 0.0   0.732 123.4  54.9 -56.8 -28.6    1.6   20.2   56.0
10    9 A D  T 3  S-      0   0 142     24,-0.1    -1,-0.3     2,-0.0    -3,-0.0   0.396 127.7-103.9 -87.2  -0.5   -2.0   21.6   56.2
11   10 A G    <   +      0   0  38     -3,-2.4    -2,-0.2     1,-0.3     2,-0.1   0.495  65.9 158.0  88.4  11.5   -3.1   18.8   53.9
12   11 A S        -      0   0  48      1,-0.1    -4,-2.4    -5,-0.0     2,-1.8  -0.414  50.3-121.1 -66.1 143.7   -3.5   20.9   50.8
```

## 13. Separate with field separator

**awk -F "--" '{print $2}' test1.dat**

```
TLLVA
LACFVLATA
GYLVG


WGILFS
LIALSI
LAILLGML
VVFVFG
VVISLQLT


VLITA
WALVAAT
FATSVI
VYVQICTF
LKFFADG
GNFFIFGD


VAAFAAL
MLLASLFSGLL
```

## 14. Separate with field separator

**awk -F "--" '{print $1" "$2" "$3}' test1.dat**

```
# 1iq6acaMS4       2 APRs
  16    22           AAE VAAFAAL SED 3
  49    59       VHG MLLASLFSGLL GQQ 0

# 1spvacaMS5       3 APRs
   9    18        QGD ITKLAVDVIV NAA 1
  61    66            TGH AVITLA GDL 1
 149   155           LPE QVYFVCY DEE 5
```

## 15. Print alternate lines

awk 'NR%2' test8.dat

```
139 1a1sa.dssp1
219 1a2za.dssp1
350 1a59a.dssp1
385 1a5aa.dssp1
100 1a6fa.dssp1
103 1a8la.dssp1
 93 1adja.dssp1
153 1ahja.dssp1
```

awk 'NR%3' test8.dat

```
139 1a1sa.dssp1
139 1a1sb.dssp1
219 1a2zb.dssp1
350 1a59a.dssp1
385 1a5aa.dssp1
385 1a5ab.dssp1
100 1a6fb.dssp1
103 1a8la.dssp1
 93 1adja.dssp1
 93 1adjb.dssp1
153 1ahjb.dssp1
```

```
139 1a1sa.dssp1
139 1a1sb.dssp1
219 1a2za.dssp1
219 1a2zb.dssp1
350 1a59a.dssp1
350 1a59b.dssp1
385 1a5aa.dssp1
385 1a5ab.dssp1
100 1a6fa.dssp1
100 1a6fb.dssp1
103 1a8la.dssp1
103 1a8lb.dssp1
 93 1adja.dssp1
 93 1adjb.dssp1
153 1ahja.dssp1
153 1ahjb.dssp1
100 1aipa.dssp1
100 1aipb.dssp1
 93 1aipc.dssp1
 93 1aipd.dssp1
341 1aj8a.dssp1
341 1aj8b.dssp1
164 1amua.dssp1
164 1amub.dssp1
 80 1amuc.dssp1
 80 1amud.dssp1
 83 1amue.dssp1
 83 1amuf.dssp1
```

## 16. Substitute LYS by ARG

**awk '{sub(/LYS/,"ARG"); print}' test6.dat**

```
ATOM      9  N    ARG A 141      -9.552  43.465  13.292  1.00 39.17          N
ATOM     10  CA   ARG A 141      -8.497  42.864  12.475  1.00 38.38          C
ATOM     11  C    ARG A 141      -8.194  41.362  12.471  1.00 37.33          C
ATOM     12  O    ARG A 141      -7.793  40.846  11.439  1.00 36.98          O
ATOM     13  CB   ARG A 141      -7.165  43.539  12.800  1.00 38.89          C
ATOM     14  CG   ARG A 141      -6.899  44.766  11.997  1.00 39.74          C
ATOM     15  CD   ARG A 141      -6.400  45.848  12.886  1.00 41.23          C
ATOM     16  CE   ARG A 141      -5.650  46.873  12.073  1.00 42.17          C
ATOM     17  NZ   ARG A 141      -5.910  48.238  12.590  1.00 42.63          N
ATOM   1497  ND1  HIS B 210      12.713  31.145  12.352  1.00 24.71          N
ATOM   1498  CD2  HIS B 210      13.114  33.174  13.046  1.00 24.85          C
ATOM   1499  CE1  HIS B 210      13.280  31.795  11.352  1.00 23.41          C
ATOM   1500  NE2  HIS B 210      13.532  33.027  11.746  1.00 23.59          N
ATOM   1501  N    ALA B 211      14.746  32.053  16.235  1.00 26.10          N
ATOM   1502  CA   ALA B 211      16.084  32.634  16.246  1.00 26.32          C
ATOM   1503  C    ALA B 211      17.083  31.563  16.650  1.00 26.17          C
ATOM   1504  O    ALA B 211      18.156  31.442  16.068  1.00 26.46          O
```

## 11.4 Built-in Variables

| *Version* | *Variable* | *Description* |
|---|---|---|
| awk | FILENAME | Current filename |
| | FS | Field separator (a space) |
| | NF | Number of fields in current record |
| | NR | Number of the current record |
| | OFMT | Output format for numbers ("%.6g") and for conversion to string |
| | OFS | Output field separator (a space) |
| | ORS | Output record separator (a newline) |
| | RS | Record separator (a newline) |
| | $0 | Entire input record |
| | $*n* | *n*th field in current record; fields are separated by FS |
| nawk | ARGC | Number of arguments on command line |
| | ARGV | An array containing the command-line arguments, indexed from 0 to ARGC - 1 |
| | CONVFMT | String conversion format for numbers ("%.6g") (POSIX) |
| | ENVIRON | An associative array of environment variables |
| | FNR | Like NR, but relative to the current file |
| | RLENGTH | Length of the string matched by match() function |
| | RSTART | First position in the string matched by match() function |
| | SUBSEP | Separator character for array subscripts ("\034") |

Unix in a nutshell, Stever et al. O'REILLY

## 17. Delete the 4th field on each line

```
awk '{$4=""; print}' test9.dat > test10.dat
```

```
9 N -9.552  39.17
10 CA -8.497  38.38
11 C -8.194  37.33
12 O -7.793  36.98
13 CB -7.165  38.89
14 CG -6.899  39.74
15 CD -6.400  41.23
```

```
 9  N     -9.552    1.00 39.17
10  CA     -8.497    1.00 38.38
11  C     -8.194    1.00 37.33
12  O     -7.793    1.00 36.98
13  CB     -7.165    1.00 38.89
14  CG     -6.899    1.00 39.74
15  CD     -6.400    1.00 41.23
```

## 18. Place the fields in order

"printf" option

```
awk '{printf ("%3d %3s %9.2f %9.2f\n", $1,$2,$3,$4)}' test10.dat
```

```
 9   N     -9.55     39.17
10  CA     -8.50     38.38
11   C     -8.19     37.33
12   O     -7.79     36.98
13  CB     -7.17     38.89
14  CG     -6.90     39.74
15  CD     -6.40     41.23
```

s: A format (STRING)
d: I format (INTEGER)
f: F format (DECIMAL)

## 19. Delete 4th field and put place all other fields in order

```
awk '{$4=""; printf ("%3d %3s %9.2f %9.2f\n", $1,$2,$3,$5)}' test9.dat
```

# Print formats

| *Character* | *Description* |
|---|---|
| c | ASCII character |
| d | Decimal integer |
| i | Decimal integer (added in POSIX) |
| e | Floating-point format ([-]*d.precision*e[+-]*dd*) |
| E | Floating-point format ([-]*d.precision*E[+-]*dd*) |
| f | Floating-point format ([-]*ddd.precision*) |
| g | e or f conversion, whichever is shortest, with trailing zeros removed |
| G | E or f conversion, whichever is shortest, with trailing zeros removed |
| o | Unsigned octal value |
| s | String |
| x | Unsigned hexadecimal number; uses a-f for 10 to 15 |
| X | Unsigned hexadecimal number; uses A-F for 10 to 15 |
| % | Literal % |

| *Conversion* | *Precision Means* |
|---|---|
| %d, %i, %o | The minimum number of digits to print |
| %u, %x, %X | |
| %e, %E, %f | The number of digits to the right of the decimal point |
| %g, %G | The maximum number of significant digits |
| %s | The maximum number of characters to print |

## 20. Print the first 3 lines

**awk 'NR<4 {print}' test2.dat**

```
HEADER    SIGNALING PROTEIN                                17-JUL-00   1FC3
TITLE     THE CRYSTAL STRUCTURE OF TRANS-ACTIVATION DOMAIN OF THE
TITLE    2 SPORULATION RESPONSE REGULATOR, SPOOA
```

## 21. Matching strings

**Print the lines contains "ATOM"**

**awk '/ATOM/ {print}' test2.dat**

```
ATOM      8  ND2 ASN A 140     -14.365  43.311  15.200  1.00 41.86          N
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17          N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38          C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33          C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98          O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89          C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74          C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23          C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17          C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63          N
ATOM   1497  ND1 HIS B 210      12.713  31.145  12.352  1.00 24.71          N
ATOM   1498  CD2 HIS B 210      13.114  33.174  13.046  1.00 24.85          C
```

# 22. Find the amino acid sequence in a PDB file

**awk '/SEQRES/ {print}' test2.dat**

```
SEQRES   1 A  120  ASN LYS PRO LYS ASN LEU ASP ALA SER ILE THR SER ILE
SEQRES   2 A  120  ILE HIS GLU ILE GLY VAL PRO ALA HIS ILE LYS GLY TYR
SEQRES   3 A  120  LEU TYR LEU ARG GLU ALA ILE ALA MET VAL TYR HIS ASP
SEQRES   4 A  120  ILE GLU LEU LEU GLY SER ILE THR LYS VAL LEU TYR PRO
SEQRES   5 A  120  ASP ILE ALA LYS LYS TYR ASN THR THR ALA SER ARG VAL
SEQRES   6 A  120  GLU ARG ALA ILE ARG HIS ALA ILE GLU VAL ALA TRP SER
SEQRES   7 A  120  ARG GLY ASN LEU GLU SER ILE SER SER LEU PHE GLY TYR
SEQRES   8 A  120  THR VAL SER VAL SER LYS ALA LYS PRO THR ASN SER GLU
SEQRES   9 A  120  PHE ILE ALA MET VAL ALA ASP LYS LEU ARG LEU GLU HIS
SEQRES  10 A  120  LYS ALA SER
```

# 23. Get the atoms of A chain

**awk '$5~/A/ {print}' test2.dat**

```
TITLE    2 SPORULATION RESPONSE REGULATOR, SPOOA
SOURCE   2 ORGANISM_SCIENTIFIC: GEOBACILLUS STEAROTHERMOPHILUS;
SEQRES   1 A  120  ASN LYS PRO LYS ASN LEU ASP ALA SER ILE THR SER ILE
SEQRES   5 A  120  ASP ILE ALA LYS LYS TYR ASN THR THR ALA SER ARG VAL
SEQRES   7 A  120  ARG GLY ASN LEU GLU SER ILE SER SER LEU PHE GLY TYR
HELIX    1   1 ASN A  140  GLY A  157  1                                  18
HELIX    2   2 ILE A  162  ASP A  178  1                                  17
HELIX    3   3 ILE A  179  ILE A  185  5                                   7
HELIX    4   4 VAL A  188  ASN A  198  1                                  11
HELIX    5   5 THR A  200  ARG A  218  1                                  19
HELIX    6   6 ILE A  224  GLY A  229  1                                   6
HELIX    7   7 GLY A  229  VAL A  234  1                                   6
HELIX    8   8 THR A  240  GLU A  255  1                                  16
ATOM      8  ND2 ASN A 140     -14.365  43.311  15.200  1.00 41.86          N
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17          N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38          C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33          C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98          O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89          C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74          C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23          C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17          C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63          N
```

**Is this the desired one?**

## 24. Use conditions

**awk '/ATOM/ && $5~/A/ {print}' test2.dat**

```
ATOM      8  ND2 ASN A 140     -14.365  43.311  15.200  1.00 41.86           N
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17           N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38           C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33           C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98           O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89           C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74           C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23           C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17           C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63           N
```

## 25. Strict condition for most probable result

**awk '$1~/ATOM/ && $5~/A/ {print}' test2.dat**

# 26. Fourth field starting with A

**awk '$4~/^A/ {print}' test2.dat**

```
COMPND    3 CHAIN: A, B, C;
HELIX     1   1 ASN A  140  GLY A  157  1                                  18
ATOM      8  ND2 ASN A 140     -14.365  43.311  15.200  1.00 41.86          N
ATOM   1501  N   ALA B 211      14.746  32.053  16.235  1.00 26.10          N
ATOM   1502  CA  ALA B 211      16.084  32.634  16.246  1.00 26.32          C
ATOM   1503  C   ALA B 211      17.083  31.563  16.650  1.00 26.17          C
ATOM   1504  O   ALA B 211      18.156  31.442  16.068  1.00 26.46          O
```

# 27. 4th field ending with "S"

**awk '$4~/S$/ {print}' test2.dat**

```
SOURCE    2 ORGANISM_SCIENTIFIC: GEOBACILLUS STEAROTHERMOPHILUS;
HELIX     9   9 LYS B  141  GLY B  157  1                                  17
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17          N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38          C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33          C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98          O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89          C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74          C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23          C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17          C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63          N
ATOM   1497  ND1 HIS B 210      12.713  31.145  12.352  1.00 24.71          N
ATOM   1498  CD2 HIS B 210      13.114  33.174  13.046  1.00 24.85          C
ATOM   1499  CE1 HIS B 210      13.280  31.795  11.352  1.00 23.41          C
ATOM   1500  NE2 HIS B 210      13.532  33.027  11.746  1.00 23.59          N
```

## 28. Atoms with no Lys residue

awk '$1~/ATOM/ && $4!~/LYS/ {print}' test2.dat

```
ATOM      8  ND2 ASN A 140     -14.365  43.311  15.200  1.00 41.86           N
ATOM   1497  ND1 HIS B 210      12.713  31.145  12.352  1.00 24.71           N
ATOM   1498  CD2 HIS B 210      13.114  33.174  13.046  1.00 24.85           C
ATOM   1499  CE1 HIS B 210      13.280  31.795  11.352  1.00 23.41           C
ATOM   1500  NE2 HIS B 210      13.532  33.027  11.746  1.00 23.59           N
ATOM   1501  N   ALA B 211      14.746  32.053  16.235  1.00 26.10           N
ATOM   1502  CA  ALA B 211      16.084  32.634  16.246  1.00 26.32           C
ATOM   1503  C   ALA B 211      17.083  31.563  16.650  1.00 26.17           C
ATOM   1504  O   ALA B 211      18.156  31.442  16.068  1.00 26.46           O
```

## 29. Get the CA coordinates

**awk '$1~/ATOM/ && $3~/CA/ {print}' test2.dat**

```
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38           C
ATOM   1502  CA  ALA B 211      16.084  32.634  16.246  1.00 26.32           C
```

## 30. Records with LYS or ALA

**awk '$1~/ATOM/ && ($4~/LYS/||$4~/ALA/) {print}' test2.dat**

**awk '$1~/ATOM/ && ($4~/LYS/||/ALA/) {print}' test2.dat**

```
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17           N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38           C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33           C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98           O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89           C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74           C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23           C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17           C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63           N
ATOM   1501  N   ALA B 211      14.746  32.053  16.235  1.00 26.10           N
ATOM   1502  CA  ALA B 211      16.084  32.634  16.246  1.00 26.32           C
ATOM   1503  C   ALA B 211      17.083  31.563  16.650  1.00 26.17           C
ATOM   1504  O   ALA B 211      18.156  31.442  16.068  1.00 26.46           O
```

## 31. Print the lines that are more than 50 characters

**awk 'length > 50' test9.dat**

```
12  O        -7.793    1.00 36.98
```

# Operators

| *Symbol* | *Meaning* |
|---|---|
| `= += -= *= /= %= ^= **=` | Assignment |
| `?:` | C conditional expression (`nawk` only) |
| `\|\|` | Logical OR (short-circuit) |
| `&&` | Logical AND (short-circuit) |
| `in` | Array membership (`nawk` only) |
| `~ !~` | Match regular expression and negation |
| `< <= > >= != ==` | Relational operators |
| (blank) | Concatenation |
| `+ -` | Addition, subtraction |
| `* / %` | Multiplication, division, and modulus (remainder) |
| `+ - !` | Unary plus and minus, and logical negation |
| `^ **` | Exponentiation |
| `++ --` | Increment and decrement, either prefix or postfix |
| `$` | Field reference |

Unix in a nutshell, Stever et al. O'REILLY

## 32. Find the atoms with residue number 141

**awk '$1~/ATOM/ && $6==141 {print}' test2.dat**

```
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17           N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38           C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33           C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98           O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89           C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74           C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23           C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17           C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63           N
```

## 33. Find the difference between X and Y coordinates of all atoms

**awk '$1~/ATOM/ {print $8-$7}' test2.dat**

```
57.676
53.017
51.361
49.556
48.639
50.704
51.665
52.248
52.523
54.148
18.432
20.06
18.515
19.495
17.307
16.55
14.48
13.286
```

```
ATOM      8  ND2 ASN A 140     -14.365  43.311  15.200  1.00 41.86           N
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17           N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38           C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33           C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98           O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89           C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74           C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23           C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17           C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63           N
ATOM   1497  ND1 HIS B 210      12.713  31.145  12.352  1.00 24.71           N
ATOM   1498  CD2 HIS B 210      13.114  33.174  13.046  1.00 24.85           C
ATOM   1499  CE1 HIS B 210      13.280  31.795  11.352  1.00 23.41           C
ATOM   1500  NE2 HIS B 210      13.532  33.027  11.746  1.00 23.59           N
ATOM   1501  N   ALA B 211      14.746  32.053  16.235  1.00 26.10           N
ATOM   1502  CA  ALA B 211      16.084  32.634  16.246  1.00 26.32           C
ATOM   1503  C   ALA B 211      17.083  31.563  16.650  1.00 26.17           C
ATOM   1504  O   ALA B 211      18.156  31.442  16.068  1.00 26.46           O
```

## 34. Replace by absolute value

```
awk '{for (i=1; i<=NF; i++) if ($i<0) $i=-$i; print}' test6.dat
```

any field

```
awk '{if ($7<0) $7=-$7; print}' test6.dat
```

7th field

```
ATOM 9 N LYS A 141 9.552 43.465 13.292 1.00 39.17 N
ATOM 10 CA LYS A 141 8.497 42.864 12.475 1.00 38.38 C
ATOM 11 C LYS A 141 8.194 41.362 12.471 1.00 37.33 C
ATOM 12 O LYS A 141 7.793 40.846 11.439 1.00 36.98 O
ATOM 13 CB LYS A 141 7.165 43.539 12.800 1.00 38.89 C
ATOM 14 CG LYS A 141 6.899 44.766 11.997 1.00 39.74 C
ATOM 15 CD LYS A 141 6.4 45.848 12.886 1.00 41.23 C
ATOM 16 CE LYS A 141 5.65 46.873 12.073 1.00 42.17 C
ATOM 17 NZ LYS A 141 5.91 48.238 12.590 1.00 42.63 N
ATOM   1497  ND1 HIS B 210      12.713  31.145  12.352  1.00 24.71          N
ATOM   1498  CD2 HIS B 210      13.114  33.174  13.046  1.00 24.85          C
ATOM   1499  CE1 HIS B 210      13.280  31.795  11.352  1.00 23.41          C
ATOM   1500  NE2 HIS B 210      13.532  33.027  11.746  1.00 23.59          N
ATOM   1501  N   ALA B 211      14.746  32.053  16.235  1.00 26.10          N
ATOM   1502  CA  ALA B 211      16.084  32.634  16.246  1.00 26.32          C
ATOM   1503  C   ALA B 211      17.083  31.563  16.650  1.00 26.17          C
ATOM   1504  O   ALA B 211      18.156  31.442  16.068  1.00 26.46          O
```

```
ATOM      9  N   LYS A 141      -9.552  43.465  13.292  1.00 39.17          N
ATOM     10  CA  LYS A 141      -8.497  42.864  12.475  1.00 38.38          C
ATOM     11  C   LYS A 141      -8.194  41.362  12.471  1.00 37.33          C
ATOM     12  O   LYS A 141      -7.793  40.846  11.439  1.00 36.98          O
ATOM     13  CB  LYS A 141      -7.165  43.539  12.800  1.00 38.89          C
ATOM     14  CG  LYS A 141      -6.899  44.766  11.997  1.00 39.74          C
ATOM     15  CD  LYS A 141      -6.400  45.848  12.886  1.00 41.23          C
ATOM     16  CE  LYS A 141      -5.650  46.873  12.073  1.00 42.17          C
ATOM     17  NZ  LYS A 141      -5.910  48.238  12.590  1.00 42.63          N
ATOM   1497  ND1 HIS B 210      12.713  31.145  12.352  1.00 24.71          N
ATOM   1498  CD2 HIS B 210      13.114  33.174  13.046  1.00 24.85          C
ATOM   1499  CE1 HIS B 210      13.280  31.795  11.352  1.00 23.41          C
ATOM   1500  NE2 HIS B 210      13.532  33.027  11.746  1.00 23.59          N
ATOM   1501  N   ALA B 211      14.746  32.053  16.235  1.00 26.10          N
ATOM   1502  CA  ALA B 211      16.084  32.634  16.246  1.00 26.32          C
ATOM   1503  C   ALA B 211      17.083  31.563  16.650  1.00 26.17          C
ATOM   1504  O   ALA B 211      18.156  31.442  16.068  1.00 26.46          O
```

# 35. Summing up the numbers in each line

**awk '{ for(i=1; i<=NF;i++) j+=$i; print j; j=0 }' test5.dat**

```
149
163
100
110.14
166
175
99.98
111.46
297
317
100
106.69
126
143
99.98
106.71
```

```
 16  11   1   8   5  11   2  10  19  16   0   6   6   5   3   7   9  11   0   3
 17  12   0   9   6  12   0  11  20  17   0   7   7   6   4   8  10  12   1   4
10.74  7.38   .67  5.37  3.36  7.38  1.34  6.71 12.75 10.74   .00  4.03  4.03  3.36  2.01  4.70  6.04  7.38   .00  2.01
11.49  8.11   .00  6.08  4.05  8.11   .00  7.43 13.51 11.49   .00  4.73  4.73  4.05  2.70  5.41  6.76  8.11   .68  2.70
 14  13   0  11  10   9   3  11  10  21   2   7  11   1  10   8  10  11   2   2
 15  14   1  12  11  10   4  12  11  22   0   0  12   2  11   9  11  12   3   3
 8.43  7.83   .00  6.63  6.02  5.42  1.81  6.63  6.02 12.65  1.20  4.22  6.63   .60  6.02  4.82  6.02  6.63  1.20  1.20
 9.55  8.92   .64  7.64  7.01  6.37  2.55  7.64  7.01 14.01   .00   .00  7.64  1.27  7.01  5.73  7.01  7.64  1.91  1.91
 41  10   4  29  16  28   1  26  22  17   6  11  13   7   7  11  20  19   1   8
 42  11   5  30  17  29   2  27  23  18   7  12  14   8   8  12  21  20   2   9
13.80  3.37  1.35  9.76  5.39  9.43   .34  8.75  7.41  5.72  2.02  3.70  4.38  2.36  2.36  3.70  6.73  6.40   .34  2.69
14.14  3.70  1.68 10.10  5.72  9.76   .67  9.09  7.74  6.06  2.36  4.04  4.71  2.69  2.69  4.04  7.07  6.73   .67  3.03
 18   4   0   8   9  12   2   4   6  17   1   1   6   5   5   7   9  11   0   1
 19   5   0   9  10  13   3   5   7  18   0   2   7   6   6   8  10  12   1   2
14.29  3.17   .00  6.35  7.14  9.52  1.59  3.17  4.76 13.49   .79   .79  4.76  3.97  3.97  5.56  7.14  8.73   .00   .79
14.18  3.73   .00  6.72  7.46  9.70  2.24  3.73  5.22 13.43   .00  1.49  5.22  4.48  4.48  5.97  7.46  8.96   .75  1.49
```

## 36. Summing up all numbers in a particular column

**awk '{a+=$7} END {print a}' test4.dat**

**132**

| | | | | | | |
|---|---|---|---|---|---|---|
| 1prxacaMS2 | 79 | WSK | DINAYN | SEE | 32.1852 | 6 |
| 1m5sacaMS3 | 20 | IKI | ARVLIT | AAT | 11.4172 | 6 |
| 1m5sacaMS3 | 164 | AGG | NFFIFG | DSQ | 96.2088 | 6 |
| 1m5sacaMS3 | 242 | DVN | AVYEIV | ING | 19.8543 | 6 |
| 1iq6acaMS4 | 50 | GML | LASLFS | GLL | 10.7893 | 6 |
| 1spvacaMS5 | 147 | ALP | EQVYFV | CYD | 96.1525 | 6 |
| 1x7dacaMS7 | 62 | KSR | YAFKYV | NGH | 58.0513 | 6 |
| 1x7dacaMS7 | 149 | GIE | EIVAYD | TDP | 47.4865 | 6 |
| 1x7dacaMS7 | 236 | NAR | VFVEYE | PQT | 74.9182 | 6 |
| 1lzlacaMS8 | 130 | DCY | AALLYI | HAH | 48.2344 | 6 |
| 1lzlacaMS8 | 228 | EDP | DVSIYA | APS | 25.9341 | 6 |
| 1ul1acaMS10 | 35 | MSI | YQFLIA | VRQ | 95.5041 | 6 |
| 1ul1ccaMS11 | 61 | HKE | AHQLFL | EPE | 10.7389 | 6 |
| 1sxjacaMS12 | 22 | DCV | QLVNFQ | CKE | 81.4939 | 6 |
| 1sxjccaMS13 | 26 | LSD | SINIIT | KET | 42.2134 | 6 |
| 1sxjccaMS13 | 111 | KSG | FLQFFL | APK | 14.4927 | 6 |
| 1rypacaMS14 | 15 | GRN | FQVEYA | VKA | 87.3043 | 6 |
| 1yx3acaMS15 | 45 | EHW | DIINFL | REY | 10.2064 | 6 |
| 1ljlacaMS16 | 0 | X | TIYFIC | TGN | 15.1024 | 6 |
| 1p1jacaMS17 | 61 | IAS | NDILYN | DKL | 93.6116 | 6 |
| 1p1jacaMS17 | 92 | KVA | MDEYYS | ELM | 44.0619 | 6 |
| 1m3sacaMS18 | 76 | EEG | DLVIIG | SGS | 54.9764 | 6 |

# References

**Linux in a Nutshell: A desktop quick reference**

**Siever et al. 2009, Oreilly**

**http://www.pement.org/awk/awk1line.txt**

**The AWK Programming Language by Alfred V. Aho, Brian W. Kernighan**

**Effective Awk Programming, 3/ed, 454 Pages by Arnold Robbins**

# Protein sequence analysis

**>sp|P01966|HBA_BOVIN**

**MVLSAADKGNVKAAWGKVGGHAAEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG**

**AKVAAALTKAVEHLDDLPGALSELSDLHAHKLRVDPVNFKLLSHSLLVTLASHLPSDFTP**

**AVHASLDKFLANVSTVLTSKYR**

**What can we do with sequence?**

# Amino acid occurrence

**It is the number of amino acids of each type present in a protein.**

**E.g. THISISAPEPTIDE**

**A: 1; C: 0; D: 1; E: 2 etc.**

```
 1   MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD
81   NLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
```

| A | B | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 15 | | 2 | 7 | 8 | 8 | 13 | 9 | | 11 | 18 | 2 | 6 | 7 | 3 | 3 | 5 | 7 | 18 | 2 | 3 |

# Amino acid composition

It is the number of amino acids of each type normalized with the total number of residues.

It is defined as:

$$Comp(i) = n_i*100.0/N$$

where i stands for the 20 amino acid residues;

$n_i$ is the number of residues of each type

N is the total number of residues.

The summation is through all the residues in the considered protein.

E.g. Comp(Ala) = 15*100/147 = 10.2

Algorithm

# Algorithm

1. Read 20 amino acid residues: E.g. aa(i), i=1,20
2. Read the sequence: seq(i),i=1,n
3. Normalize number of residues: no(i)=0
4. Compare each residue in the sequence with standard 20 residues

```
do i=1,n
do j = 1,20
if (seq(i).eq.aa(j)) no(j) = no(j)+1
```

5. Count the number with respective residues for each match
6. Result will give the occurrence of each residue
7. Normalize with total number of residues

```
comp(i)=no(i)/n
```

amino.dat

ADCEFGHIKLMNPQRSTVWY

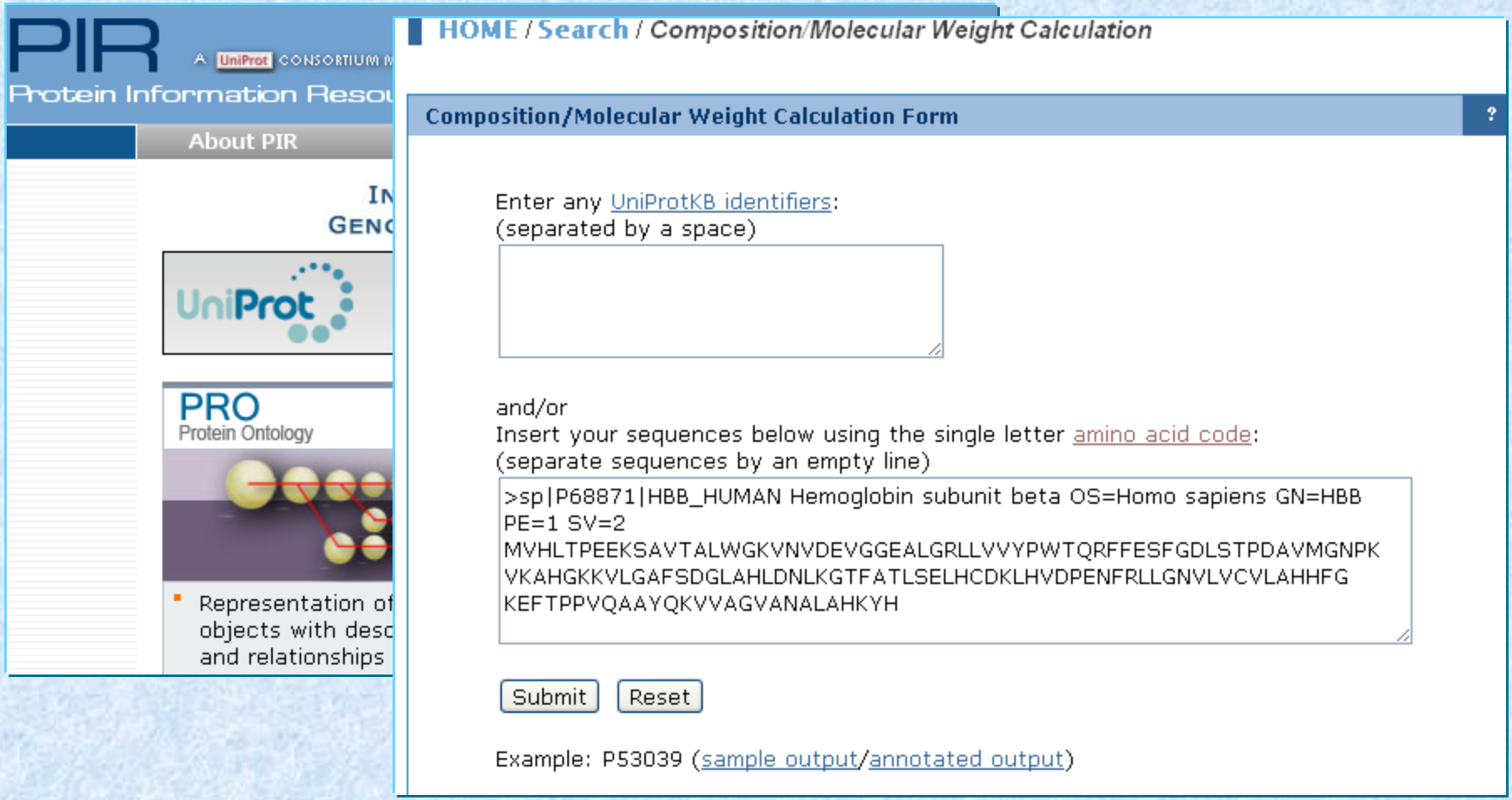
seq.dat

79

MALLPAAPGAPARATPTRWPVGCFNRPWTKWSYDEALDGIKAAGYAWTGLLTASKPSLHH
ATATPEYLAALKQKSRHAA

| | | |
|---|---|---|
| **A** | **17** | **21.52** |
| **D** | **2** | **2.53** |
| **C** | **1** | **1.27** |
| **E** | **2** | **2.53** |
| **F** | **1** | **1.27** |
| **G** | **5** | **6.33** |
| **H** | **3** | **3.80** |
| **I** | **1** | **1.27** |
| **K** | **5** | **6.33** |
| **L** | **8** | **10.13** |
| **M** | **1** | **1.27** |
| **N** | **1** | **1.27** |
| **P** | **8** | **10.13** |
| **Q** | **1** | **1.27** |
| **R** | **4** | **5.06** |
| **S** | **4** | **5.06** |
| **T** | **7** | **8.86** |
| **V** | **1** | **1.27** |
| **W** | **4** | **5.06** |
| **Y** | **3** | **3.80** |

# Composition



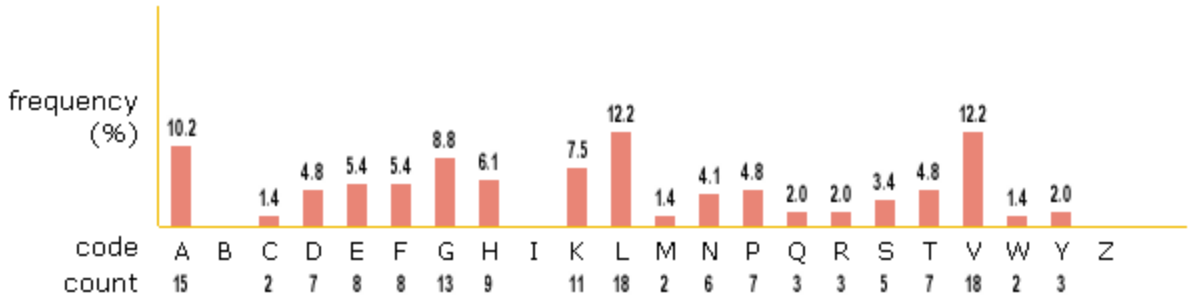
# Composition

Molecular Weight & Composition

**SEQUENCE:**

```
     >SP|P68871|HBB_HUMAN HEMOGLOBIN SUBUNIT BETA OS=HOMO SAPIENS GN=HBB PE=1 SV=2
  1  MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD
 81  NLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
```

**COMPOSITION:**




Number of residues = 147 Molecular weight = 15998.34

# Molecular weight

Molecular weight = n(i)*w(i) – 18*(N-1)

Number of residues = 147 Molecular weight = 15998.34

**CALCULATION NOTES:**

1. Molecular weight = sum of individual residues weights - water molecular weight × ( number of residues - 1 )

*where, water molecular weight = 18.015;*

2. For each residue, the table gives the molecular weight:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | Ala | 89.09 | G | Gly | 75.07 | N | Asn | 132.12 | V | Val | 117.15 |
| B | Asx | 132.61 | H | His | 155.16 | P | Pro | 115.13 | W | Trp | 204.23 |
| C | Cys | 121.15 | I | Ile | 131.17 | Q | Gln | 146.15 | Y | Tyr | 181.19 |
| D | Asp | 133.10 | K | Lys | 146.19 | R | Arg | 174.20 | Z | Glx | 146.64 |
| E | Glu | 147.13 | L | Leu | 131.17 | S | Ser | 105.09 | | | |
| F | Phe | 165.19 | M | Met | 149.21 | T | Thr | 119.12 | X | else | 128.16 |

seq.dat

79

MALLPAAPGAPARATPTRWPVGCFNRPWTKWSYDEALDGIKAAGYAWTGLLTASKPSLHH
ATATPEYLAALKQKSRHAA

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | Ala | 89.09 | G | Gly | 75.07 | N | Asn | 132.12 | V | Val | 117.15 |
| B | Asx | 132.61 | H | His | 155.16 | P | Pro | 115.13 | W | Trp | 204.23 |
| C | Cys | 121.15 | I | Ile | 131.17 | Q | Gln | 146.15 | Y | Tyr | 181.19 |
| D | Asp | 133.10 | K | Lys | 146.19 | R | Arg | 174.20 | Z | Glx | 146.64 |
| E | Glu | 147.13 | L | Leu | 131.17 | S | Ser | 105.09 | | | |
| F | Phe | 165.19 | M | Met | 149.21 | T | Thr | 119.12 | X | else | 128.16 |

| | | |
|---|---|---|
| **A** | **17** | **21.52** |
| **D** | **2** | **2.53** |
| **C** | **1** | **1.27** |
| **E** | **2** | **2.53** |
| **F** | **1** | **1.27** |
| **G** | **5** | **6.33** |
| **H** | **3** | **3.80** |
| **I** | **1** | **1.27** |
| **K** | **5** | **6.33** |
| **L** | **8** | **10.13** |

| | | |
|---|---|---|
| **M** | **1** | **1.27** |
| **N** | **1** | **1.27** |
| **P** | **8** | **10.13** |
| **Q** | **1** | **1.27** |
| **R** | **4** | **5.06** |
| **S** | **4** | **5.06** |
| **T** | **7** | **8.86** |
| **V** | **1** | **1.27** |
| **W** | **4** | **5.06** |
| **Y** | **3** | **3.80** |

79 8129

# Why composition is important?

**>tr|D9PL53|D9PL53**
**MALLPAAPGAPARATPTRWPVGCFNRPWTKWSYDEALDGIK**
**AAGYAWTGLLTASKPSLHHATATPEYLAALKQKSRHAA**

>1a91_
AMENLNMDLLY MAAAVMMGLA AIGAAIGIGI LGGKFLEGAA RQPDLIPLLR
TQFFIVMGLV DAIPMIAVGL GLYVMFAVA

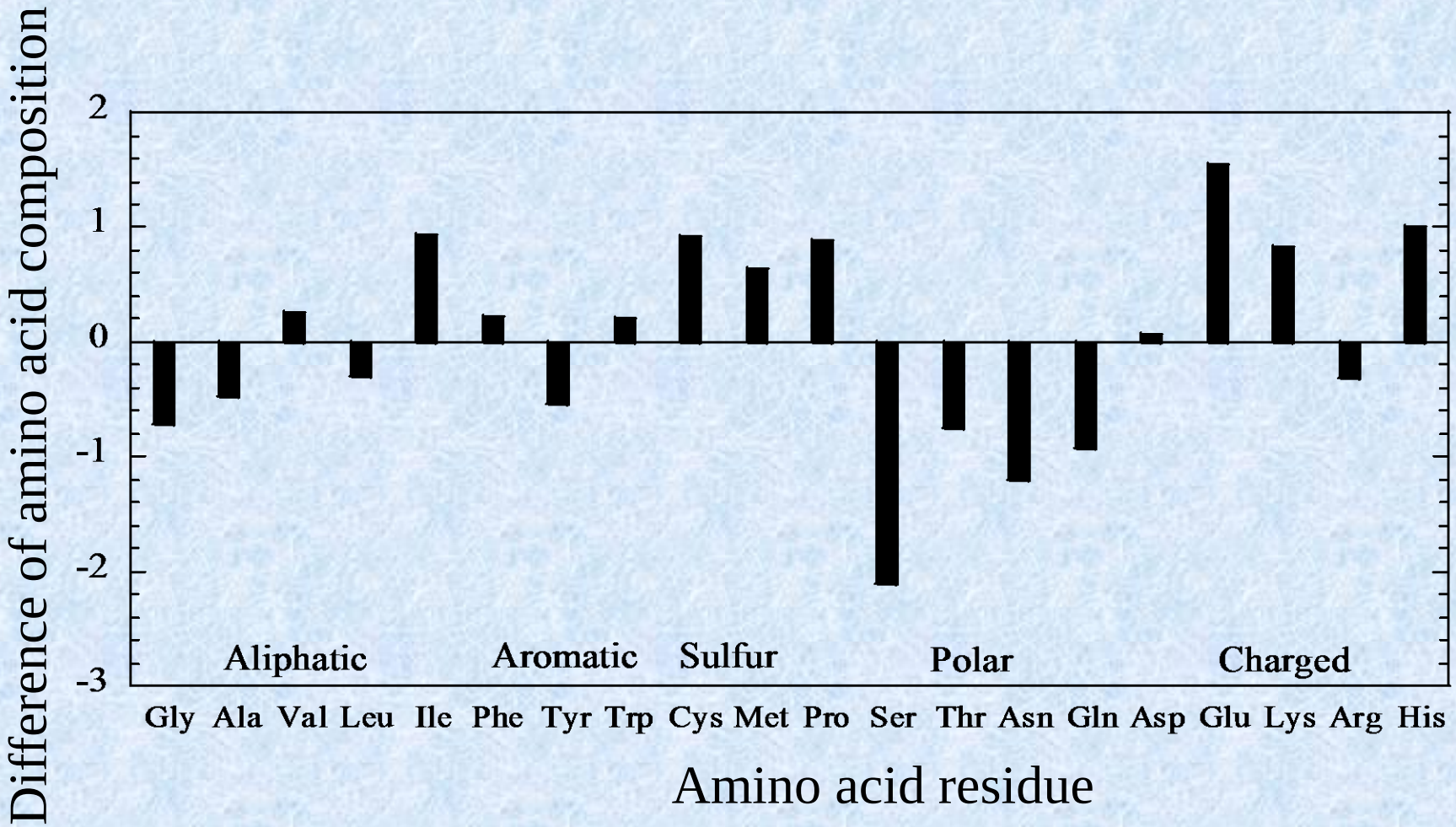
>2erv_
A ADVSAAVGAT GQSGMTYRLG LSWDWDKSWW QTSTGRLTGY WDAGYTYWEG
GDEGAGKHSL SFAPVFVYEF AGDSIKPFIE AGIGVAAFSG TRVGDQNLGS
SLNFEDRIGA GLKFANGQSV GVRAIHYSNA GLKQPNDGIE SYSLFYKIPI

# Example

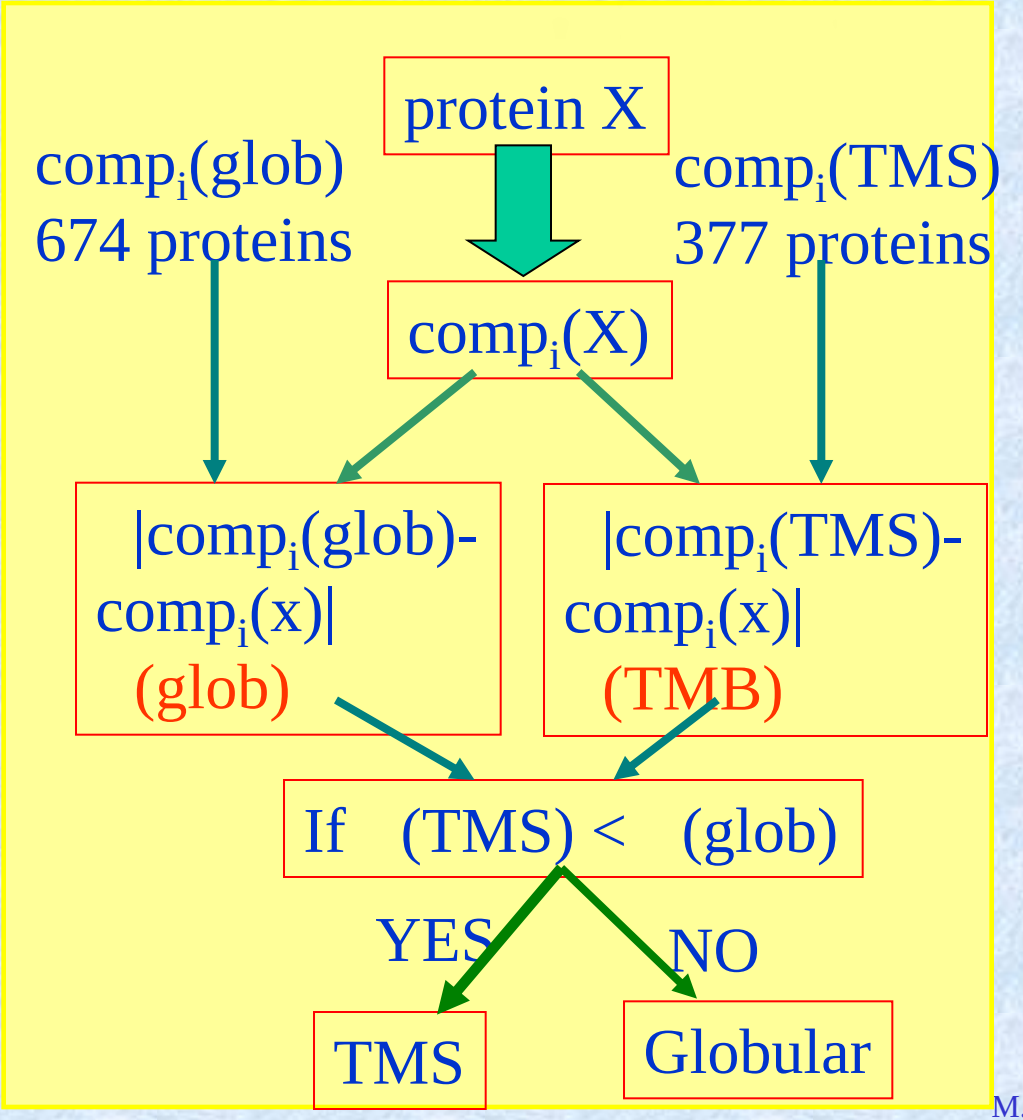
| Residue | Composition (%) | |
|---|---|---|
| | Globular | OMP |
| Ala | 8.47 | 8.95 |
| Asp | 5.97 | 5.91 |
| **Cys** | **1.39** | **0.47** |
| **Glu** | **6.32** | **4.78** |
| Phe | 3.91 | 3.68 |
| Gly | 7.82 | 8.54 |
| **His** | **2.26** | **1.25** |
| **Ile** | **5.71** | **4.77** |
| Lys | 5.76 | 4.93 |
| Leu | 8.48 | 8.78 |

| Residue | Composition (%) | |
|---|---|---|
| | Globular | OMP |
| Met | 2.21 | 1.56 |
| **Asn** | **4.54** | **5.74** |
| Pro | 4.63 | 3.74 |
| **Gln** | **3.82** | **4.75** |
| Arg | 4.93 | 5.24 |
| **Ser** | **5.94** | **8.05** |
| Thr | 5.79 | 6.54 |
| Val | 7.02 | 6.76 |
| Trp | 1.44 | 1.24 |
| Tyr | 3.58 | 4.13 |

# Example


Difference of amino acid composition
2
1
0
-1
-2
-3
Aliphatic
Aromatic
Sulfur
Polar
Charged
Gly Ala Val Leu Ile Phe Tyr Trp Cys Met Pro Ser Thr Asn Gln Asp Glu Lys Arg His
Amino acid residue

# Applications: Discrimination of beta barrel membrane proteins



# Example

**Table 2.** Steps to discriminate globular and outer membrane proteins in two typical proteins

| Residue | Adenovirus DNA-Binding Protein (1ADT) | | | | OutD protein | | | |
|---|---|---|---|---|---|---|---|---|
| | $N$ | Comp | $\sigma_{glob}$ | $\sigma_{OMP}$ | $N$ | Comp | $\sigma_{glob}$ | $\sigma_{OMP}$ |
| Ala | 10 | 11.11 | 2.64 | 2.16 | 54 | 8.31 | 0.16 | 0.64 |
| Asp | 4 | 4.44 | 1.53 | 1.47 | 40 | 6.15 | 0.18 | 0.24 |
| Cys | 1 | 1.11 | 0.28 | 0.64 | 1 | 0.15 | 1.24 | 0.32 |
| Glu | 7 | 7.78 | 1.46 | 3.00 | 31 | 4.77 | 1.55 | 0.01 |
| Phe | 5 | 5.56 | 1.65 | 1.88 | 21 | 3.23 | 0.68 | 0.45 |
| Gly | 3 | 3.33 | 4.49 | 5.21 | 46 | 7.08 | 0.74 | 1.46 |
| His | 4 | 4.44 | 2.18 | 3.19 | 3 | 0.46 | 1.80 | 0.79 |
| Ile | 1 | 1.11 | 4.60 | 3.66 | 35 | 5.38 | 0.33 | 0.61 |
| Lys | 7 | 7.78 | 2.02 | 2.85 | 28 | 4.31 | 1.45 | 0.62 |
| Leu | 10 | 11.11 | 2.63 | 2.33 | 53 | 8.15 | 0.33 | 0.63 |
| Met | 4 | 4.44 | 2.23 | 2.88 | 19 | 2.92 | 0.71 | 1.36 |
| Asn | 4 | 4.44 | 0.10 | 1.30 | 43 | 6.62 | 2.08 | 0.88 |
| Pro | 3 | 3.33 | 1.30 | 0.41 | 21 | 3.23 | 1.40 | 0.51 |
| Gln | 4 | 4.44 | 0.62 | 0.31 | 33 | 5.08 | 1.26 | 0.33 |
| Arg | 3 | 3.33 | 1.60 | 1.91 | 37 | 5.69 | 0.76 | 0.45 |
| Ser | 3 | 3.33 | 2.61 | 4.72 | 55 | 8.46 | 2.52 | 0.41 |
| Thr | 6 | 6.67 | 0.88 | 0.13 | 47 | 7.23 | 1.44 | 0.69 |
| Val | 6 | 6.67 | 0.35 | 0.09 | 64 | 9.85 | 2.83 | 3.09 |
| Trp | 2 | 2.22 | 0.78 | 0.98 | 6 | 0.92 | 0.52 | 0.32 |
| Tyr | 3 | 3.33 | 0.25 | 0.80 | 12 | 1.85 | 1.73 | 2.28 |
| Total | | | 34.18 | 39.89 | | | 23.70 | 16.09 |
| Discrimination | *Globular protein* | | | | *Outer membrane protein* | | | |

$N$: number of residues; $\sigma_{glob} = |\text{comp} - \text{comp(glob)}|$; $\sigma_{OMP} = |\text{comp} - \text{comp(OMP)}|$.

**M.M. Gromiha and M. Suwa (2005)**
***Bioinformatics* 21, 961-968.**

**http://psfs.cbrc.jp/tmbetadisc-comp/**

# TMBETA-DISC: Online servers

## Discrimination of Beta-Barrel Membrane Proteins from Amino Acid Sequen

## TMBETADISC-CO

## Discrimination of Outer Membr

## Amino Acid Compo

Discriminating outer membrane proteins from other folding types of globular and mem
outer membrane proteins (OMPs) from genomic sequences and for the successful pred

We have systematically analyzed the amino acid composition of globular proteins from
proteins. We found that the residues, Glu, His, Ile, Cys, Gln, Asn and Ser show a signif
proteins. Based on this information, we have devised a statistical method for discrimin
membrane proteins. Our approach correctly picked up the outer membrane proteins wi
proteins. On the other hand, our method has correctly excluded the globular proteins a
674 proteins.

TMBETADISC-COMP discriminates the beta-barrel membrane proteins using
membrane protein enter your sequence in a single letter code in the following b

```
MAPKDNTWYTGAKLGWSQYHDTGLINNNGPTHENKLGAGAFGGYQVNPYVGFEMGYDWLG
RMPYKGSVENGAYKAQGVQLTAKLGYPITDDLDIYTRLGGMVWRADTYSNVYGKNHDTGV
SPVFAGGVEYAITPEIATRLEYQWTNNIGDAHTIGTRPDNGMLSLGVSYRFG
```

Predict Reset

| Residue | Occurrence | Composition(%) | Globular Protein Diff. | OMP Diff. |
|---|---|---|---|---|
| A | 12 | 6.98 | 1.49 | 1.97 |
| D | 10 | 5.81 | 0.16 | 0.10 |
| C | 0 | 0.00 | 1.39 | 0.47 |
| E | 6 | 3.49 | 2.83 | 1.29 |
| F | 4 | 2.33 | 1.58 | 1.35 |
| G | 26 | 15.12 | 7.30 | 6.58 |
| H | 4 | 2.33 | 0.07 | 1.08 |
| I | 7 | 4.07 | 1.64 | 0.70 |
| K | 7 | 4.07 | 1.69 | 0.86 |
| L | 11 | 6.40 | 2.08 | 2.38 |
| M | 5 | 2.91 | 0.70 | 1.35 |
| N | 12 | 6.98 | 2.44 | 1.24 |
| P | 8 | 4.65 | 0.02 | 0.91 |
| Q | 5 | 2.91 | 0.91 | 1.84 |
| R | 6 | 3.49 | 1.44 | 1.75 |
| S | 6 | 3.49 | 2.45 | 4.56 |
| T | 14 | 8.14 | 2.35 | 1.60 |
| V | 10 | 5.81 | 1.21 | 0.95 |
| W | 5 | 2.91 | 1.47 | 1.67 |
| Y | 14 | 8.14 | 4.56 | 4.01 |
| Total | 172 | - | 37.78 | 36.66 |

Amino acid sequence seems to be an Outer Membrane Protein

# Residue pair preference (Dipeptide composition)

**It is a measure to quantify the preference of amino acid residue pairs in a sequence.**

**Dipep(i,j) = $N_{ij}$*100/ ( $N_i$+ $N_j$)**

**where i,j stands for the distribution of 20 amino acid residues at positions i and i+1.**

**$N_{i,j}$ is the number of residues of type i followed by the residue j.**

**$N_i$ and $N_j$ are the total number of residues of type i and j, respectively.**

**E.g. DIPEPTIDESAREPEPTIDEPAIRS**

**Dipep(PE) = 2*100/(5+5) = 20**

**>tr|D9PL53|D9PL53**
**MALLPAAPGAPARATPTRWPVGCFNRPWTKWSYDEALDGIK**
**AAGYAWTGLLTASKPSLHHATATPEYLAALKQKSRHAA**

| | | |
|---|---|---|
| AA | 4 | 11.76 |
| AD | 0 | .00 |
| AC | 0 | .00 |
| AE | 0 | .00 |
| AF | 0 | .00 |
| AG | 1 | 4.55 |
| AH | 0 | .00 |
| AI | 0 | .00 |
| AK | 0 | .00 |
| AL | 3 | 12.00 |

# Hydrophobicity profile

**Hydrophobicity profile is simply the plot of the hydrophobicity indices of the residues against their sequence numbers.**

**E.g. SAMPLEDATAWITHHYDINDICES**

**-1, 1, 1, -1, 2, -2, -2, 1, -1, 1, 2, 2, -1, -2, -2, 1, -2, 2, -1, -2, 2, 1, -2, -1**

>1a91_
A MENLNMDLLY MAAAVMMGLA AIGAAIGIGI
LGGKFLEGAA RQPDLIPLLR TQFFIVMGLV
DAIPMIAVGL GLYVMFAVA

**A, C, G, M, Y: 1**
**F, I, L, V, W: 2**
**D, E, H, K, R: -2**
**N, P, Q, S, T: -1**

# Sample data

**Nozaki-Tanford-Jones (Ht)**

**A: 0.87 D: 0.66 C:1.52 E: 0.67**

**F: 2.87 G: 0.10 H: 0.87 I: 3.15**

**K: 1.64 L: 2.17 M: 1.67 N: 0.09**

**P: 2.77 Q: .00 R: 0.85 S: 0.07**

**T: 0.07 V: 1.87 W: 3.77 Y: 2.67**

**Ponnuswamy-Gromiha (Hgm)**

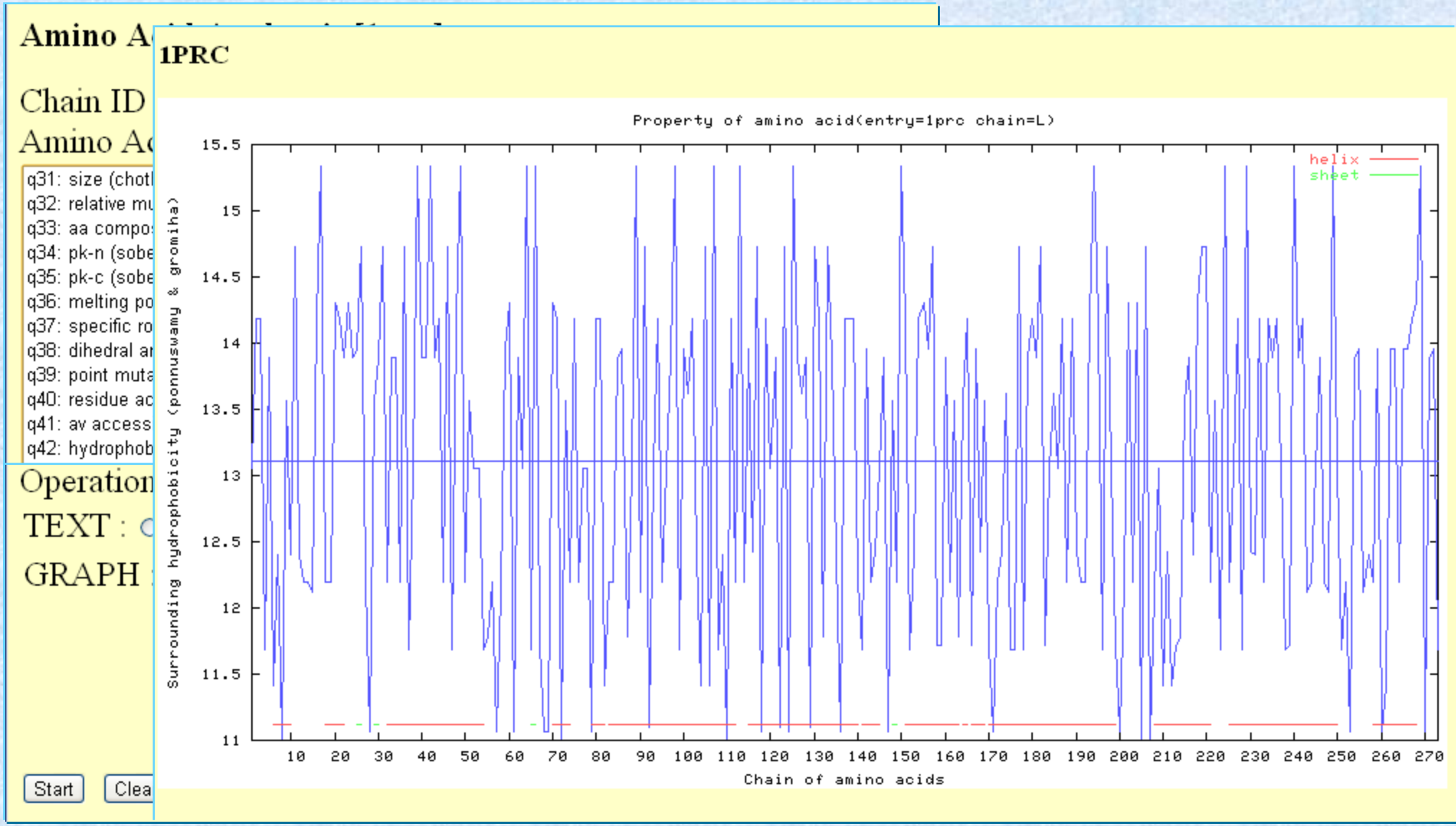
**A: 13.85 D: 11.61 C: 15.37 E: 11.38**

**F: 13.93 G: 13.34 H: 13.82 I: 15.28**

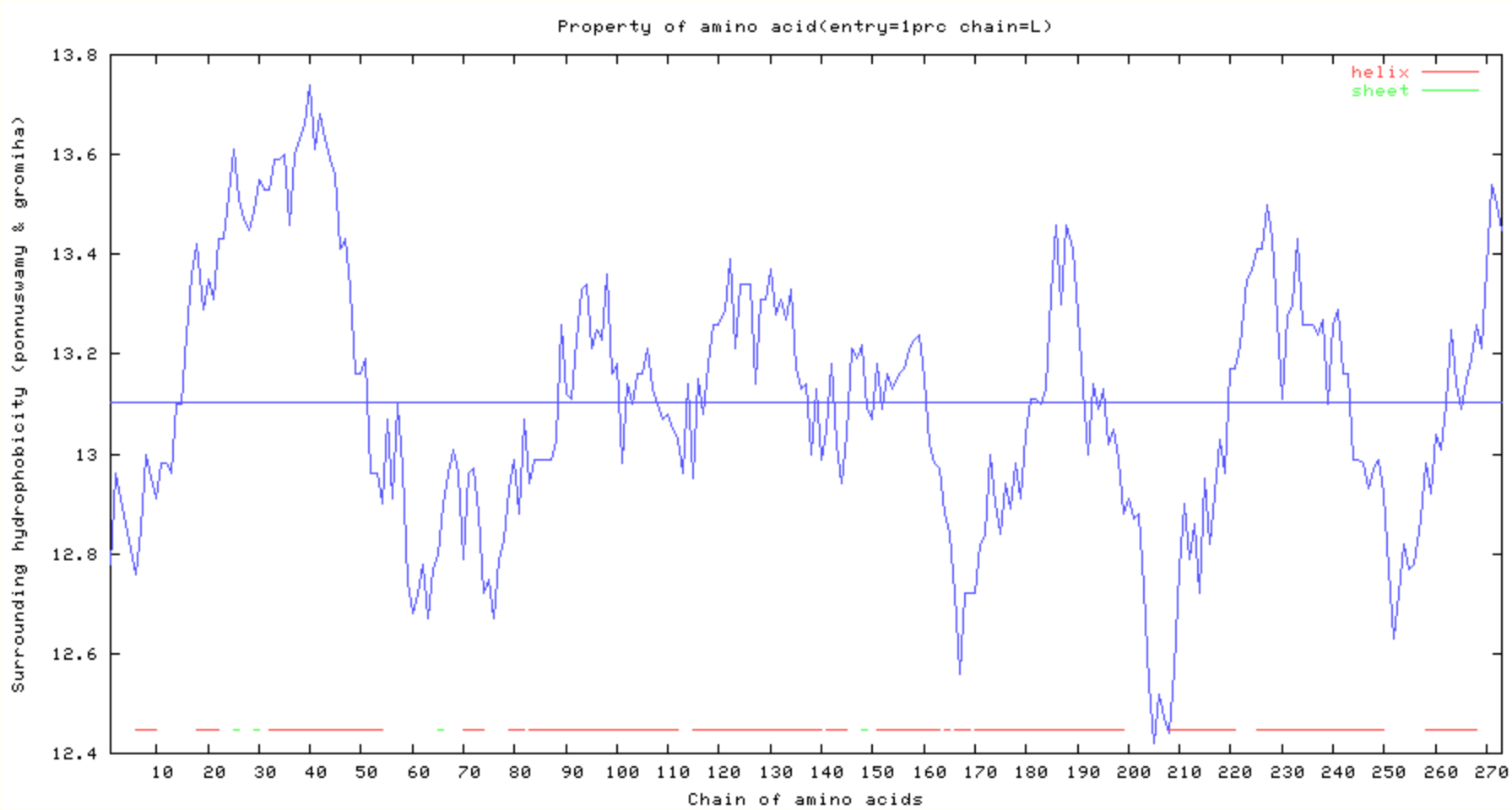
**K: 11.58 L: 14.13 M: 13.86 N: 13.02**

**P: 12.35 Q: 12.61 R: 13.10 S: 13.39**

**T: 12.70 V: 14.56 W: 15.48 Y: 13.88**


Amino A
Chain ID
Amino Ac
q31: size (chotl
q32: relative mu
q33: aa compos
q34: pk-n (sobe
q35: pk-c (sobe
q36: melting po
q37: specific ro
q38: dihedral ar
q39: point muta
q40: residue ac
q41: av access
q42: hydrophob
Operation
TEXT :
GRAPH
Start
Clea
1PRC
Property of amino acid(entry=1prc chain=L)
helix
sheet
Surrounding hydrophobicity (ponnuswamy & gromiha)
15.5
15
14.5
14
13.5
13
12.5
12
11.5
11
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270
Chain of amino acids

**1PRC**




Open
TEX
GRA

Start

# Amphipathicity

**Amphipathic character of amino acid residues is the periodicities in the ploar/nonpolar character of the amino acid sequence in a protein.**

**This has been examined by assigning a numerical hydrophobicity to each residue and searching for periodicity in the resulting one-dimensional function.**

# Amphipathicity: α-helices

The residues of an α-helical segment are considered on four adjacent edges along the direction of the helical axis. The average hydrophobicity of the residues constituting the edge i (i = 1,4) is given by

$$\alpha_i = (\ \ h_{i+j})/n,$$

where n is the total number of residues in the edge,

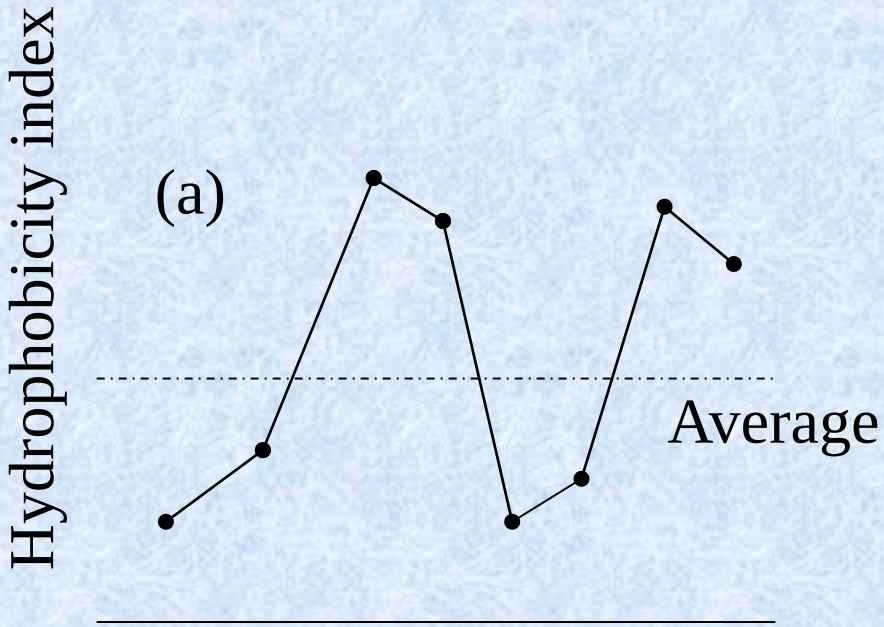
j increases at an interval of 4 from 0 to m, m being the number of residues in the helix;

h is the hydrophobic index of the residue.

The power of amphipathicity of a helix is taken to be

$$A_\alpha = |(a_1+a_2) - (a_3+a_4)| \text{ or } |(a_1+a_4) - (a_2+a_3)|.$$

It has been reported that 75% of the helical segments in known structures are amphipathic in nature.



# Amphipathicity: β-strands

A β-strand segment is considered to have two faces and the average hydrophobicity of residues constituting the face i (i = 1, 2) is given by

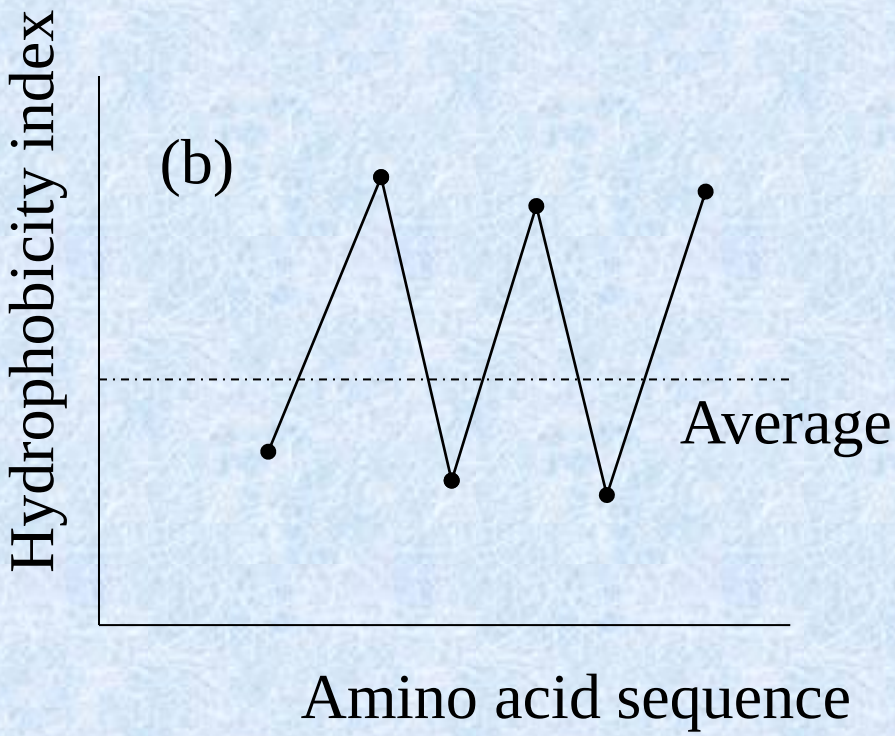
$$\beta_i = (\ \ h_{i+j})/n,$$

where n is the total number of residues in the face, j increases at an interval of 2 from 0 to m, m being the number of residues in the strand;

The amphipathicity index of a strand is computed using the equation,

$$A_\beta = |\beta_1 - \beta_2|.$$

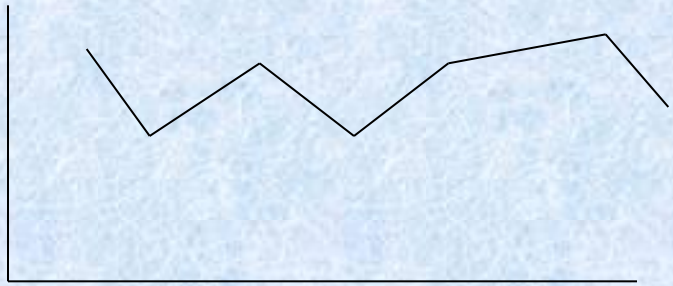
The structural analysis showed that about 65% of the β-strands possess amphipathic character.



# Patterns

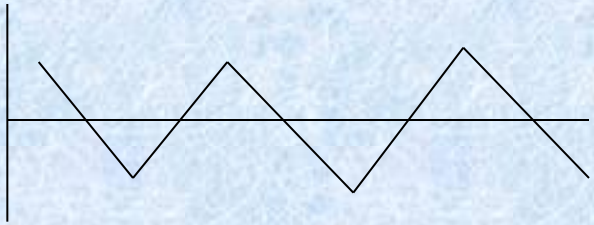
- Identify the pattern of hydrophobic residues for membrane spanning helical proteins

E.g. AILVGYWFFVVA



- Amphiphathic character of β-strands by alternative hydrophobic-hydrophilic residues

AKINIHVTFKIKLP



# Pattern Definition

**[LIVM]-[VIC]-x(2) -G-[DENQTA]-x-[GAC]-x(2)-[LIVMFY](4)-x (2)-G**

**1. Use capital letters for amino acid residues**

**2. Use "[…]" for a choice of multiple amino acids in a particular position. [LIVM] means that L, I, V, or M can be in the first position**

**3. Use "{…}" to exclude amino acids. {CF} means C and F should not be in that particular position**

**4. Use "x" or "X" for a position that can be any amino acid.**

**5. Use "(n)", where n is a number, for multiple positions; x(3) is the same as "xxx"**

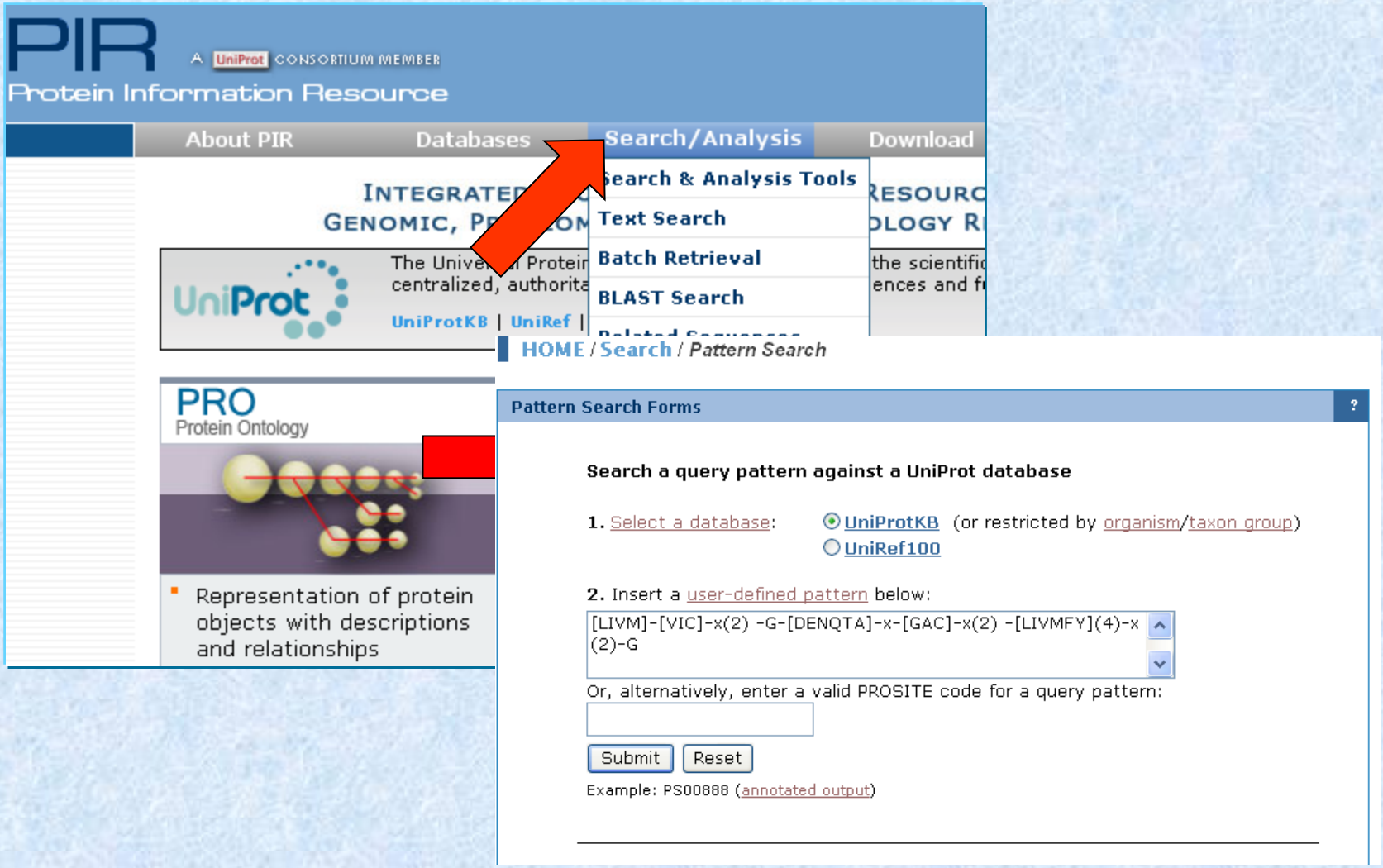
# PIR: Pattern Definition

**[LIVM]-[VIC]-x(2) -G-[DENQTA]-x-[GAC]-x(2)-[LIVMFY](4)-x (2)-G**

**Illustrates a 17 amino acid peptide that has:**

**L, I, V, or M at position 1;**
**V, I, or C at position 2;**
**any residue at positions 3 and 4;**
**G at position 5 and so on ….**

# PIR: Pattern search



# [LIVM]-[VIC]-x(2) -G-[DENQTA]-x-[GAC]-x(2) -[LIVMFY](4)-x(2)-G

Pattern Search Result (UniProtKB)

Query Pattern On ✓ | Help?

725 proteins | 37 pages | 20 / page | ⏮ ⏪ 22 | 23 | 24 | 25 | 26 ⏩ ⏭ Save Result As: TABLE | FASTA

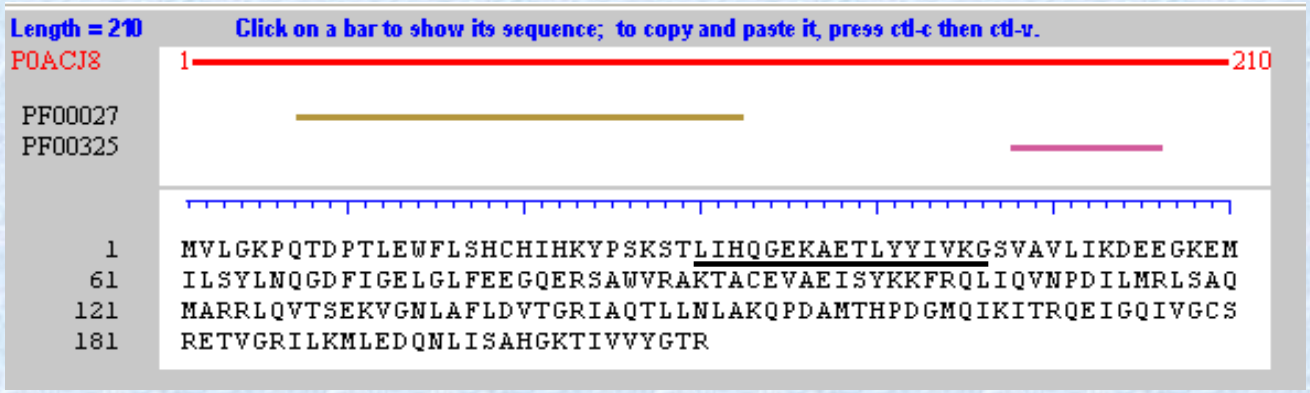
1 selected (show) • BLAST • FASTA • Pattern Match • Multiple Alignment • Domain Display

| Protein AC/ID | Protein Name | Length | Organism Name | PIRSF ID | Match Range |
|---|---|---|---|---|---|
| O96777/O96777_HELVI (ProClass, UniProtKB/TrEMBL) | Cyclic nucleotide and voltage-activated ion channel | 678 | Heliothis virescens (Tobacco budworm moth) | | 477-493; |
| O97119/O97119_LIMPO (ProClass, UniProtKB/TrEMBL) | Cyclic nucleotide-gated ion channel LCNG1 | 900 | Limulus polyphemus (Atlantic horseshoe crab) | | 532-548; |
| P0A1L7/SPAQ_SALTY (ProClass, UniProtKB/Swiss-Prot) | Surface presentation of antigens protein spaQ | 86 | Salmonella typhimurium | PIRSF004669 | 4-20; |
| P0A1L8/SPAQ_SALTI (ProClass, UniProtKB/Swiss-Prot) | Surface presentation of antigens protein spaQ | 86 | Salmonella typhi | PIRSF004669 | 4-20; |
| P0A1M2/SPAQ_SALSE (ProClass, UniProtKB/Swiss-Prot) | Surface presentation of antigens protein spaQ | 86 | Salmonella senftenberg | PIRSF004669 | 4-20; |
| P0A1M3/SPAQ_SALTP (ProClass, UniProtKB/Swiss-Prot) | Surface presentation of antigens protein spaQ | 86 | Salmonella typhisuis | PIRSF004669 | 4-20; |
| P0ACJ8/CRP_ECOLI (ProClass, UniProtKB/Swiss-Prot) | Catabolite gene activator; (AltName: Full=cAMP receptor protein; AltName: Full=cAMP regulatory protein) | 210 | Escherichia coli (strain K12) | PIRSF003151 | 30-46; |
| P0ACJ9/CRP_ECOL6 (ProClass, UniProtKB/Swiss-Prot) | Catabolite gene activator; (AltName: Full=cAMP receptor protein; AltName: Full=cAMP regulatory protein) | 210 | Escherichia coli O6 | PIRSF003151 | 30-46; |
| P0ACK0/CRP_ECO57 (ProClass, UniProtKB/Swiss-Prot) | Catabolite gene activator; (AltName: Full=cAMP receptor protein; AltName: Full=cAMP regulatory protein) | 210 | Escherichia coli O157:H7 | PIRSF003151 | 30-46; |
| P29973/CNGA1_HUMAN (ProClass, UniProtKB/Swiss-Prot) | cGMP-gated cation channel alpha-1; (AltName: Full=CNG channel alpha-1; Short=CNG-1; Short=CNG1; AltName: Full=Cyclic nucleotide-gated channel alpha-1; AltName: Full=Cyclic nucleotide-gated channel, photoreceptor; AltName: Full=Cyclic nucleotide-gated cation channel 1; AltName: Full=Rod photoreceptor cGMP-gated channel subunit alpha) | 690 | Homo sapiens (Human) | PIRSF002403 | 506-522; |
| P29974/CNGA1_MOUSE (ProClass, UniProtKB/Swiss-Prot) | cGMP-gated cation channel alpha-1; (AltName: Full=CNG channel alpha-1; Short=CNG-1; Short=CNG1; AltName: Full=Cyclic nucleotide-gated channel alpha-1; AltName: Full=Cyclic nucleotide-gated channel, photoreceptor; AltName: Full=Cyclic nucleotide-gated cation channel 1; AltName: Full=Rod photoreceptor cGMP-gated channel subunit alpha) | 684 | Mus musculus (Mouse) | PIRSF002403 | 498-514; |
| P36600/KAPR_SCHPO (ProClass, UniProtKB/Swiss-Prot) | cAMP-dependent protein kinase regulatory subunit; ( Short=PKA regulatory subunit) | 412 | Schizosaccharomyces pombe (Fission yeast) | PIRSF000548 | 171-187;305-321; |
| P49605/KAPR_USTMA (ProClass, UniProtKB/Swiss-Prot) | cAMP-dependent protein kinase regulatory subunit; ( Short=PKA regulatory subunit) | 525 | Ustilago maydis (Smut fungus) | PIRSF000548 | 241-257;375-391; |

Links to iProClass and UniProtKB

Link to NCBI taxonomy

Link to PIRSF report




Pattern: LIHQGEKAETLYYIVKG

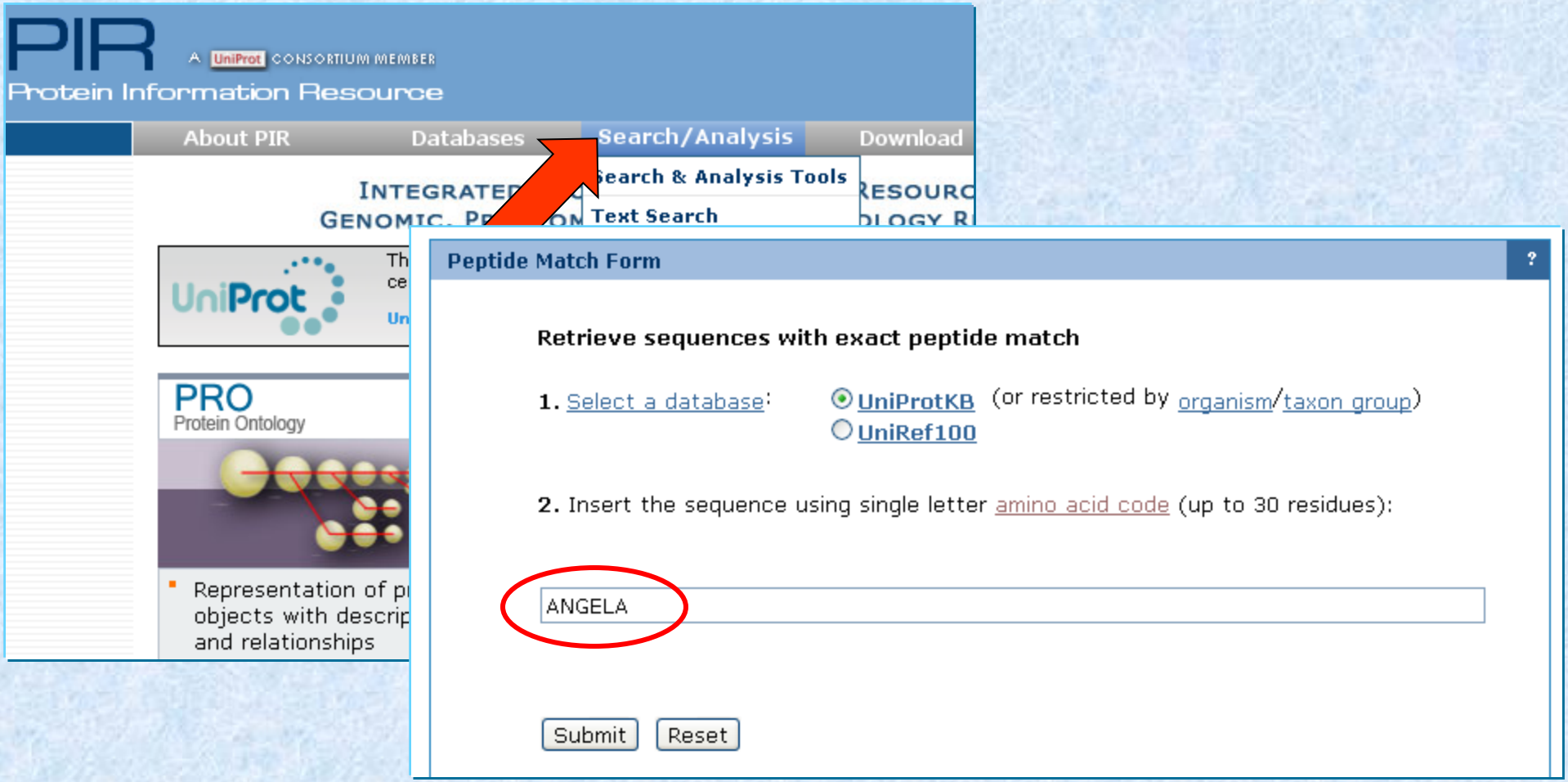
**[LIVM]-[VIC]-x(2) -G-[DENQTA]-x-[GAC]-x(2)-[LIVMFY](4)-x (2)-G**

**New β-signal motif : $P_o x G h_y x H_y x H_y$**
**[K,R,H,Q,N,S,T].G [I,V,L,F,M,Y,W,A,C].[I,V,L,F,M,Y,W].[I,V,L,F,M,Y,W]**

## Algorithm

**K. Imai, M.M. Gromiha and P. Horton (2008) Cell**

# PIR: Specific Peptide search



# PIR: Specific Peptide search

PIR A UniProt CONSORTIUM MEMBER
Protein Information Resource

Protein Search | Site Search

About PIR | Databases | Search/Analysis | Download | Support

Peptide Match Result (UniProtKB)

Query Peptide On | Help?

415 proteins | 21 pages | 20 / page | 1 | 2 | 3 | 4 | 5

Save Result As: TABLE | FASTA

check&analyze

BLAST • FASTA • Pattern Match • Multiple Alignment • Domain Display

| Protein AC/ID | Protein Name | Length | Organism | PIRSF ID | Match Range |
|---|---|---|---|---|---|
| A0K205/A0K205_ARTS2 ProClass UniProtKB/TrEMBL | NmrA family protein; BioThesaurus | 286 | Arthrobacter sp. (strain FB24) | | 254 - 259 TYTAIANGELAGPTSD |
| A0P1M1/A0P1M1_9RHOB ProClass UniProtKB/TrEMBL | HupV protein (Fragment); BioThesaurus | 117 | Labrenzia aggregata IAM 12614 | | 85 - 90 GLSMPANGELAVNLIH |
| A0Q324/A0Q324_CLONN ProClass UniProtKB/TrEMBL | Thermolysin metallopeptid... BioThesaurus | 452 | Clostridium novyi (strain NT) | PIRSF029893 | 446 - 451 QIIKTANGELAL |
| A0Y5Y4/A0Y5Y4_9GAMM ProClass UniProtKB/TrEMBL | Putative uncharacterized ... BioThesaurus | 627 | Alteromonadales bacterium TW-7 | | 455 - 460 VVINKANGELAGDLSL |
| A1AR21/A1AR21_PELPD ProClass UniProtKB/TrEMBL | Putative uncharacterized ... BioThesaurus | 341 | Pelobacter propionicus (strain DSM 2379) | | 126 - 131 VRSSAANGELAGIFDR |
| A1CYJ3/A1CYJ3_NEOFI ProClass UniProtKB/TrEMBL | ABC transporter, putative; BioThesaurus | 1461 | Neosartorya fischeri (strain ATCC 1020 / DSM 3700 / FGSC A1164 / NRRL 181) (Aspergillus fischerianus) | | 349 - 354 KDEGAANGELAEEDAD |
| A1KBE0/A1KBE0_AZOSB ProClass UniProtKB/TrEMBL | Putative uncharacterized ... BioThesaurus | 309 | Azoarcus sp. (strain BH72) | | 146 - 151 SSDAGANGELAKMARW |
| A1S409/A1S409_SHEAM | Signal transduction | 680 | Shewanella amazonensis (strain ATCC BAA-1098 / SB2B) | | 254 - 259 |

# Position specific scoring matrices (Profiles)

**Position specific scoring matrices (PSSM) or profiles express the patterns inherent in a multiple sequence alignment of a set of homologous sequences.**

**The basic idea to use profiles is to match the query sequences from the database against the sequences in the alignment table, giving higher weight to positions that are conserved than to those are variable.**

**These profiles are obtained with a set of probability scores for each amino acid (or gap) at each position of the alignment.**

# Profiles: Applications

(i) they permit greater accuracy in alignments of distantly-related sequences,

(ii) the conservation patterns facilitate identification of other homologous sequences,

(iii) patterns from the sequences are useful in classifying subfamilies within a set of homologues,

(iv) most structure prediction methods are reliable if based on multiple sequence alignment rather than on a single sequence etc.

## a) Alignment Matrix

| | | | | | |
|---|---|---|---|---|---|
| A | A | T | T | G | A |
| A | G | G | T | C | C |
| A | G | G | A | T | G |
| A | G | G | C | G | T |

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | 4 | 1 | 0 | 1 | 0 | 1 |
| C | 0 | 0 | 0 | 1 | 1 | 1 |
| G | 0 | 3 | 3 | 0 | 2 | 1 |
| T | 0 | 0 | 1 | 2 | 1 | 1 |

consensus: A G G T G N

$$\ln \frac{(n_{i,j} + p_i) / (N+1)}{p_i} \approx \ln \frac{f_{i,j}}{p_i}$$

## b) Weight Matrix

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | 1.2 | 0 | -1.6 | 0 | -1.6 | 0 |
| C | -1.6 | -1.6 | -1.6 | 0 | 0 | 0 |
| G | -1.6 | .96 | .96 | -1.6 | .59 | 0 |
| T | -1.6 | -1.6 | 0 | .59 | 0 | 0 |

test sequence: A G G T G C

**Fig. 1.** Examples of the simple matrix model for summarizing a DNA alignment. (**a**) An alignment matrix describing the alignment of the four 6-mers on top. The matrix contains the number of times, $n_{i,j}$, that letter $i$ is observed at position $j$ of this alignment. Below the matrix is the consensus sequence corresponding to the alignment (N indicates that there is no nucleotide preference). (**b**) A weight matrix derived from the alignment in (a). The formula used for transforming the alignment matrix to a weight matrix is shown above the arrow. In this formula, $N$ is the total number of sequences (four in this example), $p_i$ is the *a priori* probability of letter $i$ (0.25 for all the bases in this example) and $f_{i,j} = n_{i,j}/N$ is the frequency of letter $i$ at position $j$. The numbers enclosed in blocks are summed to give the overall score of the test sequence. The overall score is 4.3, which is also the maximum possible score with this weight matrix.

**Bioinformatics. 1999;15(7-8):563-77**

# Method

EQDRLLVELEQP.....AK

PSI-BLAST

PSI-BLAST PSSM

| PROTEIN | A | C | D | :: | Y |
|---|---|---|---|---|---|
| E | -306 | -575 | 428 | :: | -433 |
| Q | -208 | -423 | -285 | :: | -335 |
| D | -180 | -35 | 127 | :: | -48 |
| R | -298 | -549 | 66 | :: | -296 |
| L | -257 | -377 | -569 | :: | -341 |
| L | 307 | -219 | -605 | :: | 626 |
| V | -289 | -31 | -207 | :: | 316 |
| E | -108 | -533 | 405 | :: | -481 |
| L | -248 | -390 | -586 | :: | 199 |
| E | -364 | -632 | 75 | :: | -460 |
| Q | -375 | -472 | -455 | :: | -286 |
| P | -3 | -517 | -261 | :: | -508 |
| : | :: | :: | :: | :: | :: |
| A | 536 | -287 | -397 | :: | -376 |
| K | -240 | -489 | -236 | :: | -358 |

Given amino acid sequence get PSI-BLAST results

**x-min/max-min**

(Normalize PSSM in range of 0-1)

**Normalized PSSM**

| PROTEIN | A | C | D | :: | Y |
|---|---|---|---|---|---|
| E | 0.21 | ***0.08*** | 0.59 | :: | 0.15 |
| Q | 0.26 | 0.15 | 0.22 | :: | 0.20 |
| D | 0.28 | 0.35 | 0.43 | :: | 0.34 |
| R | 0.22 | 0.09 | 0.40 | :: | 0.22 |
| L | **0.24** | 0.18 | 0.08 | :: | 0.19 |
| L | 0.21 | 0.26 | 0.06 | :: | 0.69 |
| V | 0.22 | 0.35 | 0.26 | :: | 0.53 |
| E | 0.31 | ***0.10*** | 0.57 | :: | 0.12 |
| L | 0.24 | 0.17 | 0.07 | :: | 0.47 |
| E | 0.18 | ***0.05*** | 0.41 | :: | 0.13 |
| Q | 0.18 | 0.13 | 0.14 | :: | 0.22 |
| P | 0.37 | 0.11 | 0.24 | :: | 0.11 |
| : | :: | :: | :: | :: | :: |
| A | 0.64 | 0.22 | 0.17 | :: | 0.18 |
| K | 0.25 | 0.12 | 0.25 | :: | 0.19 |

**Normalize it**

**X-min/(max-min)**

**E.g. 5, 6, 9, 2, 1**

**(5-1)/(9-1) = 4/8 = 0.5**

**(9-1)/(9-1) = 8/8/ = 1.0**

**(1-1)/(9-1) = 0/8 = 0.0**

# PSSM-400

**PSSM-400**

**Value of LA=** Σ value of L in column A (shown in bold)

**Value of EC=** Σ value of E in column C (shown in bold and italics)

EC: 0.08, 0.10, 0.05

LA: 0.24, 0.21, 0.24

QD ?

# Applications

**Protein secondary structure prediction**

**Discrimination of proteins belonging to different classes, types etc.**

**Identifying the binding sites, functionally important residues etc.**

# Large scale analysis

**Non redundant sequences**

**No two protein sequences have the sequence identity of more than a specific cutoff (say, 40%).**

**Redundancy cause a bias in any analysis.**

**E.g. Consider two sequences**

**ADIKLAAIKL and KILASDPQWE: Average A is 4/20 = 0.20**

**If one of these sequences appear twice:**

**ADIKLAAIKL, ADIKLAAIKL and KILASDPQWE: Average A is 7/30= 0.23**

**(over-represented)**

**ADIKLAAIKL, KILASDPQWE and KILASDPQWE: Average A is 5/30 = 0.17**

**(under-represented)**

# Programs

CD-HIT

Blastclust

PISCES

**http://cd-hit.org/**

**CD-HIT: Cluster Database at High Identity with Tolerance.**

**The program takes a fasta format sequence database as input and produces a set of 'non-redundant' (nr) representative sequences as output.**

**It uses clustering algorithm and eliminates the redundant sequences.**

**The main advantages of this program are given below:**

**(i) it can handle huge datasets,**

**(ii) it is easy to download and**

**(iii) the results can be obtained quickly.**

**CD-HIT can be used to create the non-redundant dataset of less than 40% sequence identity.**

# Algorithm

**Greedy incremental algorithm: selects representative protein sequence sets**

**Sequences with the identity of more than the threshold will be discarded.**

**Longest sequences, the first and proceed with shorter ones.**

**Sequence identity is the number of identical residues divided by the length of the shorter sequence**

# Short-word filtering system

**Explicit alignment is time consuming**

**Algorithm without aligning.**

**Sequences with >90% sequence identity**

**Decapeptides: query and database (at least 1)**

**Pentapepides: 85%**

**Tetrapeptides: 80%**

**Tripeptides: 75%**

**Dipeptides: 65% -> efficiency decreases**

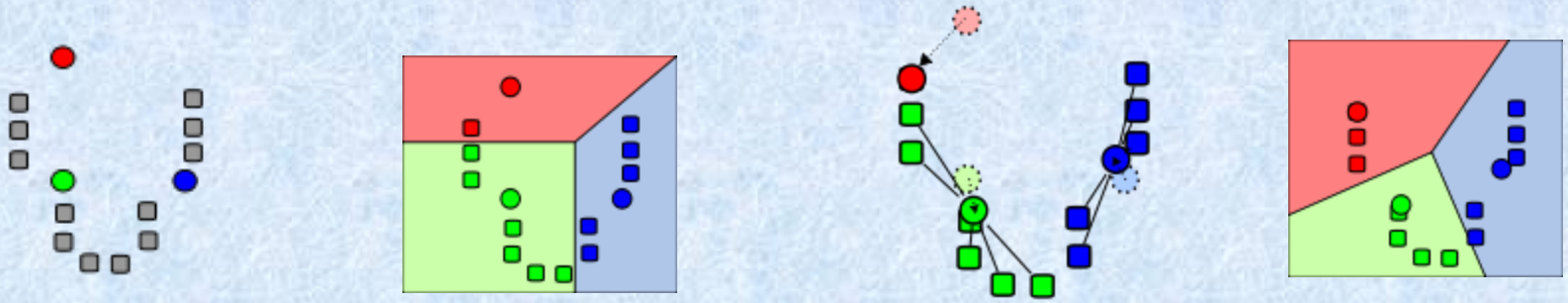
**Compare word size and number of same words with sequence identity**

# Clustering methods

***k*-means clustering is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean.**

observations ($\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$), $k$-means clustering aims to partition the $n$ observations into $k$ sets ($k \leq n$)$\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS):

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$



# Clustering methods based on composition

**Hamming distance**

$$D^H = \quad |Comp(1)_i - comp(2)_i|, \; i=1,20$$

**Euclidean distance**

$$D^E = \{ \quad [Comp(1)_i - comp(2)_i]^2\}^{1/2}$$

# CD-HIT Installation

## Installation

**Most CD-HIT programs were written in C++.**

**Download current CD-HIT at http://bioinformatics.org/cd-hit/**

**Example**

**cd-hit-v4.5.4-2011-03-07.tgz**

**Unpack the file with**

**"tar xvf cd-hit-v4.5.4-2011-03-07.tgz --gunzip"**

**Change directory by "cd cd-hit-v4.5.4-2011-03-07"**

**Compile the programs by "make"**

**Run the program**

# Run CD-HIT

**./cd-hit -i db -o db90 -c 0.9 -n 5**

**db: input file name**

**db90:output file name**

**0.9, means 90% identity (clustering threshold)**

**5 is the size of word**

**Choice of word size:**

**-n 5 for thresholds 0.7 ~ 1.0**

**-n 4 for thresholds 0.6 ~ 0.7**

**-n 3 for thresholds 0.5 ~ 0.6**

**-n 2 for thresholds 0.4 ~ 0.5**

# Example

## ./cd-hit -i hemoglobin_fasta -o db85 -c 0.85 -n 5

```
>sp|P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapiens GN=HBA1 PE=1 SV=2
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR
>sp|P01946|HBA_RAT Hemoglobin subunit alpha-1/2 OS=Rattus norvegicus GN=Hba1 PE=1 SV=3
MVLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFSHIDVSPGSAQVKAHG
KKVADALAKAADHVEDLPGALSTLSDLHAHKLRVDPVNFKFLSHCLLVTLACHHPGDFTP
AMHASLDKFLASVSTVLTSKYR
>sp|P01942|HBA_MOUSE Hemoglobin subunit alpha OS=Mus musculus GN=Hba PE=1 SV=2
MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHG
KKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHPADFTP
AVHASLDKFLASVSTVLTSKYR
>sp|P01966|HBA_BOVIN Hemoglobin subunit alpha OS=Bos taurus GN=HBA PE=1 SV=2
MVLSAADKGNVKAAWGKVGGHAAEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
AKVAAALTKAVEHLDDLPGALSELSDLHAHKLRVDPVNFKLLSHSLLVTLASHLPSDFTP
AVHASLDKFLANVSTVLTSKYR
>sp|P01958|HBA_HORSE Hemoglobin subunit alpha OS=Equus caballus GN=HBA PE=1 SV=2
MVLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
KKVGDALTLAVGHLDDLPGALSNLSDLHAHKLRVDPVNFKLLSHCLLSTLAVHLPNDFTP
AVHASLDKFLSSVSTVLTSKYR
>sp|P69907|HBA_PANTR Hemoglobin subunit alpha OS=Pan troglodyte
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR
>sp|P01959|HBA_EQUAS Hemoglobin subunit alpha OS=Equus asinus G
MVLSAADKTNVKAAWSKVGGNAGEFGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
KKVGDALTLAVGHLDDLPGALSNLSDLHAHKLRVDPVNFKLLSHCLLSTLAVHLPNDFTP
AVHASLDKFLSTVSTVLTSKYR
>sp|P01965|HBA_PIG Hemoglobin subunit alpha OS=Sus scrofa GN=HB
VLSAADKANVKAAWGKVGGQAGAHGAEALERMFLGFPTTKTYFPHFNLSHGSDQVKAHGQ
KVADALTKAVGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHHPDDFNPS
VHASLDKFLANVSTVLTSKYR
>sp|P06635|HBA_PONPY Hemoglobin subunit alpha OS=Pongo pygmaeus
MVLSPADKTNVKTAWGKVGAHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKDHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR
>sp|P60529|HBA_CANFA Hemoglobin subunit alpha OS=Canis familiar
VLSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHFDLSPGSAQVKAHGK
KVADALTTAVAHLDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPTEFTPA
VHASLDKFFAAVSTVLTSKYR
```

```
[gromiha@INSIGHT1 cd-hit-v4.5.4-2011-03-07]$ more db85
>sp|P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapiens GN=HBA1 PE=1 SV=2
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR
>sp|P01946|HBA_RAT Hemoglobin subunit alpha-1/2 OS=Rattus norvegicus GN=Hba1 PE=1 SV=3
MVLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFSHIDVSPGSAQVKAHG
KKVADALAKAADHVEDLPGALSTLSDLHAHKLRVDPVNFKFLSHCLLVTLACHHPGDFTP
AMHASLDKFLASVSTVLTSKYR
>sp|P01965|HBA_PIG Hemoglobin subunit alpha OS=Sus scrofa GN=HBA PE=1 SV=1
VLSAADKANVKAAWGKVGGQAGAHGAEALERMFLGFPTTKTYFPHFNLSHGSDQVKAHGQ
KVADALTKAVGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHHPDDFNPS
VHASLDKFLANVSTVLTSKYR
>sp|P60529|HBA_CANFA Hemoglobin subunit alpha OS=Canis familiaris GN=HBA PE=1 SV=1
VLSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHFDLSPGSAQVKAHGK
KVADALTTAVAHLDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPTEFTPA
VHASLDKFFAAVSTVLTSKYR
[gromiha@INSIGHT1 cd-hit-v4.5.4-2011-03-07]$
```

Connected to 10.93.219.140 SSH2 - aes12

# Blastclust

**Blastclust is a program within the standalone BLAST package used to cluster either protein or nucleotide sequences.**

**The program begins with pairwise matches and places a sequence in a cluster if the sequence matches at least one sequence already in the cluster.**

**In the case of proteins, the blastp algorithm is used to compute the pairwise matches.**

**The general command to create a set of non-redundant set of protein sequences is *blastclust -i infile -o outfile -p T -L .9 -b T -S 95*,**

**where <u>infile</u> and <u>outfile</u> are input and output files, respectively.**

**T stands for protein;**

**the coverage of the length and sequence identity cutoff are 90% (-L .9) and 95% (-S 95), respectively.**

# PISCES

**PISCES is a protein sequence culling server to produce subsets of non-redundant sequences using Protein Data Bank entries or Uniprot sequences in FASTA format.**

**Sequence identities for PDB sequences are determined by the combination of Combinatorial Extension structural alignment and PSI-BLAST alignment.**

**non-PDB sequences are culled with sequence identities from PSI-BLAST. PISCES does not search the non-redundant sequence database, but rather use the user's input sequences as the database.**

**This server will usually be used to cull a related set of sequences, for instance those from a PSI-BLAST search.**

**It takes the amino acid sequence in FASTA format and sends the list of non-redundant protein sequences by e-mail.**

**http://dunbrack.fccc.edu/pisces/**

# PISCES

- PISCES can also therefore provide meaningful re
30%) compared to servers that use only sequence p

## What do you want to do?

- Cull sequences from the whole PDB by resolu
- Cull PDB list which can be created by using P
- Cull from your own list of PDB chains.
- Cull from your own list of GenBank, SwissProt paste the hits listed at the top of BLAST outpu from GenBank.
- Cull from your own file of sequences in FAST use the fragments of sequences from the Sbjc upload)

Submit Re

*>>PISCES --server: Taking input parameters for culling protein sequences*

## Please browse or paste FASTA format sequences or BLAST/PSI-BLAST output file

Paste or type in your FASTA sequences in the following textbox (Help?):

```
>479227|Genbank|Outer membrane
integral membrane protein|outD protein
MLLLSGSVLLMASSLAWSAEFSASFKGTDIQEFINTVSK
NLNKTVIIDPSVSGTITVRSY
DMMNEEQYYQFFLSVLDVYGFTVIPMDNNVLKIIRSKDA
```

Or

**Upload file:**

Browse...

## Set sequence identity threshold:

| | |
|---|---|
| Maximum percentage identity: | 25 |
| Minimum chain length: | 40 |
| Maximum chain length: | 10000 |

Submit Reset

```
> 7467903|Genbank|Outer membrane integral membrane
MSKFTITIFITTLLFTGSVIALDLEQALTEGYKNNEELKAAQIKFLNAIE
QFPQAFSGFMPNVGLQINRQNSKTKYNKKYVNRLGITPRETASTQGILTI
EQSLFNGGASIAALKAAQSGFRASRSEYYAGEQKVLLNLITAYLDCVESK
EKYDISESRVRTNIQQVKTVEEKLRLGEATAIDIAAARAGLAAAETNKLA
AYADFQGKKANFIKVFGIEANDITMPDLPDRLPISLDEFTRKAAKFNPDI
NSARHNVTVTKALEMVQKGKLLPQVSVKLLSGGTNYNPQEPVIQNINNRI
YTTTLSVNIPIYPEGGAQYSRIRSAKNQTRNSVVQLDSAIKQIKAGVVSV
WEGFETAKSRIVAANQGVEAAQISYNGIVQEEIVGSKTILDVLDAEQKLY
EAKITRVDAYKNSVLASYQMKLLTGELTAKSLKLKVKYFSPEEEFNNLKK
KMFIGF
> 11559475|Genbank|Outer membrane integral membrane
MTRNRFVMRRIATTLLVAGIIVSQAAYAQVTLNFVNADIDQVAKAIGAAT
GKTIIVDPRVKGQLNLVAERPVPEDQALKTLQSALRMQGFALVQDHGVLK
VVPEADAKLQGVPTYIGNAPQARGDQVITQVFELHNESANNLLPVLRPLI
SPNNTVTAYPANNTIVVTDYADNVRRIAQIISGVDSAAGAQVQVVPLRNA
NAIDLAAQLQKMLDPGAIGNSDATLKVSVTADPRTNALLLRASNASRLAA
AKRLVQQLDAPSAVPGNMHVVPLRNADAVKLAKTLRGMLGKGGNDSGSSA
SSNDANSFNQNGGSSASGNFSTGTSGTPPLPSGGLGGSSSSSYGGSGGSS
GGGLGTGGLLGGDKDKSGDDNQPGGMIQADSATNSLIITASDPVYRNLRS
VIDQLDARRAQVYIEALIVELNSTTQGNLGIQWQVASGQFLGGTNLAPTA
GNGLGNSIINLTAGGLTNAAGGITGGGLASNLGQLSQGLNIGWLHNMFGV
QGLGALLQYFAGVSDANVLSTPNLITLDNEEAKIVVGQNVPIATGSYSNL
TSGTTSNAFNTYDRRDVGLTLHVKPQITDGGILKLQLYTEDSAVVNGTTN
SQTGPTFTKRSIQSTILADNGEIIVLGGLMQDNYQVSNSKVPLLGDIPWI
GQLFRSESKVRAKTNLMVFLRPVIISDRSTAQEVTSNRYDYIQGVTGAYK
SDNNVIRDKDDPVVPPMPLGPSQGGTAAGNLFDLDKMRRQQLQRQVVPVP
AQPLPEATPAQPQGVPLQAVPQQPLTTAPGASQ
> 7469324|Genbank|Outer membrane integral membrane
MRSNSVKNFRFWLTTEIATCCLLALAPAQAETVSQSNTLDGDLRTAIAGD
SSRDWLQFEKSLEQSLKQKEEIDSWKPSLELMQAKSLVKPGQKLTNIELL
VQELEALSDPLALNFPEPNQTSVAQMAPPSRPMPPPPAGSGQVMFPNPEI
IIQQQGGVPQRGASPQVGNPSILSPAVPVAPVRSRAVPPPVGDLAISNIN
ASFDMIDLGQRGQVNVPSLVLREAPAREVLAVLTRYAGMNLIFTDNQNNE
GTPTPGTPPGGQVAPPQAQSTITLDIQNESVQDVFNYVLMASGLKASRRG
NTIFAGANLLPSARNIITRTIRLNQASAESVASTLASQGAEVNILFEGQE
DVQLAENAPPRVIKQPPTLVPLTVQKPANDSSVLILEGLVVSTDPRLNTV
TLVGEPRNVELASSMITQMDARRRQVAVNVKIIDINLNNIQDYDSSFSFG
IGDSFFVQDSGSAVMRFGDTAPVQEIDINNNLGRITNPPAIVNPFQDGEI
FFDLNRITNIEVPLGPGTIPINFFTSGSGAVSNNPLFNGVTEFPIVEVDE
QGLLTITQPEFGLPSFYQYPKKFQAQIDAQIRSGNAKILTDPTLIVQEGE
AAQVKLTESVIASVDTQVDTQGDTAVRTITPVLEDVGLTLNVIVDRIDDN
GFITLRVNPIVASPAGTQVFDSGAGAINEITLINKRELTSGVVRLRDDQT
FILSGIISELQRSTTSKVPILGDLPVIGALFRQSTDTTDRSEVIILMTPK
IIHDSTEAQFGFRYNPDAATAEFLRQKGFPVQAQP
```

# Sequence id

| IDs | length |
|---|---|
| 7467903\|Genbank\|Outer | 456 |
| 11559475\|Genbank\|Outer | 783 |
| 7469324\|Genbank\|Outer | 785 |
| 15596906\|Genbank\|Outer | 295 |
| P26466\|SwissProt\|Outer | 452 |
| 15597487\|Genbank\|Outer | 452 |
| 5640161\|Genbank\|Outer | 889 |
| P13949\|SwissProt\|Outer | 201 |
| P10170\|SwissProt\|Outer | 260 |
| P16945\|SwissProt\|Outer | 292 |
| 7208425\|Genbank\|Outer | 560 |
| 15598604\|Genbank\|Outer | 891 |
| 13470835\|Genbank\|Outer | 794 |
| P19196\|SwissProt\|Outer | 835 |
| 12620518\|Genbank\|Outer | 230 |
| P31600\|SwissProt\|Outer | 990 |
| 15596468\|Genbank\|Outer | 616 |
| P15727\|SwissProt\|Outer | 482 |
| 7520765\|Genbank\|Outer | 778 |
| P16466\|SwissProt\|Outer | 1577 |
| 3228547\|Genbank\|Outer | 700 |
| P06970\|SwissProt\|Outer | 812 |
| P22340\|SwissProt\|Outer | 505 |
| P44601\|SwissProt\|Outer | 565 |
| P35077\|SwissProt\|Outer | 584 |
| 12721580\|Genbank\|Outer | 444 |
| 7470479\|Genbank\|Outer | 654 |
| P13794\|SwissProt\|Outer | 350 |
| P06111\|SwissProt\|Outer | 257 |
| P24126\|SwissProt\|Outer | 530 |
| P29041\|SwissProt\|Outer | 759 |
| 5759281\|Genbank\|Outer | 462 |

# Phylogeny

**Phylogeny is the description of biological relationships, usually expressed as a tree.**

**A statement of phylogeny among objects assumes homology and depends on classification.**

**Phylogenetic analysis is an investigation of the evolutionary relationships among a group of related sequences by producing a tree representation of the relationships.**

**In fact, phylogenetic relationships among many kinds of organisms are difficult to determine in any other way.**

**Simply, organisms with high degrees of molecular similarity are expected to be closely related than those that are dissimilar.**

**Due to the availability of molecular data, taxonomists are forced to rely on comparisons of phenotypes (how organisms looked) to infer their genotypes (the genes that gave rise to their physical appearance).**

Humans, flies, mollusks: light detecting organ-eye

Protein/DNA sequences

# Phylogenetic trees

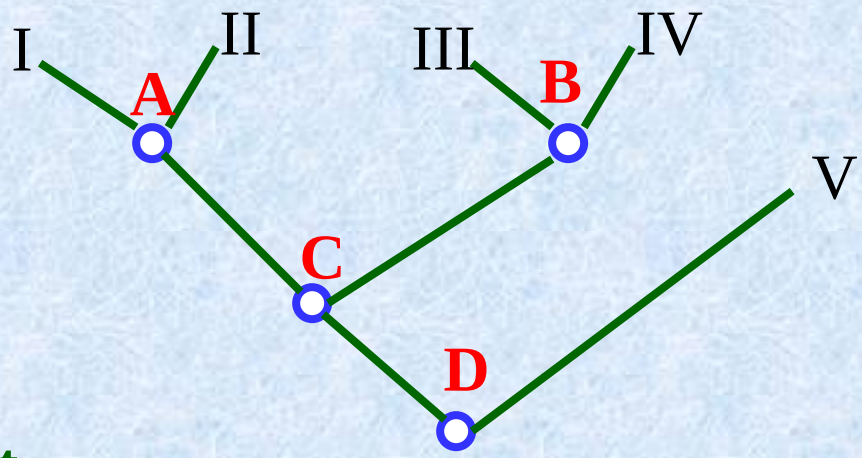
**Graphical representation of the evolutionary relationship among three or more genes or organisms.**

**Relatedness of data; divergence times; nature of common ancestors**

**Nodes and branches:**

**Terminal nodes: at the tip of the branches (gene or organism: available data)**

**Internal node: common ancestor (data are not available)**



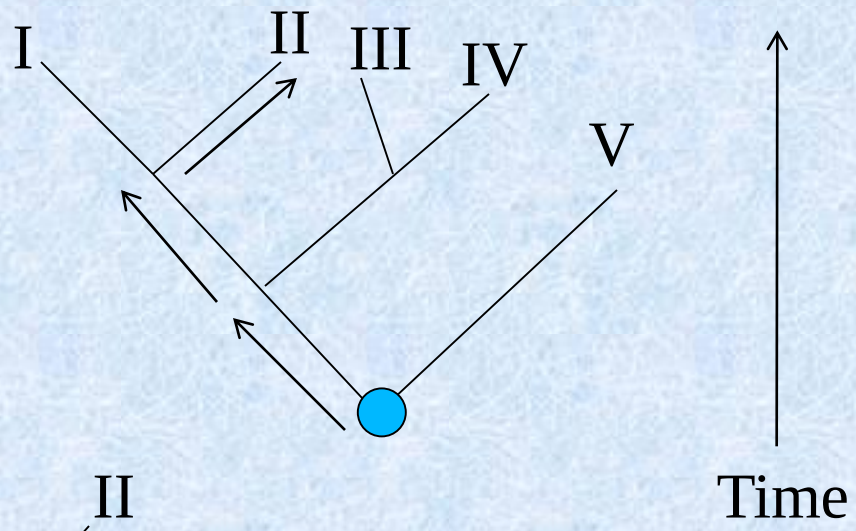

**Bifurcating**
**Multifurcating**

**Scaled trees**
**Unscaled trees**

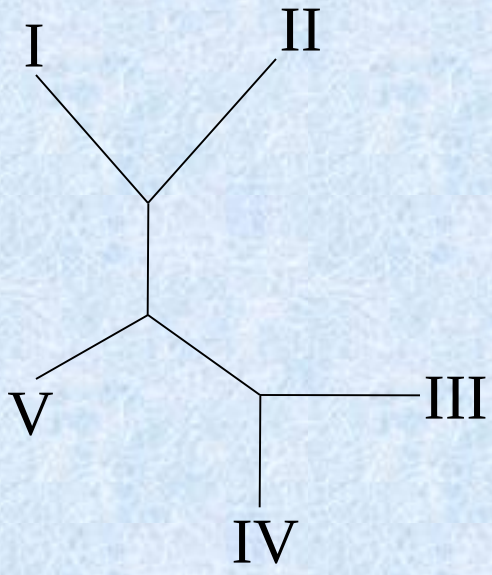
**Newick format**
**(((I,II),(III,IV)), V)**

# Rooted and Unrooted trees

**Rooted trees: a single node is designated as a common ancestor and a unique path leads it.**

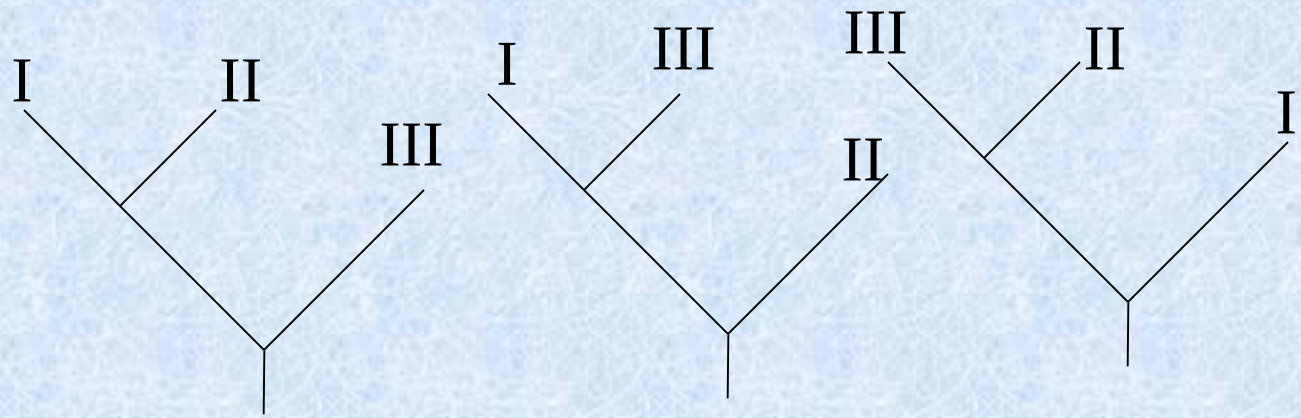



**Unrooted trees: only relationship and not direction**

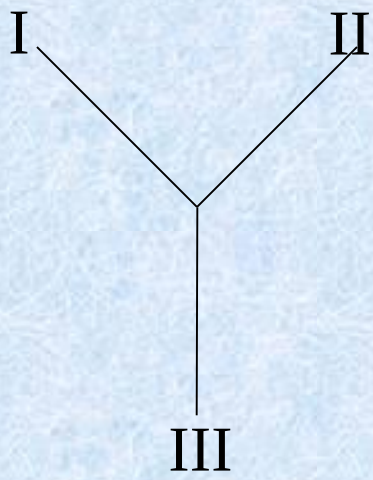


# Possible rooted and unrooted trees
# 3 species

Rooted: 3




Unrooted: 1




$N_R = (2n-3)!/2^{n-2}(n-2)!$

$N_U=(2n-5)!/2^{n-3}(n-3)!$

**2 data: Rooted: 1; Unrooted:1**

**3 data: R: 3, U: 1**

**4 data: R: 15, U: 3**

# Tree construction

1. UPGMA (Unweighted Pair Group Method with Arithmetic mean
2. Transformed Distance Method
3. Neighbor's Relation Method
4. Neighbor Joining Methods
5. Maximum Likelihood Approaches

# UPGMA

**Statistically based method**

**Requires data that can be condensed to genetic distance (distance matrix)**

**E.g. Species, A, B, C and D**

| Species | A | B | C |
|---|---|---|---|
| B | $d_{AB}$ | - | - |
| C | $d_{AC}$ | $d_{BC}$ | - |
| D | $d_{AD}$ | $d_{BD}$ | $d_{CD}$ |

**$d_{AB}$: the number of mismatching nucleotides (divided by total number of sites, where matches could have been found)**

**If A and B are related (minimum mismatches)**

**Form a group (AB)**

**Combine with other species, C and D**

$$d_{(AB),C)} = 1/2\ (d_{AC} + d_{BC})$$

$$d_{(AB),D} = 1/2\ (d_{AD} + d_{BD})$$

**The species separated by the smallest distance in the new matrix can be clustered together.**

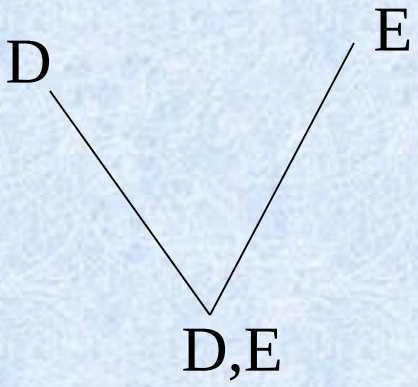
**For scaled branches, the distances will be averaged.**

# Example

```
A: GTGCTGCACG GCTCAGTATA GCATTTACCC TTCCATCTTC AGATCCTGAA
B: ACGCTGCACG GCTCAGTGCG GTGCTTACCC TCCCATCTTC AGATCCTGAA
C: GTGCTGCACG GCTCGGCGCA GCATTTACCC TCCCATCTTC AGATCCTATC
D: GTATCACACG ACTCAGCGCA GCATTTGCCC TCCCGTCTTC AGATCCTAAA
E: GTATCACATA GCTCAGCGCA GCATTTGCCC TCCCGTCTTC AGATCTAAAA
```

**Pairwise comparison (distance matrix)**

| Species | A | B | C | D |
|---|---|---|---|---|
| B | 9 | - | - | - |
| C | 8 | 11 | - | - |
| D | 12 | 15 | 10 | - |
| E | 15 | 18 | 13 | 5 |



**Pairwise comparison (distance matrix)**

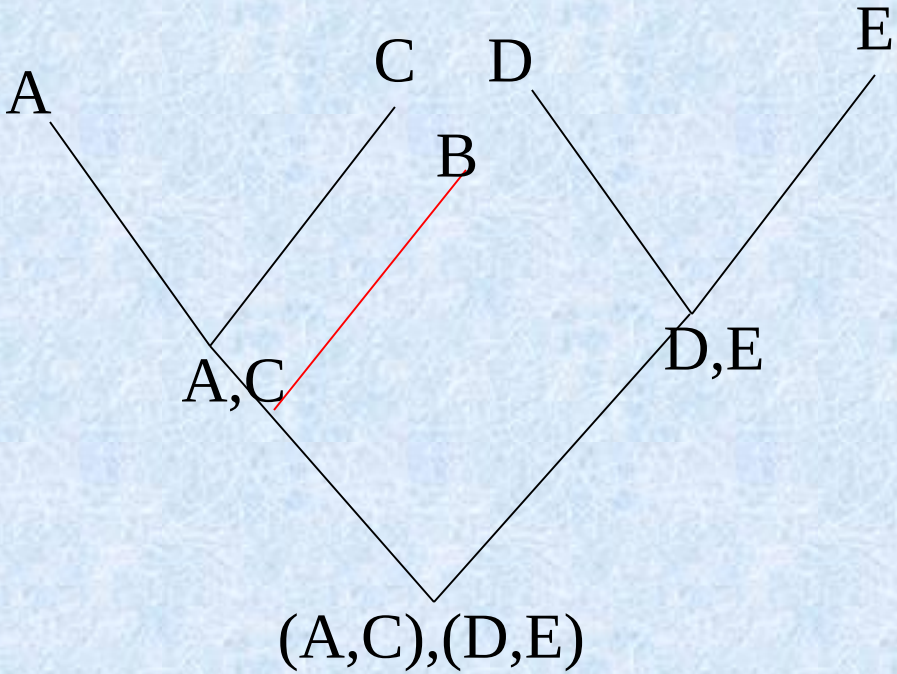
| Species | A | B | C |
|---|---|---|---|
| B | 9 | - | - |
| C | 8 | 11 | - |
| DE | 13.5 | 16.5 | 11.5 |

(AC),B = ½(AB+BC)

**Pairwise comparison**

| Species | B | AC |
|---|---|---|
| AC | 10 | - |
| DE | 16.5 | 12.5 |

(((A,C),B),D,E)



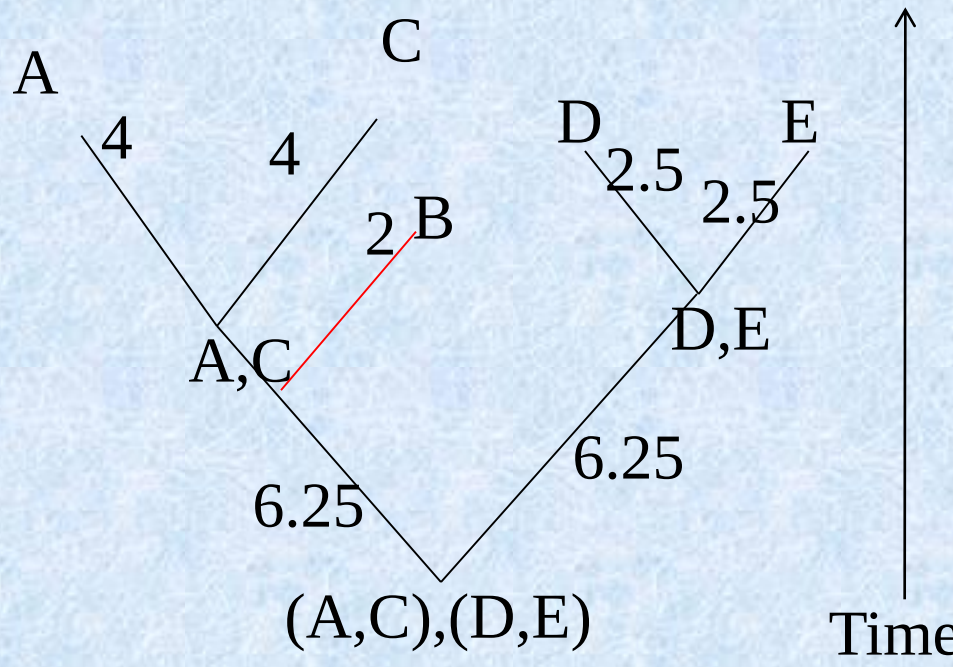

**(((A,C),B),(D,E))**

# Estimation of branch lengths

**Length of the branches can also be calculated with distance matrix**

**Pairwise comparison (distance matrix)**

| Species | A | B | C | D |
|---|---|---|---|---|
| B | 9 | - | - | - |
| C | 8 | 11 | - | - |
| D | 12 | 15 | 10 | - |
| E | 15 | 18 | 13 | 5 |



# Neighbor's Relation Method

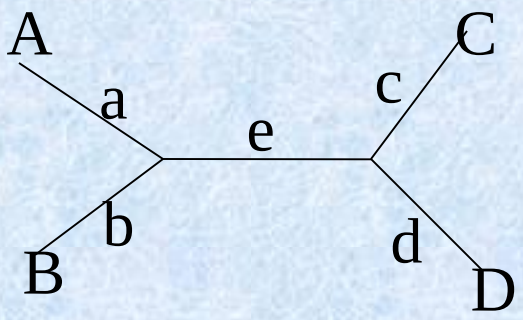
**Another popular variant of UPGMA: tree is constructed with the smallest possible branch lengths overall.**

**Any unrooted tree, pairs of species that are separated from each other by just one internal node are said to be neighbors.**

**[UPGMA: Branch length is not additive.**

**E.g: $d_{AE}$ = 4 + 6.25 + 6.25 + 2.5 = 19**

**Actual case: 15]**




**If additivity holds good:**

$$d_{AC}+d_{BD} = d_{AD}+d_{BC} = a+b+c+d+2e = d_{AB}+d_{CD}+2e$$

**a,b,c,d: lengths of terminal branches**

**e: length of central branch**

**Four point condition:**

$d_{AB}+d_{CD} < d_{AC}+d_{BD}$

**and**

$d_{AB}+d_{CD} < d_{AD}+d_{BC}$

**Considers all possible pairwise arrangements of four species and determines the arrangement, which satisfies the four point condition.**

**For four species, considers all possible values,**

**(i)** $d_{AB}+d_{CD}$**; (ii)** $d_{AC}+d_{BD}$ **and (iii)** $d_{AD}+d_{BC}$

**Smallest sum with pairing is 1 and others are 0**

**Repeat for all possible four pairs**

**Ones with highest scores are grouped.**

**New distance matrix can be generates as was done for UPGMA**

# Neighbor joining method

**Tree is of star-like and all species comes as a single central node regardless of the number.**

**Difference with other methods is the way it determines the sum of branch lengths with each reiteration of the process.**

$$S_{12} = (1/(2(N-2)) \Sigma(d_{1k}+d_{2k}) + (1/2)d_{12} + (1/N-2)\Sigma d_{ij}$$

**Any pair of species can take positions 1 and 2; k is an accepted outgroup**

**Simplified into** $Q_{12} = (N-2)d_{12} - \Sigma d_{1i} - \Sigma d_{2i}$

**All possible pairs are considered and the pairs with smallest distance is taken.**

**Construct new distance matrix as done with UPGMA and repeat the process.**

# Maximum likelihood approaches

**It represent an alternative and purely statistically based method of phylogenetic reconstruction.**

**Probabilities are considered for every individual nucleotide substitution.**

**Transitions (purine to purine/ pyramidine to pyramidine) and transversions**

**Multiple substitutions occurred at one or more sites, which are not necessarily independent.**

**It is necessary to take into account of all these facts, which needs heavy computational power.**

**With current facilities, it is possible to use the method for tree construction.**

# Program to construct trees

- **as Windows executables** (not counting executing in a "DOS box"). Programs available as source code which is Windows-specific are li
(Note that compilers available on Windows systems, particularly the free Cygwin and MinGW compilers, can also be used to compile ma
generic source code). Programs run in interpreted environments such as Perl, Python, R or MATLAB can also be run under Windows if t
programs are listed above under Unix.

| | | | | |
|---|---|---|---|---|
| PHYLIP | DNASIS | Mesquite | MrModeltest | MESA |
| PAUP* | MINSPNET | Phyledit | SymmeTREE | MultiPhyl |
| TREECON | BioEdit | SYN-TAX | TreeJuxtaposer | NimbleTree |
| GDA | ProSeq | PTP | Network | ArboDraw |
| SeqPup | PAL | DIVA | Spectronet | SPAGeDi |
| MOLPHY | WINCLADA | TreeFitter | Phylogen | CBCAnalyzer |
| GeneDoc | NONA | Phylo_win | Phylap | DualBrothers |
| COMPONENT | Phylogenetic Independence | SplitsTree | Dnatree | PaupUp |
| TREEMAP | PEBBLE | PAST | IMa2 | Notung |
| COMPARE | HY-PHY | GeneStudio Pro | ProtTest | SSA |
| RAPDistance | TreeExplorer | Treefinder | GEODIS | Multidivtime |
| TreeView | Genie | PPH | TreeSetViz | ParaFit |
| Phylodendron | Vanilla | MetaPIGA | TreeMe | IDC |
| POPGENE | MEGA | Phyltools | ModelGenerator | TreeMaker |
| TFPGA | TNT | MSA | Simplot | CodonRates |
| GeneTree | GelCompar II | Mgenome | PHYLOGR | BAli-Phy |
| MVSP | Bionumerics | APE | ProfDist | CoMET |
| RSTCALC | TCS | PHASE | START2 | TreeDyn |
| Genetix | FORESTER | PHYML | IQPNNI | DigTree |
| NJplot | Populations | YCDMA | STC | Geneious |
| unrooted | T-REX | NSA | TreeSAAP | Brownie |
| Arlequin | MrBayes | BEAST | Swaap | Mac5 |
| DAMBE | EDIBLE | Clann | Swaap PH | BayesPhylogenies |
| DnaSP | Winboot | Jevtrace | TreeGraph 2 | BayesTraits |
| PAML | r8s | MrMTgui | DIVERGE | MrEnt |

**Phylip**

# Phylip

**Phylip is a program to create phylogenetic tree for a given set of amino acid sequences.**

**It takes the multiple alignment of the sequences as input**

**Multiple sequence alignments can be done with ClustalW, MAFFT etc.**

**MAFFT is widely used to prepare the input multiple alignment file suitable for Phylip**

# MAFFT

mafft.cbrc.jp/alignment/software/

MAFFT version 6

Multiple alignment program for amino acid or nucleotide sequences

CBRC AIST

**Download version**
- Mac OS X
- Windows
- Linux
- Source

Online version
- Alignment
- Phylogeny
- Rough tree

Merits / limitations
Algorithms
Tips
Benchmarks
Feedback

## About

MAFFT is a multiple sequence alignment program for unix-like operating systems. It offers a range of multiple alignment methods, L-INS-i (accurate; for alignment of <~200 sequences), FFT-NS-2 (fast; for alignment of <~10,000 sequences), *etc.*

## Download and Installation

- **Mac OS X**
- **Linux**
- **Windows**
- **Source**
- Changelog

The latest version is 6.857. ***New!*** (2011/05/30)

## Input Format

Fasta format. example1 (LSU rRNA), example2 (protein)

The type of input sequences (amino acid or nucleotide) is automatically recognized.

## Usage

```
% mafft [arguments] input > output
```

# MAFFT

**UPPERCASE / lowercase**:

○ Same as input

**Parameters**:

Scoring matrix for amino acid sequences: BLOSUM62

Scoring matrix for nucleotide sequences: 200PAM / κ=2

† Switch it to '1PAM / κ=2' when aligning closely related DNA sequences.

Gap opening penalty: 1.53 (1.0 – 3.0)

Offset value: 0.0 (0.0 – 1.0)

† If long gaps are not expected, set it as 0.1 or larger value.

**Mafft-homologs** (Collects homologs from SwissProt by BLAST and performs profile-based alignments; Protein only): Help

☐ On

☐ Show homologs (if any)

Number of homologs: 50 (5 – 200)

Threshold: *E*= 1e-10 (1e-5 – 1e-40)

**Plot LAST hits** (DNA only):

◉ The top sequence vs the others ○ The longest sequence vs the others

◉ Plot and alignment ○ Plot only ○ Alignment only

Threshold: score=39 (E=8.4e-11)

Submit Reset

with <200 sequences × <1,000 nucleotides) Help

# MAFFT Results

Jalview

Reformat to GCG, PHYLIP, MSF, NEXUS, uppercase/lowercase, *etc.* with Readseq

Phylogenetic Tree

MAFFT-G-INS-i Result

```
CLUSTAL format alignment by MAFFT (

sp|P69905|HBA_H MVLSPADKTNVKAAWGKVG
sp|P69907|HBA_P MVLSPADKTNVKAAWGKVG
sp|P06635|HBA_P MVLSPADKTNVKTAWGKVG
sp|P01966|HBA_B MVLSAADKGNVKAAWGKVG
sp|P01958|HBA_H MVLSAADKTNVKAAWSKVG
sp|P01959|HBA_E MVLSAADKTNVKAAWSKVG
sp|P01942|HBA_M MVLSGEDKSNIKAAWGKIG
sp|P01946|HBA_R MVLSADDKTNIKNCWGKIG
sp|P01965|HBA_P -VLSAADKANVKAAWGKVG
sp|P60529|HBA_C -VLSPADKTNIKSTWDKIG
                 ***  ** *:*  *.*:*

sp|P69905|HBA_H KKVADALTNAVAHVDDMPN
sp|P69907|HBA_P KKVADALTNAVAHVDDMPN
sp|P06635|HBA_P KKVADALTNAVAHVDDMPN
sp|P01966|HBA_B AKVAAALTKAVEHLDDLPG
sp|P01958|HBA_H KKVGDALTLAVGHLDDLPG
sp|P01959|HBA_E KKVGDALTLAVGHLDDLPG
sp|P01942|HBA_M KKVADALASAAGHLDDLPG
sp|P01946|HBA_R KKVADALAKAADHVEDLPG
sp|P01965|HBA_P QKVADALTKAVGHLDDLPG
sp|P60529|HBA_C KKVADALTTAVAHLDDLPG
                 **. **: *. *::*:*.
```

Submit Reset

**Options**

Output sequence format:

Pretty

- IG|Stanford
- GenBank|gb
- NBRF
- EMBL|em
- GCG
- DNAStrider
- Pearson|Fasta|fa
- Phylip3.2
- Phylip|Phylip4
- Plain|Raw
- PIR|CODATA
- MSF
- PAUP|NEXUS
- Pretty
- XML
- Clustal
- FlatFeat|FFF
- GFF
- ACEDB

Remove gap symbols: -

Calculate checksum of sequences

Select all, or sequences by number:

Translate bases (list as from-base:to-base pairs)

*seq by D.G. Gilbert, 2.1.26 (18-Oct-2007)*

*tp://iubio.bio.indiana.edu/soft/molbio/readseq/java/*

```
 10 142
sp|P69905|    MVLSPADKTN VKAAWGKVGA HAGEYGAEAL ERMFLSFPTT KTYFPHFDLS
sp|P69907|    MVLSPADKTN VKAAWGKVGA HAGEYGAEAL ERMFLSFPTT KTYFPHFDLS
sp|P06635|    MVLSPADKTN VKTAWGKVGA HAGDYGAEAL ERMFLSFPTT KTYFPHFDLS
sp|P01966|    MVLSAADKGN VKAAWGKVGG HAAEYGAEAL ERMFLSFPTT KTYFPHFDLS
sp|P01958|    MVLSAADKTN VKAAWSKVGG HAGEYGAEAL ERMFLGFPTT KTYFPHFDLS
sp|P01959|    MVLSAADKTN VKAAWSKVGG NAGEFGAEAL ERMF
sp|P01942|    MVLSGEDKSN IKAAWGKIGG HGAEYGAEAL ERMF
sp|P01946|    MVLSADDKTN IKNCWGKIGG HGGEYGEEAL QRMF
sp|P01965|    -VLSAADKAN VKAAWGKVGG QAGAHGAEAL ERMF
sp|P60529|    -VLSPADKTN IKSTWDKIGG HAGDYGGEAL DRTF

              HGSAQVKGHG KKVADALTNA VAHVDDMPNA LSAL
              HGSAQVKGHG KKVADALTNA VAHVDDMPNA LSAL
              HGSAQVKDHG KKVADALTNA VAHVDDMPNA LSAL
              HGSAQVKGHG AKVAAALTKA VEHLDDLPGA LSEL
              HGSAQVKAHG KKVGDALTLA VGHLDDLPGA LSNL
              HGSAQVKAHG KKVGDALTLA VGHLDDLPGA LSNL
              HGSAQVKGHG KKVADALASA AGHLDDLPGA LSAL
              PGSAQVKAHG KKVADALAKA ADHVEDLPGA LSTL
              HGSDQVKAHG QKVADALTKA VGHLDDLPGA LSAL
              PGSAQVKAHG KKVADALTTA VAHLDDLPGA LSAL

              LLSHCLLVTL AAHLPAEFTP AVHASLDKFL ASVS
              LLSHCLLVTL AAHLPAEFTP AVHASLDKFL ASVS
              LLSHCLLVTL AAHLPAEFTP AVHASLDKFL ASVSTVLTSK YR
              LLSHSLLVTL ASHLPSDFTP AVHASLDKFL ANVSTVLTSK YR
              LLSHCLLSTL AVHLPNDFTP AVHASLDKFL SSVSTVLTSK YR
              LLSHCLLSTL AVHLPNDFTP AVHASLDKFL STVSTVLTSK YR
              LLSHCLLVTL ASHHPADFTP AVHASLDKFL ASVSTVLTSK YR
              FLSHCLLVTL ACHHPGDFTP AMHASLDKFL ASVSTVLTSK YR
              LLSHCLLVTL AAHHPDDFNP SVHASLDKFL ANVSTVLTSK YR
              LLSHCLLVTL ACHHPTEFTP AVHASLDKFF AAVSTVLTSK YR
```

Submit Reset

**Options**

Output sequence format:
Phylip|Phylip4

Return biosequence data:
- Download to file
- View in browser

Change sequence case to
- No change
- lower
- UPPER

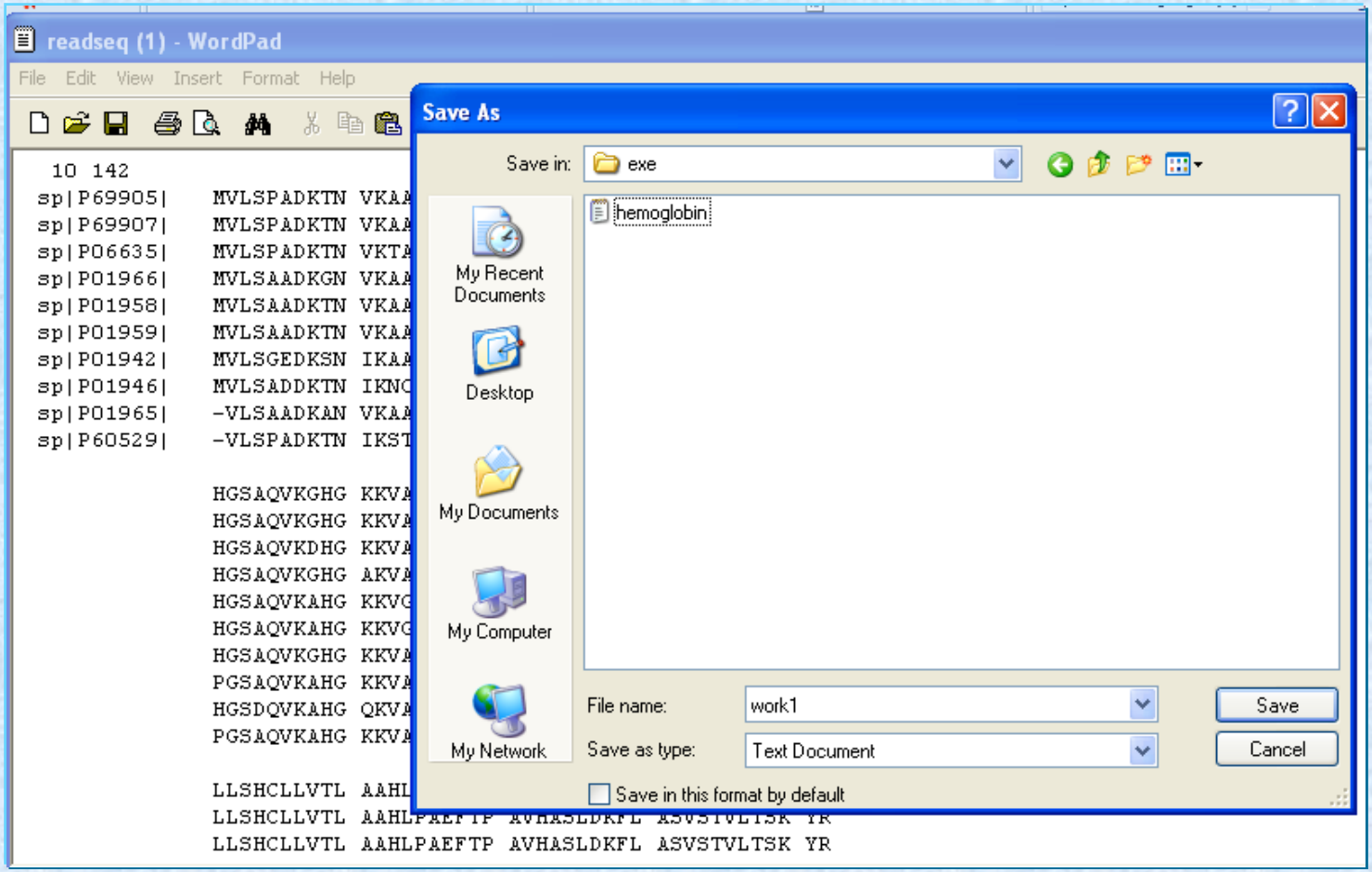
- Remove gap symbols: -
- Calculate checksum of sequences

Select all, or sequences by number:

- Translate bases (list as from-base:to-base pairs)

**Saved in a temporary file "readseq".**

**Open it and save as "work1"**




**Folder: Phylip-3.69/exe/work1**

# Procedure to run Phylip

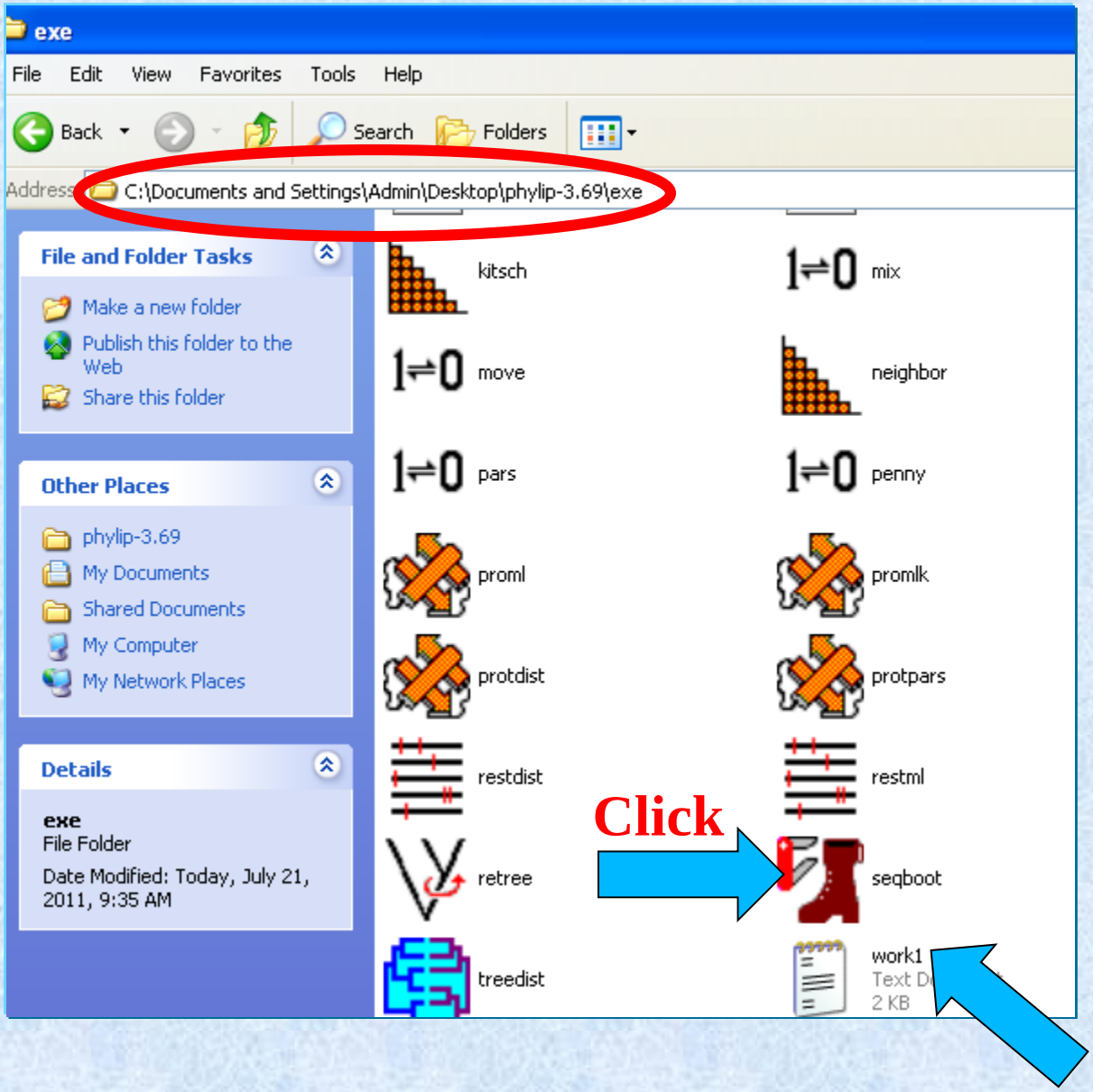
1. **Bootstrapping: to check the confidence level**

**In statistics, bootstrapping is a computer-based method for assigning measures of accuracy to sample estimates.**

**Bootstrapping is the practice of estimating properties of an estimator (such as its variance) by measuring those properties when sampling from an approximating distribution.**

**One standard choice for an approximating distribution is the empirical distribution of the observed data.**

**This can be implemented by constructing a number of resamples of the observed dataset, each of which is obtained by random sampling with replacement from the original dataset.**


exe
File Edit View Favorites Tools Help
Back
Search
Folders
Address
C:\Documents and Settings\Admin\Desktop\phylip-3.69\exe
File and Folder Tasks
Make a new folder
Publish this folder to the Web
Share this folder
Other Places
phylip-3.69
My Documents
Shared Documents
My Computer
My Network Places
Details
exe
File Folder
Date Modified: Today, July 21, 2011, 9:35 AM
kitsch
mix
move
neighbor
pars
penny
proml
promlk
protdist
protpars
restdist
restml
retree
seqboot
treedist
work1
2 KB
Click

```
C:\Documents and Settings\Admin\Desktop\phylip-3.69\exe\seqboot.exe
seqboot.exe: can't find input file "infile"
Please enter a new file name> work1.txt
```

```
Bootstrapping algorithm, version 3.69

Settings for this run:
  D       Sequence, Morph, Rest., Gene Freqs?  Molecular sequences
  J   Bootstrap, Jackknife, Permute, Rewrite?  Bootstrap
  %     Regular or altered sampling fraction?  regular
  B       Block size for block-bootstrapping?  1 (regular bootstrap)
  R                    How many replicates?    100
  W               Read weights of characters?  No
  C                 Read categories of sites?  No
  S      Write out data sets or just weights?  Data sets
  I              Input sequences interleaved?  Yes
  0       Terminal type (IBM PC, ANSI, none)?  IBM PC
  1        Print out the data at start of run  No
  2      Print indications of progress of run  Yes

  Y to accept these or type the letter for one to change
```

```
completed replicate number    1
completed replicate number    2
completed replicate number    3
completed replicate number    4
completed replicate number    5
completed replicate number    6
completed replicate number    7
completed replicate number    8
completed replicate number    9
completed replicate number   10

Output written to file "outfile"

Done.

Press enter to quit.
```

```
Bootstrapping algorithm, version 3.69

Settings for this run:
  D       Sequence, Morph, Rest., Gene Freqs?  Molecular s
  J   Bootstrap, Jackknife, Permute, Rewrite?  Bootstrap
  %     Regular or altered sampling fraction?  regular
  B       Block size for block-bootstrapping?  1 (regular bootstrap)
  R                    How many replicates?    10
  W               Read weights of characters?  No
  C                 Read categories of sites?  No
  S      Write out data sets or just weights?  Data sets
  I              Input sequences interleaved?  Yes
  0       Terminal type (IBM PC, ANSI, none)?  IBM PC
  1        Print out the data at start of run  No
  2      Print indications of progress of run  Yes

  Y to accept these or type the letter for one to change
R
Number of replicates?
10
```

**outfile** contains 10 replications

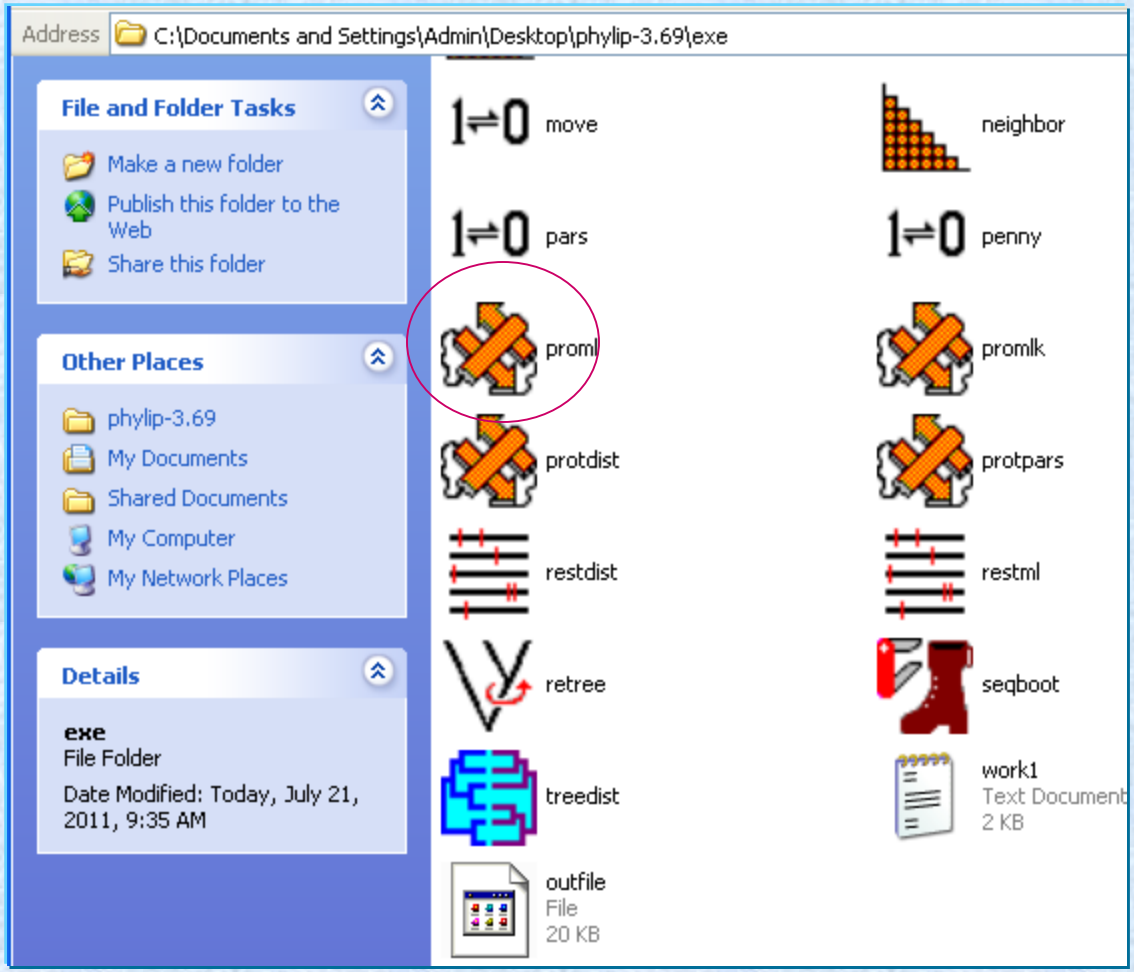
```
  Y to accept these or type the letter for one to change
y

Random number seed (must be odd)?
5
```

```
   10   142
sp|P69905| MLPPADKKTT TVVAGKGAHH HGGEGAELRM MMMFFSTTKK KTYFFFFDDS SHHGGAAKGK
sp|P69907| MLPPADKKTT TVVAGKGAHH HGGEGAELRM MMMFFSTTKK KTYFFFFDDS SHHGGAAKGK
sp|P06635| MLPPADKKTT TVVTGKGAHH HGGDGAELRM MMMFFSTTKK KTYFFFFDDS SHHGGAAKGK
sp|P01966| MLAAADKKGG GVVAGKGGHH HAAEGAELRM MMMFFSTTKK KTYFFFFDDS SHHGGAAKGA
sp|P01958| MLAAADKKTT TVVASKGGHH HGGEGAELRM MMMFFGTTKK KTYFFFFDDS SHHGGAAKGK
sp|P01959| MLAAADKKTT TVVASKGGNN NGGEGAELRM MMMFFGTTKK KTYFFFFDDS SHHGGAAKGK
sp|P01942| MLGGEDKKSS SIIAGKGGHH HAAEGAELRM MMMFFSTTKK KTYFFFFDDS SHHGGAAKGK
sp|P01946| MLAADDKKTT TIINGKGGHH HGGEGEELRM MMMFFATTKK KTYIIIIDDS SPPGGAAKGK
sp|P01965| -LAAADKKAA AVVAGKGGQQ QGGAGAELRM MMMFFGTTKK KTYFFFFNNS SHHGGDDKGQ
sp|P60529| -LPPADKKTT TIISDKGGHH HGGDGGELRT TTTFFSTTKK KTYFFFFDDS SPPGGAAKGK

           KKKVADDDLA AAADPSALLH HAHHRRDDNN KLLLHCCCCV TTTLLAAAHL LAAAEFFTHA
           KKKVADDDLA AAADPSALLH HAHHRRDDNN KLLLHCCCCV TTTLLAAAHL LAAAEFFTHA
           KKKVADDDLA AAADPSALLH HAHHRRDDNN KLLLHCCCCV TTTLLAAAHL LAAAEFFTHA
           AKKVAAAALA AAEDPSELLH HAHHRRDDNN KLLLHSSSSV TTTLLAASHL LSSSDFFTHA
           KKKVGDDDLA AAGDPSNLLH HAHHRRDDNN KLLLHCCCCS TTTLLAAVHL LNNNDFFTHA
           KKKVGDDDLA AAGDPSNLLH HAHHRRDDNN KLLLHCCCCS TTTLLAAVHL LNNNDFFTHA
           KKKVADDDLA AAGDPSALLH HAHHRRDDNN KLLLHCCCCV TTTLLAASHH HAAADFFTHA
           KKKVADDDLA AADEPSTLLH HAHHRRDDNN KFFFHCCCCV TTTLLAACHH HGGGDFFTHA
           QKKVADDDLA AAGDPSALLH HAHHRRDDNN KLLLHCCCCV TTTLLAAAHH HDDDDFFNHA
           KKKVADDDLA AAADPSALLH HAYYRRDDNN KLLLHCCCCV TTTLLAACHH HTTTEFFTHA

           ASSLLFFVVS STLLSKYYYR RR
           ASSLLFFVVS STLLSKYYYR RR
           ASSLLFFVVS STLLSKYYYR RR
           ASSLLFFVVS STLLSKYYYR RR
           ASSLLFFVVS STLLSKYYYR RR
           ASSLLFFVVS STLLSKYYYR RR
           ASSLLFFVVS STLLSKYYYR RR
           ASSLLFFVVS STLLSKYYYR RR
           ASSLLFFVVS STLLSKYYYR RR
           ASSLLFFVVS STLLSKYYYR RR

   10   142
sp|P69905| MLAAAWVGGA AGEYYGGAAL EEERFFLLLF FYFFFFPPHD DDLLHAAQQQ QVVVKGGKKA
sp|P69907| MLAAAWVGGA AGEYYGGAAL EEERFFLLLF FYFFFFPPHD DDLLHAAQQQ QVVVKGGKKA
sp|P06635| MLAATWVGGA AGDYYGGAAL EEERFFLLLF FYFFFFPPHD DDLLHAAQQQ QVVVKDGKKA
sp|P01966| MLAAAWVGGA AAEYYGGAAL EEERFFLLLF FYFFFFPPHD DDLLHAAQQQ QVVVKGGKKA
sp|P01958| MLAAAWVGGA AGEYYGGAAL EEERFFLLLF FYFFFFPPHD DDLLHAAQQQ QVVVKAGKKG
```

**outfile**
**10 different sets**

# Phylogenetic tree using Maximum likelihood method

**The program is proml**



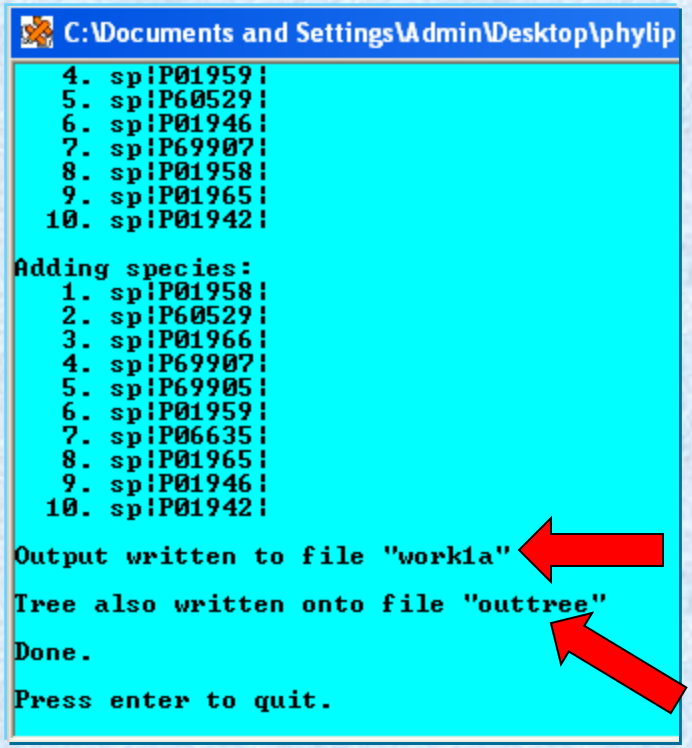

**outfile obtained from seqboot is the input for proml**

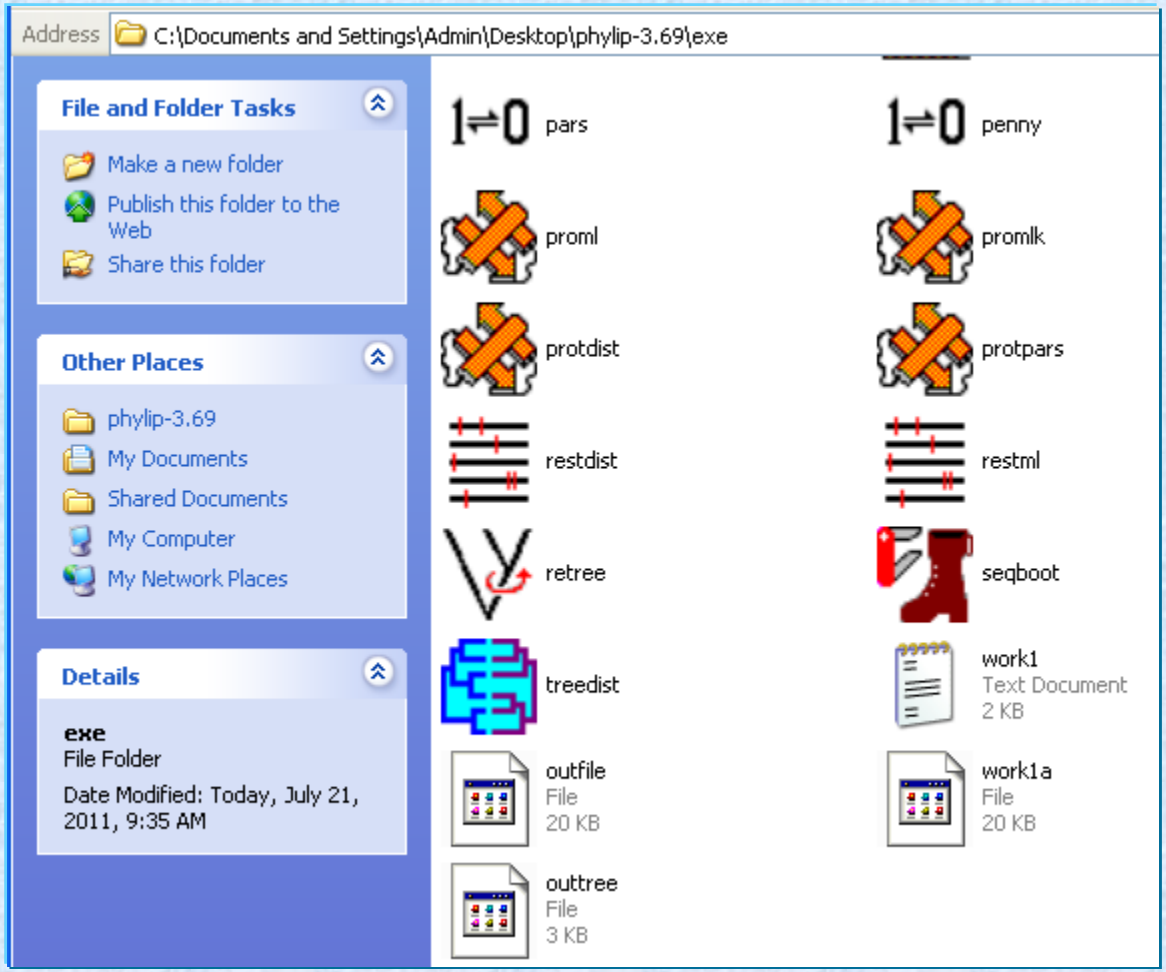
```
C:\Documents and Settings\Admin\Desktop\phylip-3.69\exe\proml.exe
proml.exe: can't find input file "infil
Please enter a new file name> outfile

proml.exe: the file "outfile" that you wanted to
     use as output file already exists.
     Do you want to Replace it, Append to it,
     write to a new File, or Quit?
     (please type R, A, F, or Q)
F
Please enter a new file name> work1a
```

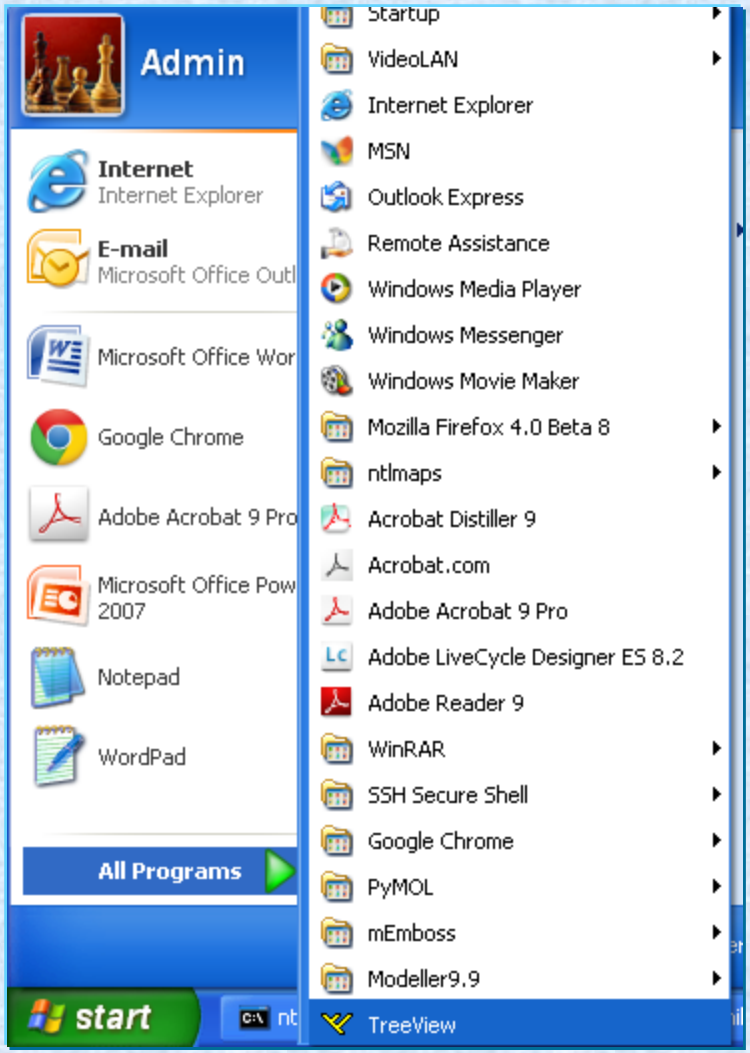
```
Amino acid sequence Maximum Likelihood method, version 3.69

Settings for this run:
  U                 Search for best tree?  Yes
  P    JTT, PMB or PAM probability model?  Jones-Taylor-Thornton
  C               One category of sites?  Yes
  R           Rate variation among sites?  constant rate of change
  W                       Sites weighted?  No
  S        Speedier but rougher analysis?  Yes
  G               Global rearrangements?  No
  J   Randomize input order of sequences?  No. Use input order
  O                        Outgroup root?  No, use as outgroup species  1
  M           Analyze multiple data sets?  No
  I          Input sequences interleaved?  Yes
  0   Terminal type (IBM PC, ANSI, none)?  IBM PC
  1    Print out the data at start of run  No
  2  Print indications of progress of run  Yes
  3                        Print out tree  Yes
  4       Write out trees onto tree file?  Yes
  5   Reconstruct hypothetical sequences?  No

  Y to accept these or type the letter for one to change
m
Multiple data sets or multiple weights? (type D or W)
d
How many data sets?
10
```

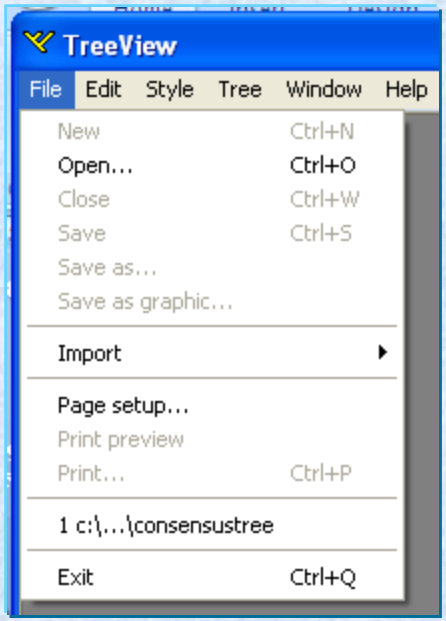
```
Random number seed (must be odd)?
5
Number of times to jumble?
3
```

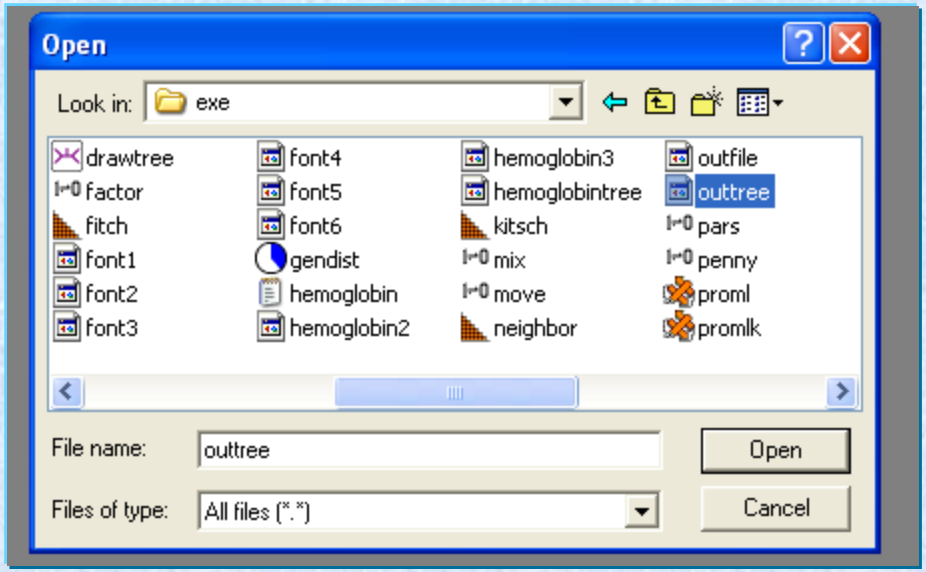
```
  Y to accept these or type the letter for one to change
Y
```

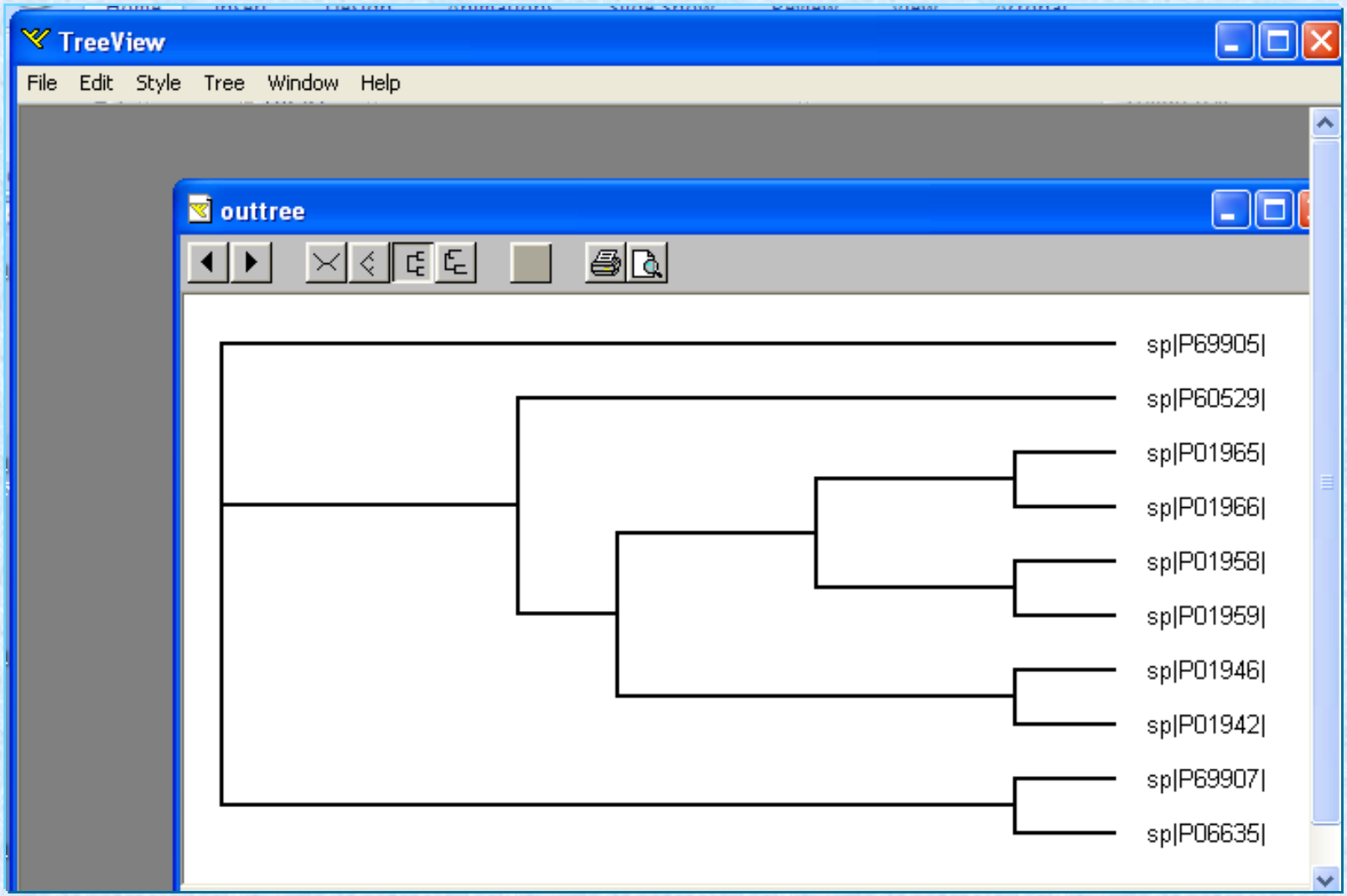

C:\Documents and Settings\Admin\Desktop\phylip
4. sp¦P01959¦
5. sp¦P60529¦
6. sp¦P01946¦
7. sp¦P69907¦
8. sp¦P01958¦
9. sp¦P01965¦
10. sp¦P01942¦
Adding species:
1. sp¦P01958¦
2. sp¦P60529¦
3. sp¦P01966¦
4. sp¦P69907¦
5. sp¦P69905¦
6. sp¦P01959¦
7. sp¦P06635¦
8. sp¦P01965¦
9. sp¦P01946¦
10. sp¦P01942¦
Output written to file "work1a"
Tree also written onto file "outtree"
Done.
Press enter to quit.

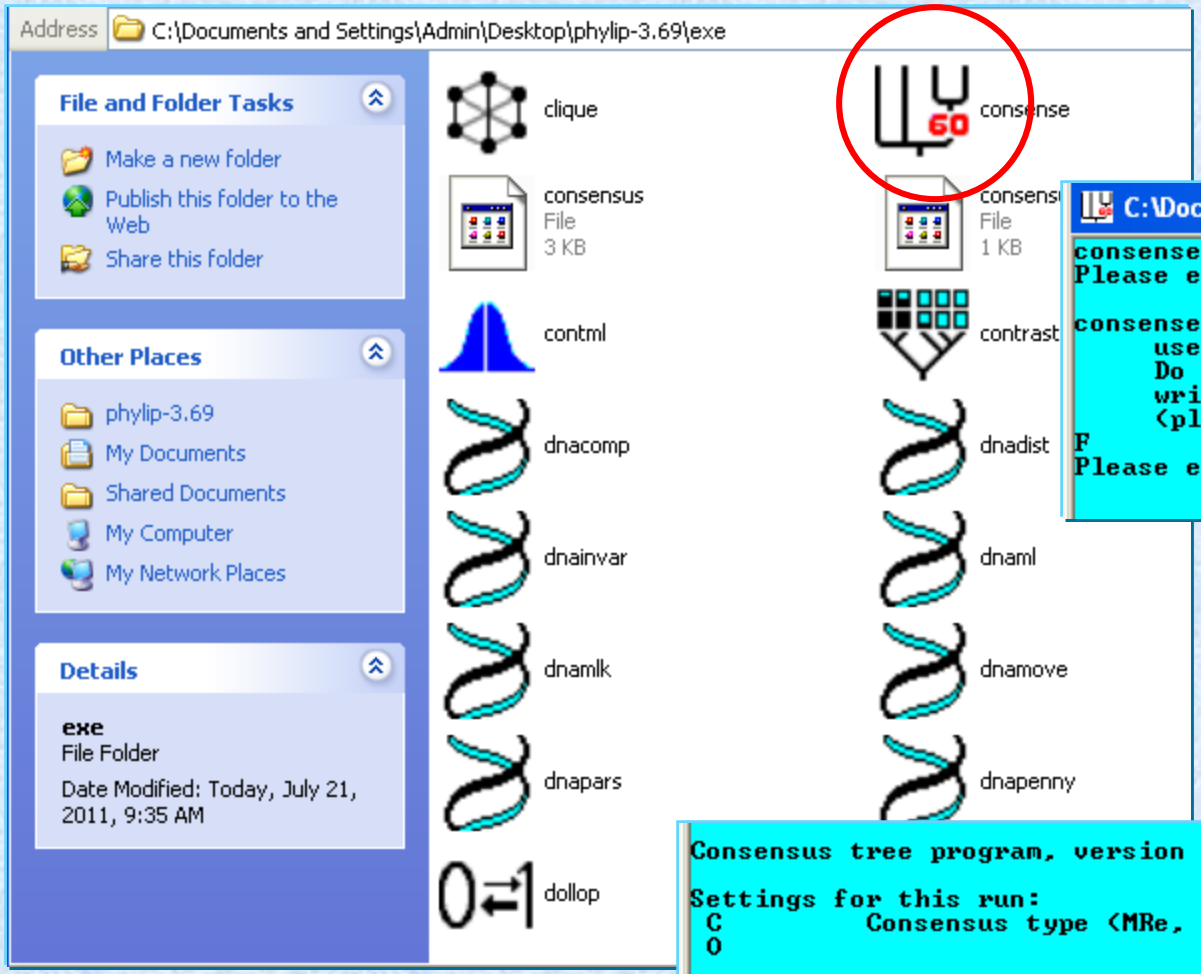

Address
C:\Documents and Settings\Admin\Desktop\phylip-3.69\exe
File and Folder Tasks
Make a new folder
Publish this folder to the Web
Share this folder
Other Places
phylip-3.69
My Documents
Shared Documents
My Computer
My Network Places
Details
exe
File Folder
Date Modified: Today, July 21, 2011, 9:35 AM
pars
penny
proml
promlk
protdist
protpars
restdist
restml
retree
seqboot
treedist
work1
Text Document
2 KB
outfile
File
20 KB
work1a
File
20 KB
outtree
File
3 KB


Admin
Internet
Internet Explorer
E-mail
Microsoft Office Outl
Microsoft Office Wor
Google Chrome
Adobe Acrobat 9 Pro
Microsoft Office Pow 2007
Notepad
WordPad
All Programs
start
Startup
VideoLAN
Internet Explorer
MSN
Outlook Express
Remote Assistance
Windows Media Player
Windows Messenger
Windows Movie Maker
Mozilla Firefox 4.0 Beta 8
ntlmaps
Acrobat Distiller 9
Acrobat.com
Adobe Acrobat 9 Pro
Adobe LiveCycle Designer ES 8.2
Adobe Reader 9
WinRAR
SSH Secure Shell
Google Chrome
PyMOL
mEmboss
Modeller9.9
TreeView


TreeView
File Edit Style Tree Window Help
New Ctrl+N
Open... Ctrl+O
Close Ctrl+W
Save Ctrl+S
Save as...
Save as graphic...
Import
Page setup...
Print preview
Print... Ctrl+P
1 c:\...\consensustree
Exit Ctrl+Q


Open
Look in: exe
drawtree
factor
fitch
font1
font2
font3
font4
font5
font6
gendist
hemoglobin
hemoglobin2
hemoglobin3
hemoglobintree
kitsch
mix
move
neighbor
outfile
outtree
pars
penny
proml
promlk
File name: outtree
Open
Files of type: All files (*.*)
Cancel


TreeView
File
Edit
Style
Tree
Window
Help
outtree
sp|P69905|
sp|P60529|
sp|P01965|
sp|P01966|
sp|P01958|
sp|P01959|
sp|P01946|
sp|P01942|
sp|P69907|
sp|P06635|

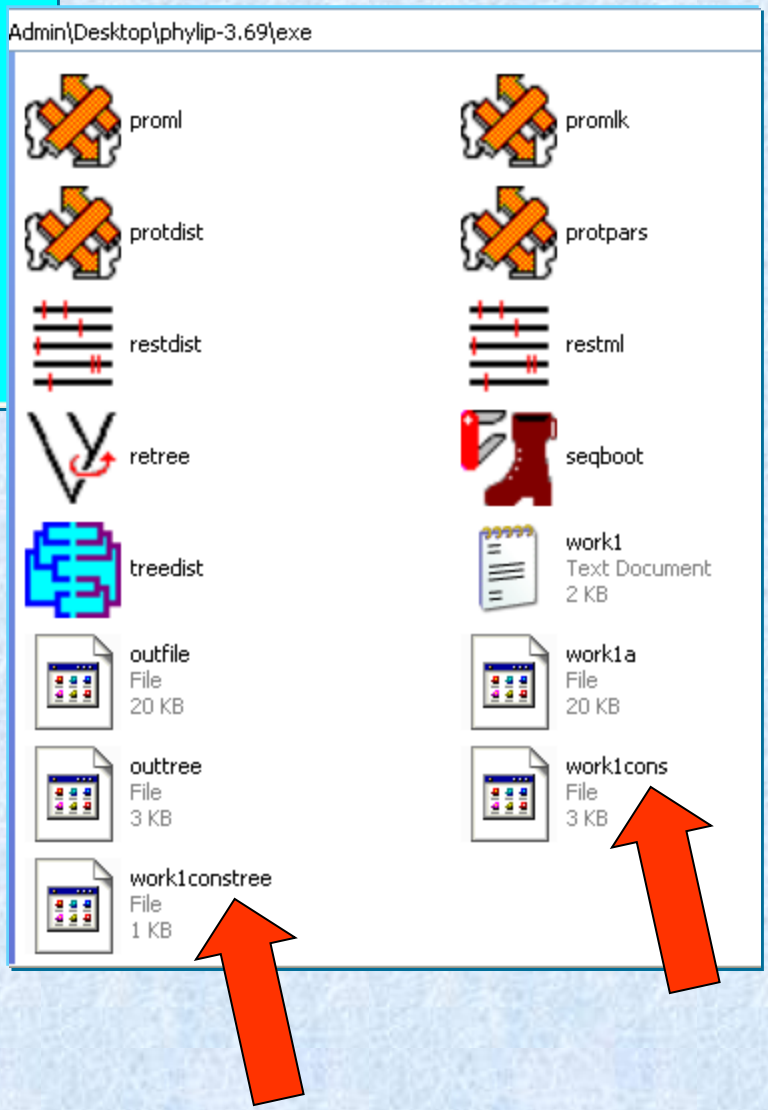


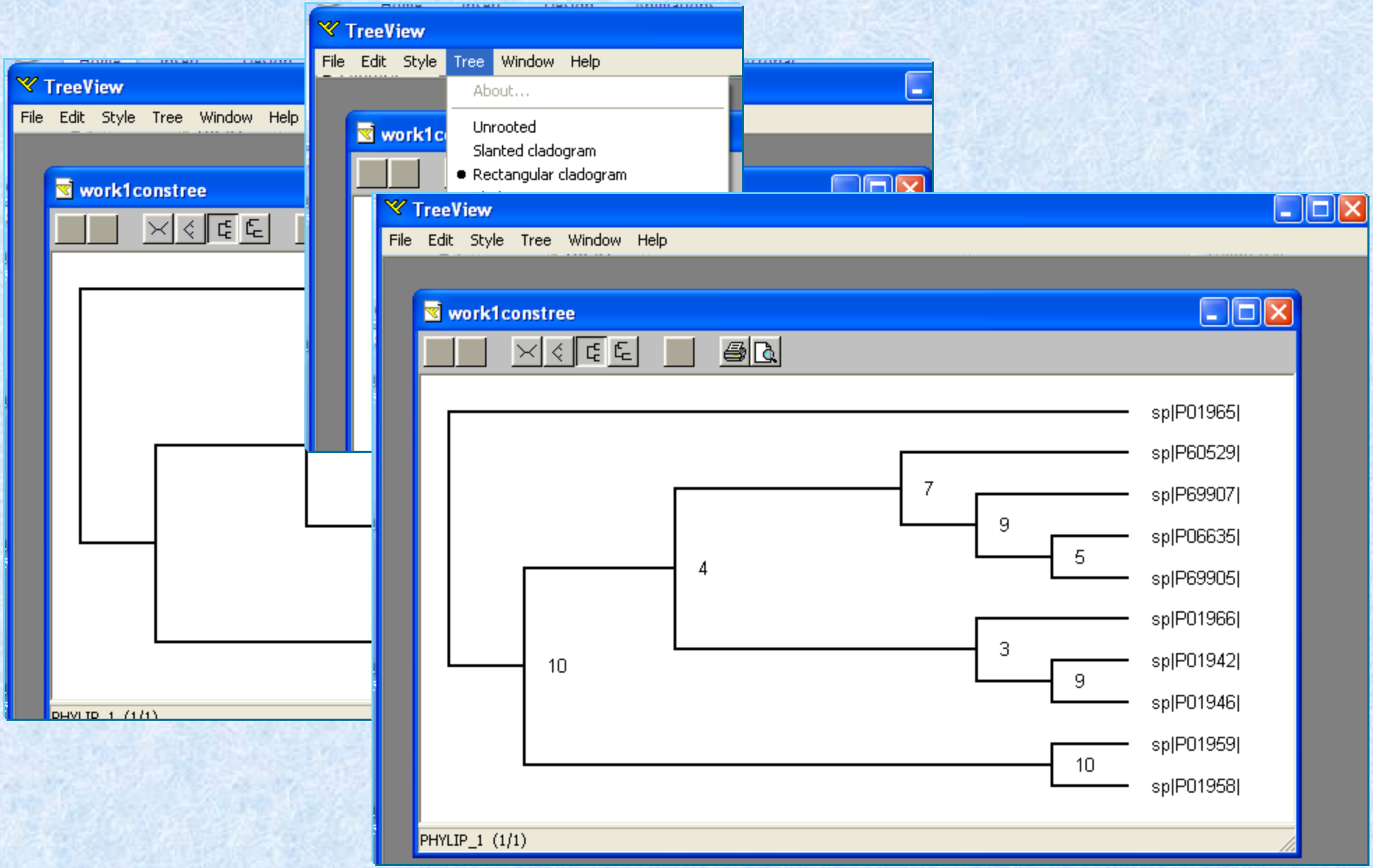

```
C:\Documents and Settings\Admin\Desktop\phylip-3.69\exe\consei
consense.exe: can't find input tree file "intree"
Please enter a new file name> outtree

consense.exe: the file "outfile" that you wanted to
     use as output file already exists.
     Do you want to Replace it, Append to it,
     write to a new File, or Quit?
     (please type R, A, F, or Q)
F
Please enter a new file name> work1cons
```

```
Consensus tree program, version 3.69

Settings for this run:
 C          Consensus type (MRe, strict, MR, Ml):  Majority rule (extended)
 O                                Outgroup root:  No, use as outgroup species  1

 R                 Trees to be treated as Rooted:  No
 T             Terminal type (IBM PC, ANSI, none):  IBM PC
 1                  Print out the sets of species:  Yes
 2          Print indications of progress of run:  Yes
 3                                  Print out tree:  Yes
 4                Write out trees onto tree file:  Yes

Are these settings correct? (type Y or the letter for one to change)
y_
```

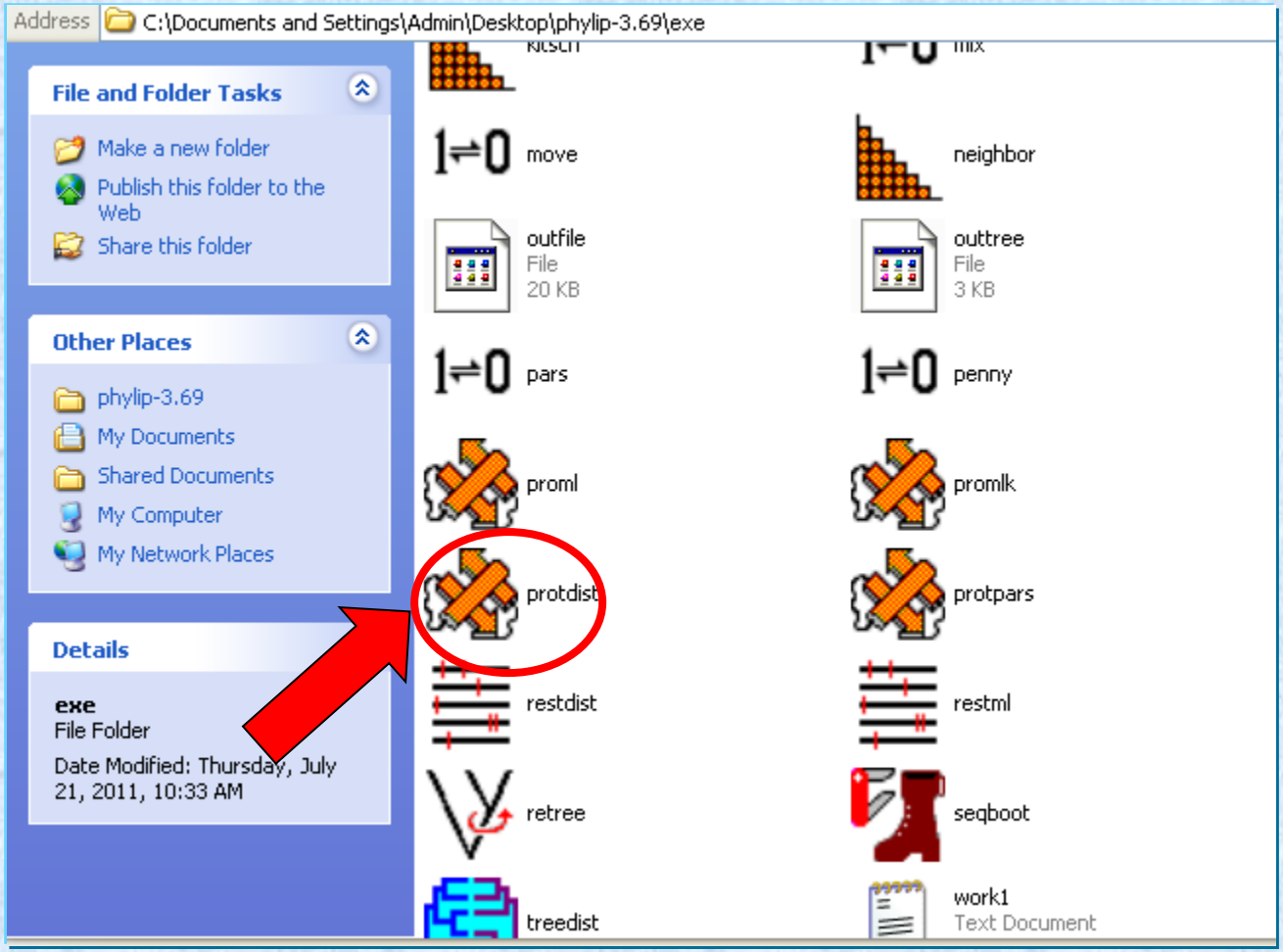
```
Consensus tree program, version 3.69

Settings for this run:
 C         Consensus type (MRe, strict, MR, Ml):  Majority rule (extended)
 O                              Outgroup root:  No, use as outgroup species  1

 R               Trees to be treated as Rooted:  No
 T           Terminal type (IBM PC, ANSI, none):  IBM PC
 1               Print out the sets of species:  Yes
 2         Print indications of progress of run:  Yes
 3                              Print out tree:  Yes
 4                Write out trees onto tree file:  Yes

Are these settings correct? (type Y or the letter for one to change)
y

consense.exe: the file "outtree" that you wanted to
     use as output tree file already exists.
     Do you want to Replace it, Append to it,
     write to a new File, or Quit?
     (please type R, A, F, or Q)
f
Please enter a new file name> work1constree_
```

```
Consensus tree written to file "work1constree"

Output written to file "work1cons"

Done.

Press enter to quit.
```




TreeView
File Edit Style Tree Window Help
About...
Unrooted
Slanted cladogram
Rectangular cladogram
work1constree
sp|P01965|
sp|P60529|
sp|P69907|
sp|P06635|
sp|P69905|
sp|P01966|
sp|P01942|
sp|P01946|
sp|P01959|
sp|P01958|
7
9
5
4
3
9
10
10
PHYLIP_1 (1/1)

# NJ and UPGMA methods



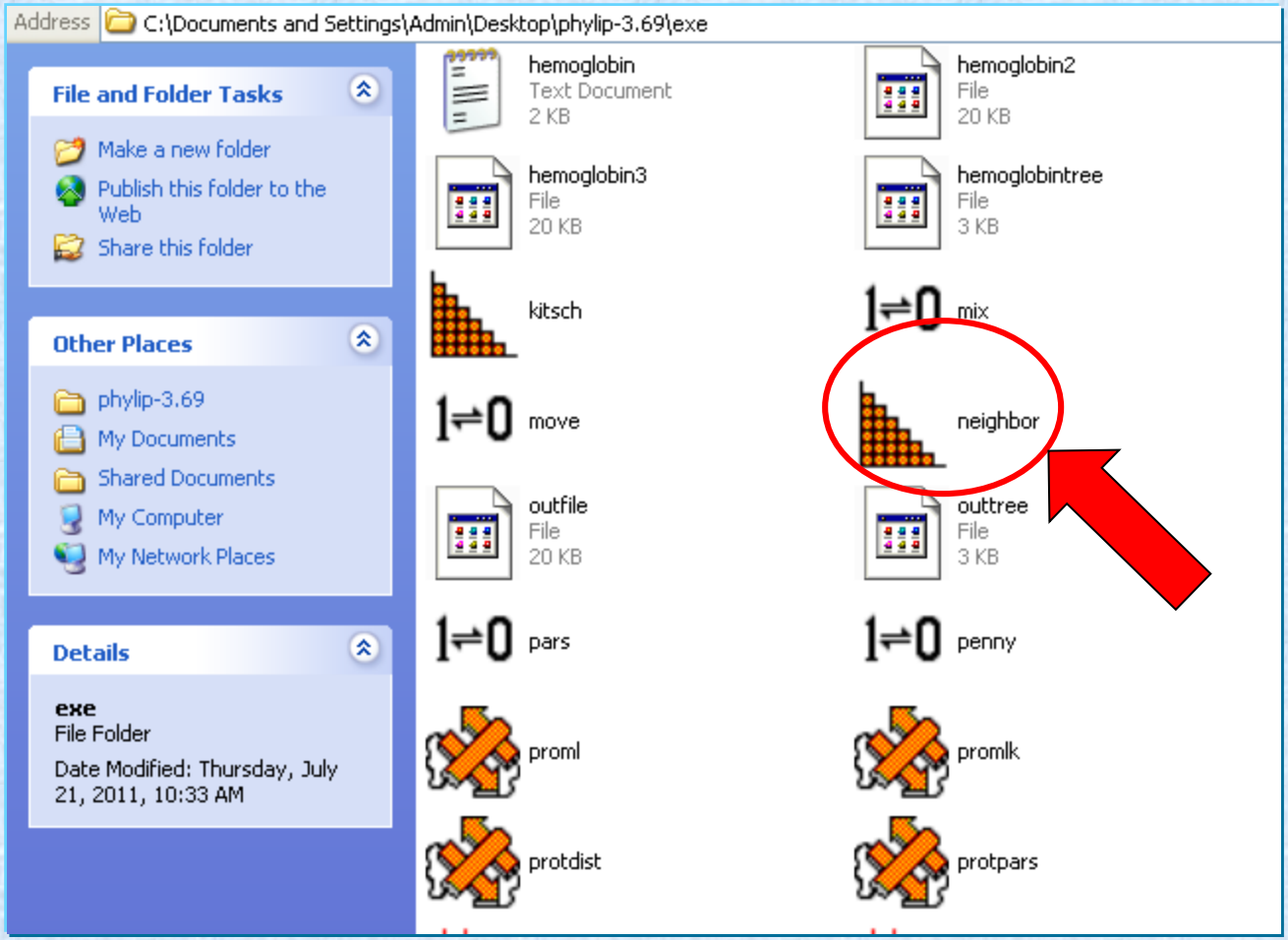
```
protdist.exe: can't find input file "infile"
Please enter a new file name> outfile

protdist.exe: the file "outfile" that you wanted to
     use as output file already exists.
     Do you want to Replace it, Append to it,
     write to a new File, or Quit?
     (please type R, A, F, or Q)
f
Please enter a new file name> work1-protdist
```

```
Protein distance algorithm, version 3.69

Settings for this run:
  P  Use JTT, PMB, PAM, Kimura, categories model?  Jones-Taylor-Thornton matrix
  G  Gamma distribution of rates among positions?  No
  C           One category of substitution rates?  Yes
  W                    Use weights for positions?  No
  M                   Analyze multiple data sets?  No
  I                  Input sequences interleaved?  Yes
  0                Terminal type (IBM PC, ANSI)?  IBM PC
  1            Print out the data at start of run  No
  2           Print indications of progress of run  Yes

Are these settings correct? (type Y or the letter for one to change)
m
Multiple data sets or multiple weights? (type D or W)
d
How many data sets?
10
```

```
Protein distance algorithm, version 3.69

Settings for this run:
  P  Use JTT, PMB, PAM, Kimura, categories model?  Dayhoff PAM matrix
  G  Gamma distribution of rates among positions?  No
  C           One category of substitution rates?  Yes
  W                    Use weights for positions?  No
  M                   Analyze multiple data sets?  Yes, 10 data sets
  I                  Input sequences interleaved?  Yes
  0                Terminal type (IBM PC, ANSI)?  IBM PC
  1            Print out the data at start of run  No
  2           Print indications of progress of run  Yes

Are these settings correct? (type Y or the letter for one to change)
```

# work1-protdist

```
sp|Q9WVA2|  0.000000  0.954472  1.211022  1.534523  2.384227
2.197177
  2.466581  2.303897  2.378322  2.245130  2.098663  2.066428
1.923666
  1.946335
sp|Q9Y5J9|  0.954472  0.000000  1.114781  1.699316  2.007548
2.014938
  2.236423  1.873154  2.401207  2.523799  2.217026  2.237688
1.818816
  1.896587
sp|Q09783|  1.211022  1.114781  0.000000  1.303653  2.771608
2.657542
  2.794046  2.191316  2.722125  2.545494  2.061901  2.150728
2.356218
  2.009404
sp|Q75DU7|  1.534523  1.699316  1.303653  0.000000  2.572591
2.581259
  2.958788  2.454755  2.303727  2.134035  2.242738  1.951797
1.613806
  2.068258
sp|Q75F72|  2.384227  2.007548  2.771608  2.572591  0.000000
0.263512
  0.532918  0.949154  1.037593  1.164635  1.410843  1.465146
1.892952
  2.173354
sp|Q6CJX3|  2.197177  2.014938  2.657542  2.581259  0.263512
0.000000
  0.454437  0.916289  1.004302  1.030748  1.405926  1.425100
1.939355
```

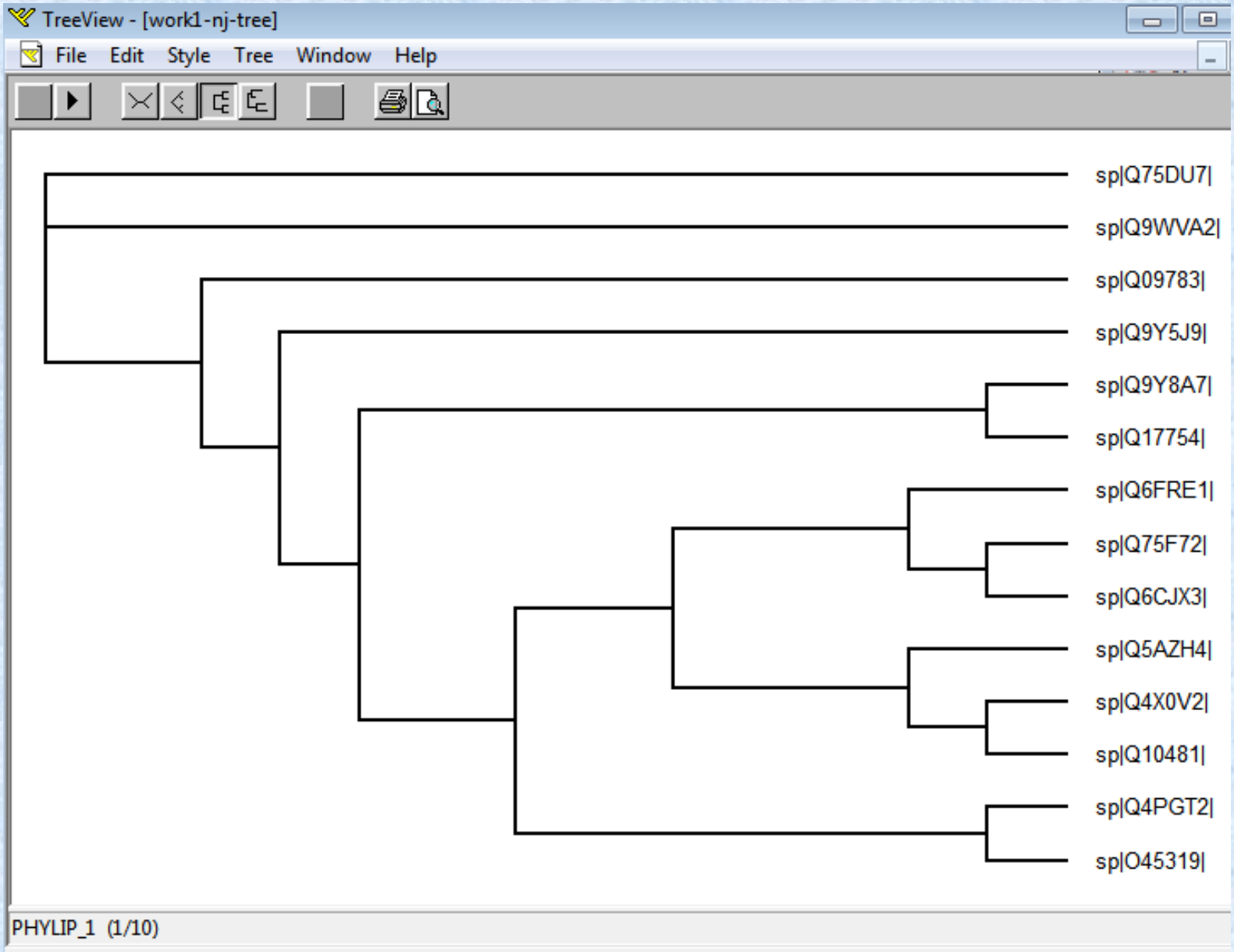

Address
C:\Documents and Settings\Admin\Desktop\phylip-3.69\exe
File and Folder Tasks
Make a new folder
Publish this folder to the Web
Share this folder
Other Places
phylip-3.69
My Documents
Shared Documents
My Computer
My Network Places
Details
exe
File Folder
Date Modified: Thursday, July 21, 2011, 10:33 AM
hemoglobin
Text Document
2 KB
hemoglobin2
File
20 KB
hemoglobin3
File
20 KB
hemoglobintree
File
3 KB
kitsch
mix
move
neighbor
outfile
File
20 KB
outtree
File
3 KB
pars
penny
proml
promlk
protdist
protpars

```
neighbor.exe: can't find input file "infile"
Please enter a new file name> work1-protdist

neighbor.exe: the file "outfile" that you wanted to
     use as output file already exists.
     Do you want to Replace it, Append to it,
     write to a new File, or Quit?
     (please type R, A, F, or Q)
F
Please enter a new file name> work1-nj
```

```
Neighbor-Joining/UPGMA method version 3.69

Settings for this run:
  N       Neighbor-joining or UPGMA tree?  Neighbor-joining
  O                        Outgroup root?  No, use as outgroup species  1
  L         Lower-triangular data matrix?  No
  R         Upper-triangular data matrix?  No
  S                       Subreplicates?  No
  J     Randomize input order of species?  No. Use input order
  M           Analyze multiple data sets?  No
  0   Terminal type (IBM PC, ANSI, none)?  IBM PC
  1    Print out the data at start of run  No
  2  Print indications of progress of run  Yes
  3                        Print out tree  Yes
  4       Write out trees onto tree file?  Yes


  Y to accept these or type the letter for one to change
m
How many data sets?
10

Random number seed (must be odd)?
5
```

```
Settings for this run:
  N       Neighbor-joining or UPGMA tree?  Neighbor-joining
  O                        Outgroup root?  No, use as outgroup species  1
  L         Lower-triangular data matrix?  No
  R         Upper-triangular data matrix?  No
  S                       Subreplicates?  No
  J     Randomize input order of species?  Yes (random number seed =       5)
  M           Analyze multiple data sets?  Yes, 10 sets
  0   Terminal type (IBM PC, ANSI, none)?  IBM PC
  1    Print out the data at start of run  No
  2  Print indications of progress of run  Yes
  3                        Print out tree  Yes
  4       Write out trees onto tree file?  Yes


  Y to accept these or type the letter for one to change
y

neighbor.exe: the file "outtree" that you wanted to
     use as output tree file already exists.
     Do you want to Replace it, Append to it,
     write to a new File, or Quit?
     (please type R, A, F, or Q)
f
Please enter a new file name> work1-nj-tree
```

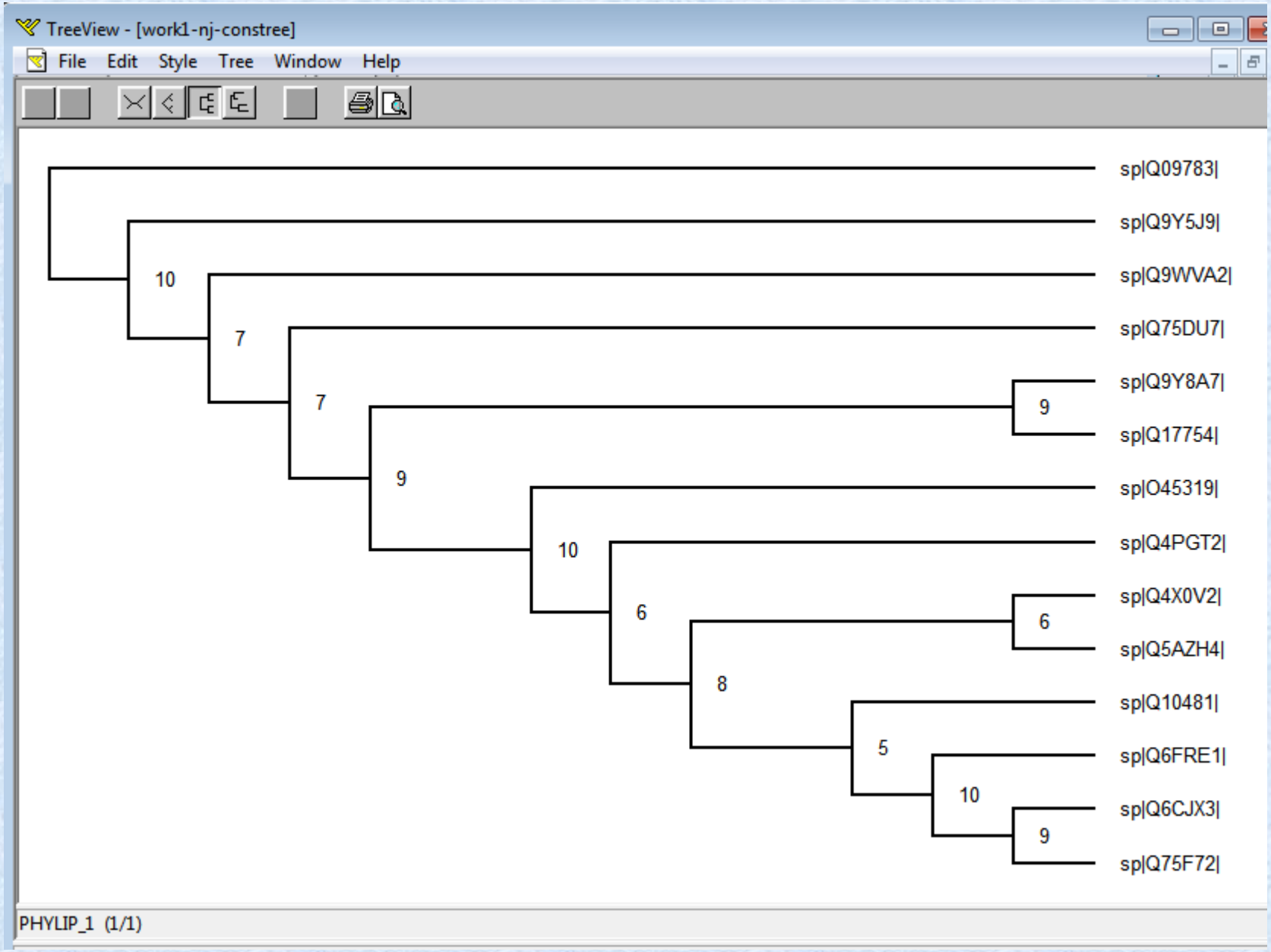
```
((sp|Q09783|:0.61703,(sp|Q9Y5J9|:0.71119,((sp|Q9Y8A7|:0.96547,
sp|Q17754|:0.59919):0.21205,(((sp|Q6FRE1|:0.32455,(sp|Q75F72
|:0.10986,
sp|Q6CJX3|:0.11552):0.07268):0.29205,((sp|Q4X0V2|:0.40667,
sp|Q10481|:0.30864):0.10478,sp|Q5AZH4
|:0.45459):0.12561):0.17539,
(sp|Q4PGT2|:0.67722,sp|O45319
|:0.67765):0.02625):0.50607):0.47300):0.27258):0.09925,
sp|Q75DU7|:0.56167,sp|Q9WVA2|:0.64156);
((sp|Q09783|:0.68210,sp|Q9Y5J9|:0.38881):0.03306,(((sp|Q9Y8A7
|:0.64578,
sp|Q17754|:0.49015):0.06214,((sp|O45319|:0.50317,sp|Q4PGT2
|:0.42263):0.12931,
(sp|Q5AZH4|:0.36428,(sp|Q4X0V2|:0.39579,(((sp|Q75F72|:0.12829,
sp|Q6CJX3|:0.06919):0.12728,sp|Q6FRE1|:0.29582):0.35209,
sp|Q10481|:0.38144):0.03146):0.07114):0.13883):0.70546):0.52085,
sp|Q75DU7|:0.77374):0.34769,sp|Q9WVA2|:0.46575);
((sp|Q9Y5J9|:0.30392,sp|Q09783|:0.67455):0.07237,((sp|Q17754
|:0.56886,
((sp|O45319|:0.52738,(sp|Q4PGT2|:0.49822,(sp|Q10481|:0.35123,
(((sp|Q6CJX3|:0.08839,sp|Q75F72|:0.12819):0.10053,sp|Q6FRE1
|:0.26834):0.28216,
(sp|Q5AZH4|:0.33943,sp|Q4X0V2
|:0.30149):0.12912):0.02663):0.11009):0.03799):0.40163,
sp|Q75DU7|:0.96568):0.13386):0.11587,sp|Q9Y8A7
|:0.62374):0.33497,sp|Q9WVA2|:0.37622);
((sp|Q75DU7|:0.81103,((sp|O45319|:0.60856,((((sp|Q6FRE1
|:0.36825,
sp|Q6CJX3|:0.11354):0.02000,sp|Q75F72|:0.15488):0.34593,
sp|Q10481|:0.29363):0.02586,((sp|Q4PGT2|:0.56535,sp|Q5AZH4
|:0.38322):0.10438,
sp|Q4X0V2|:0.40429):0.07203):0.20545):0.71213,(sp|Q9Y8A7
```


TreeView - [work1-nj-tree]
File Edit Style Tree Window Help
sp|Q75DU7|
sp|Q9WVA2|
sp|Q09783|
sp|Q9Y5J9|
sp|Q9Y8A7|
sp|Q17754|
sp|Q6FRE1|
sp|Q75F72|
sp|Q6CJX3|
sp|Q5AZH4|
sp|Q4X0V2|
sp|Q10481|
sp|Q4PGT2|
sp|O45319|
PHYLIP_1 (1/10)

```
consense.exe: can't find input tree file "intree"
Please enter a new file name> work1-nj-tree

consense.exe: the file "outfile" that you wanted to
     use as output file already exists.
     Do you want to Replace it, Append to it,
     write to a new File, or Quit?
     (please type R, A, F, or Q)
f
Please enter a new file name> work1-nj-cons
```

```
Consensus tree program, version 3.695

Settings for this run:
 C         Consensus type (MRe, strict, MR, Ml):  Majority rule (extended)
 O                                Outgroup root:  No, use as outgroup species

 R               Trees to be treated as Rooted:  No
 T           Terminal type (IBM PC, ANSI, none):  IBM PC
 1               Print out the sets of species:  Yes
 2         Print indications of progress of run:  Yes
 3                              Print out tree:  Yes
 4               Write out trees onto tree file:  Yes

Are these settings correct? (type Y or the letter for one to change)
y

consense.exe: the file "outtree" that you wanted to
     use as output tree file already exists.
     Do you want to Replace it, Append to it,
     write to a new File, or Quit?
     (please type R, A, F, or Q)
f
Please enter a new file name> work1-nj-constree
```

```
((((((sp|Q9Y8A7|:10.0,sp|Q17754|:10.0):9.00,((((sp|Q4X0V2
|:10.0,sp|Q5AZH4|:10.0):6.00,
(((sp|Q6CJX3|:10.0,sp|Q75F72|:10.0):9.00,sp|Q6FRE1
|:10.0):10.0,sp|Q10481|:10.0):5.00):8.00,
sp|Q4PGT2|:10.0):6.00,sp|O45319|:10.0):10.0):9.00,sp|Q75DU7
|:10.0):7.00,sp|Q9WVA2|:10.0):7.00,
sp|Q9Y5J9|:10.0):10.0,sp|Q09783|:10.0);
```


TreeView - [work1-nj-constree]
File Edit Style Tree Window Help
sp|Q09783|
sp|Q9Y5J9|
sp|Q9WVA2|
sp|Q75DU7|
sp|Q9Y8A7|
sp|Q17754|
sp|O45319|
sp|Q4PGT2|
sp|Q4X0V2|
sp|Q5AZH4|
sp|Q10481|
sp|Q6FRE1|
sp|Q6CJX3|
sp|Q75F72|
10
7
7
9
9
10
6
6
8
5
10
9
PHYLIP_1 (1/1)

# Protein secondary structure

**Regular, recurring arrangements in space of adjacent amino acid residues in a polypeptide chain.**

**Maintained by hydrogen bonds between amide hydrogen and carbonyl oxygen of the peptide backbone.**

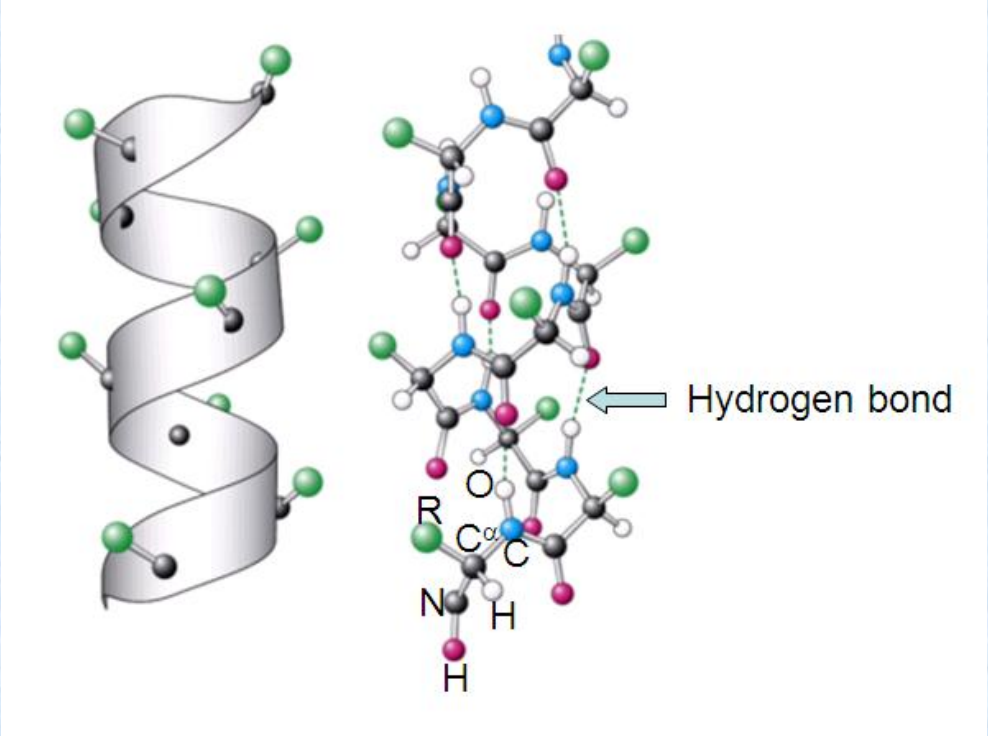
**1. α-helices**

**2. β-strands (sheets)**

**3. Turns (bends)**

**4. Coil (irregular)**

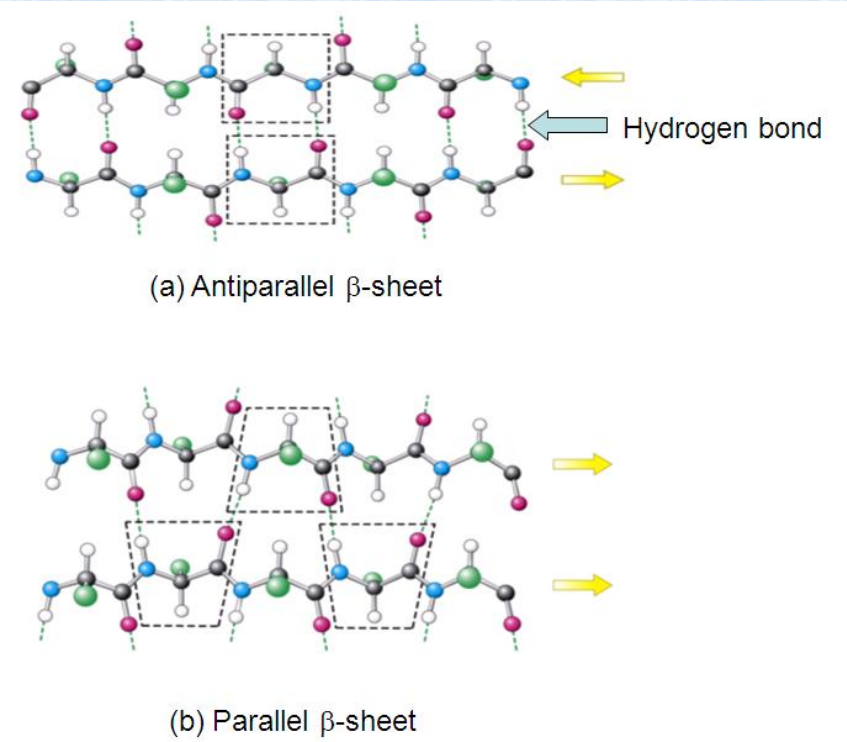
# Secondary structures: α-helix

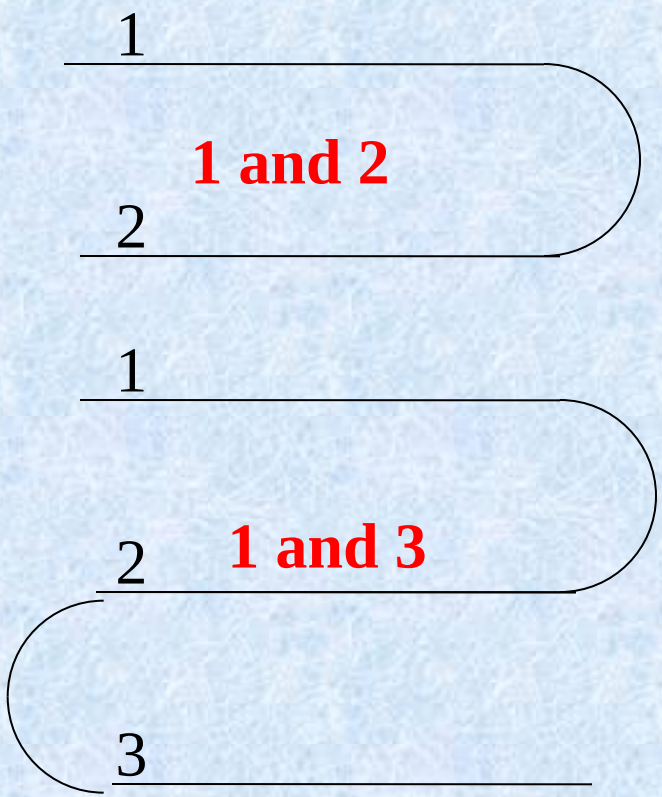



The α-helix is stabilized by hydrogen bonds between the CO and the NH groups of the main chain (i and i+4). 3.6 residues/turn; 5.4 Å/turn (repeating unit of α-helix)

Residues are closely packed.

# Secondary structures: β-strand




**β-strand is fully extended and loosely packed; they can hydrogen bond together to form β-sheets (parallel or anti parallel) depending on the direction.**

# Turns

One-third of all residues in globular proteins are contained in turns that serve to reverse the direction of the polypeptide chain.

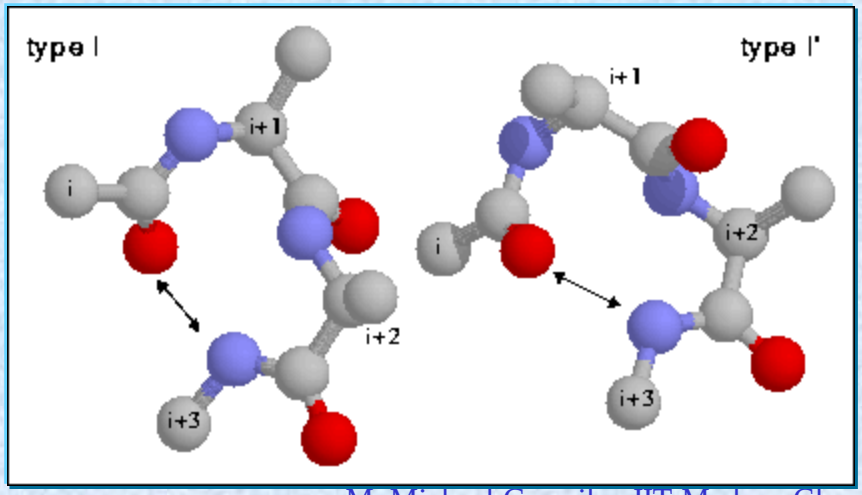
Turns contain a hydrogen bond between the carbonyl oxygen of residue i and the amide nitrogen of i+3.

These three types of turns are designated I, II, and III.

Type I turns occur most frequently (2-3 times more frequently than type II).

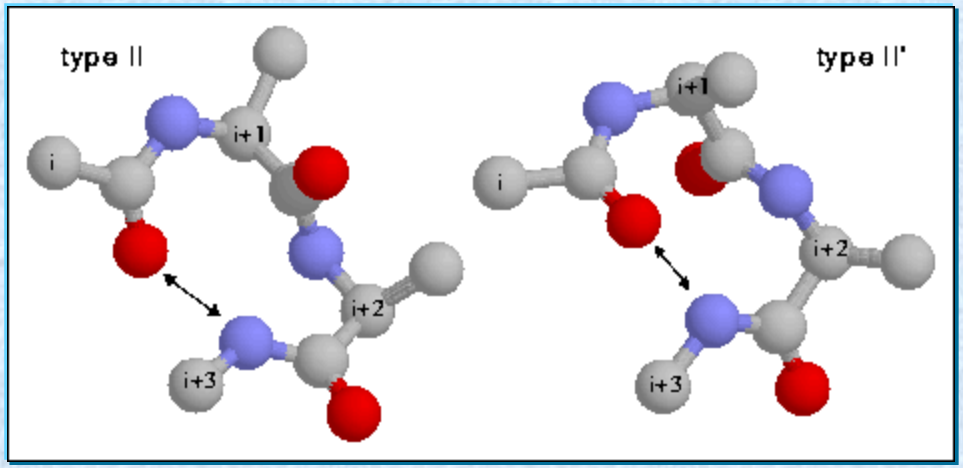
The backbone dihedral angles of residue are (-60, -30) and (-90, 0) of residues i+1 and i+2 of the type I turn.

Proline is often found in position i+1 in type I turns as its phi angle is restricted to -60.



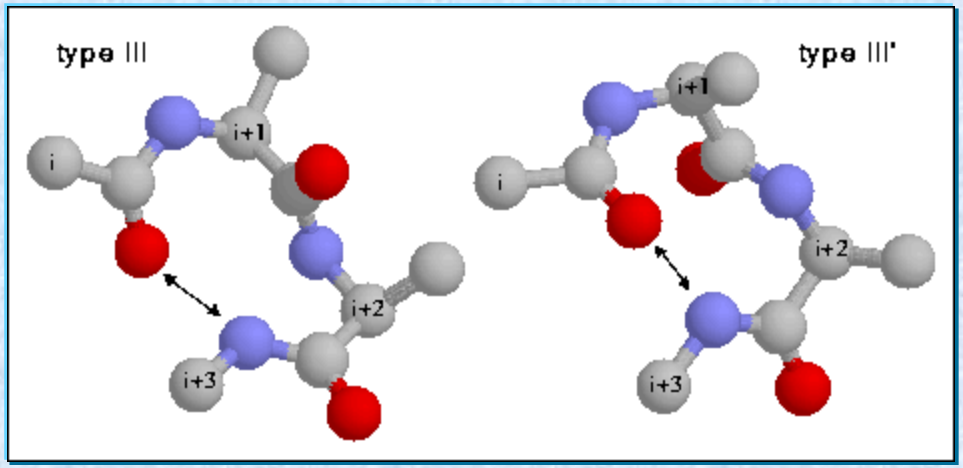
# Turns

The backbone dihedral angles of residue are (-60, 120) and (80, 0) of residues i+1 and i+2, respectively of the type II turn.

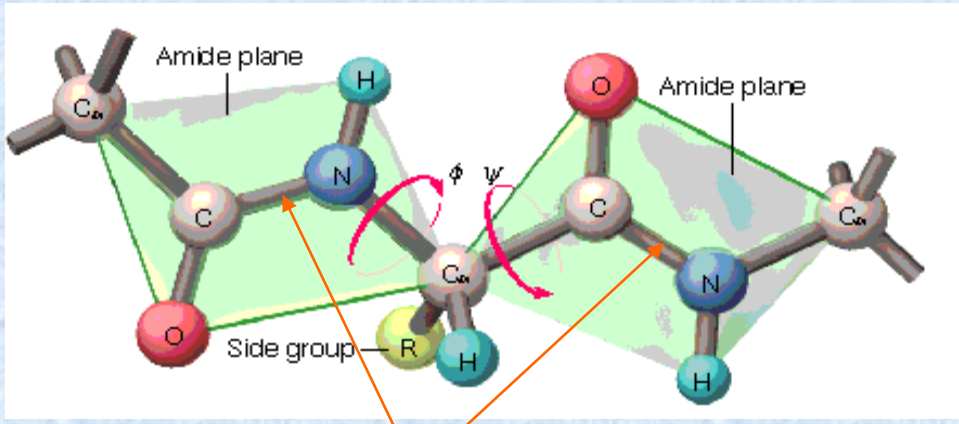


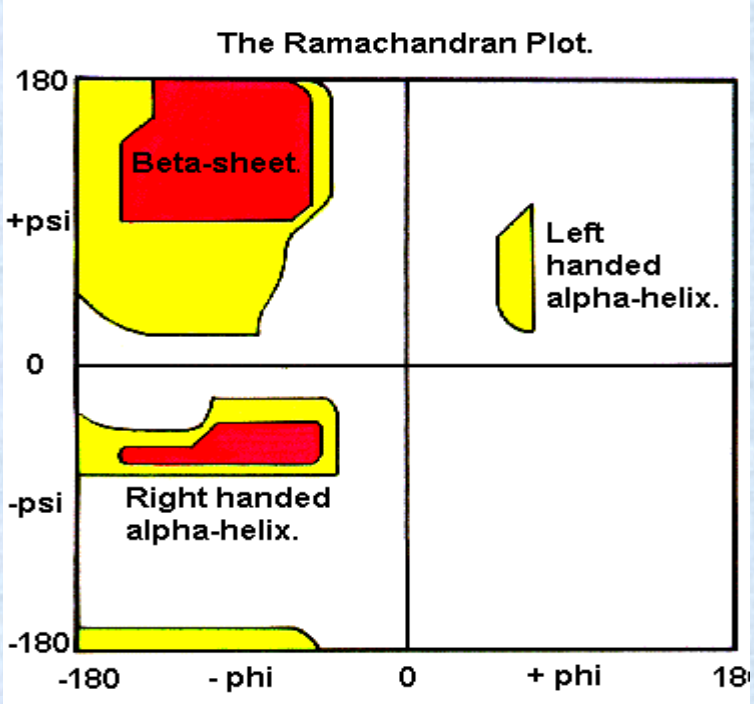

Glycine is favored in this position in the type II' as it requires a positive phi value.

The backbone dihedral angles of residue are (-60, -30) and (-60, -30) of residues i+1 and i+2, of the classical type III turn.



# Ramachandran plot




N-$C_\alpha$ and $C_\alpha$-C bonds are free to rotate.
These rotations are represented by the torsional angles, and , respectively.

G N Ramachandran used computer models of small polypeptides to systematically vary and for finding stable conformations.

Atoms were treated as hard spheres (dimensions -> van der Waals radii).

and angles that cause spheres to collide -> sterically disallowed conformation

Red: Allowed; White: Disallowed

# Torsional angles

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ATOM | 27 | CG | GLU | A | 4 | 7.583 | 14.248 | 24.732 | 1.00 | 14.95 | C |
| ATOM | 28 | CD | GLU | A | 4 | 7.577 | 15.734 | 24.340 | 1.00 | 15.89 | C |
| ATOM | 29 | OE1 | GLU | A | 4 | 6.511 | 16.344 | 24.159 | 1.00 | 17.02 | O |
| ATOM | 30 | OE2 | GLU | A | 4 | 8.572 | 16.043 | 23.730 | 1.00 | 15.37 | O |
| ATOM | 31 | N | GLY | A | 5 | 7.687 | 11.127 | 22.185 | 1.00 | 12.43 | N |
| ATOM | 32 | CA | GLY | A | 5 | 8.655 | 10.062 | 21.811 | 1.00 | 14.67 | C |
| ATOM | 33 | C | GLY | A | 5 | 9.176 | 10.126 | 20.356 | 1.00 | 16.35 | C |
| ATOM | 34 | O | GLY | A | 5 | 10.347 | 9.825 | 20.076 | 1.00 | 11.57 | O |
| ATOM | 35 | N | GLU | A | 6 | 8.315 | 10.552 | 19.440 | 1.00 | 13.50 | N |
| ATOM | 36 | CA | GLU | A | 6 | 8.800 | 10.689 | 18.054 | 1.00 | 12.50 | C |
| ATOM | 37 | C | GLU | A | 6 | 9.632 | 11.962 | 17.942 | 1.00 | 13.96 | C |
| ATOM | 38 | O | GLU | A | 6 | 10.617 | 11.986 | 17.195 | 1.00 | 14.23 | O |
| ATOM | 39 | CB | GLU | A | 6 | 7.614 | 10.784 | 17.076 | 1.00 | 13.07 | C |

(x1,y1,z1)
(x2,y2,z2)
(x3,y3,z3)

## Dihedral Angles

### Surface One

| | x | y | z |
|---|---|---|---|
| Point 1 | 9.1760 | 10.126 | 20.3560 |
| Point 2 | 8.3150 | 10.5520 | 19.4400 |
| Point 3 | 8.8000 | 10.6890 | 18.0540 |

| Distances | |
|---|---|
| 1-2 | 1.32734810807 |
| 2-3 | 1.47478473005 |
| 3-1 | 2.39948932066 |

Equation of Plane Ax+By+Cz+D=0

-0.46494399999999 x + -1.637606 y + -0.324567 z + 27.4556103519997 = 0

### Surface Two

| | x | y | z |
|---|---|---|---|
| Point 1 | 8.3150 | 10.5520 | 19.4400 |
| Point 2 | 8.8000 | 10.6890 | 18.0540 |
| Point 3 | 9.6320 | 11.9620 | 17.9420 |

| Distances | |
|---|---|
| 1-2 | 1.47478473005 |
| 2-3 | 1.52489245522 |
| 3-1 | 2.44266104893 |

Equation of Plane Ax+By+Cz+D=0

1.749034 x + -1.09883199999999 y + 0.50342100000000 z + -12.734846686000 = 0

**Dihedral Angle =** 1.34556071239086 radians

77.094949898613 degrees

$$A = \begin{vmatrix} 1 & y1 & z1 \\ 1 & y2 & z2 \\ 1 & y3 & z3 \end{vmatrix} \quad B = \begin{vmatrix} x1 & 1 & z1 \\ x2 & 1 & z2 \\ x3 & 1 & z3 \end{vmatrix} \quad C = \begin{vmatrix} x1 & y1 & 1 \\ x2 & y2 & 1 \\ x3 & y3 & 1 \end{vmatrix} \quad D = -\begin{vmatrix} x1 & y1 & z1 \\ x2 & y2 & z2 \\ x3 & y3 & z3 \end{vmatrix}$$

## Dihedral Angle Calculator

A dihedral angle is the angle between two planes.

To calculate this angle, you can follow these steps:

1. **Calculate the equation for each plane**. It will be in the form:

$$Ax + By + Cz + D = 0$$

2. Then, knowing the equation of the two planes, you can **calculate the dihedral angle**:

$$\cos\alpha = \frac{A_1A_2 + B_1B_2 + C_1C_2}{\sqrt{A_1^2 + B_1^2 + C_1^2}\sqrt{A_2^2 + B_2^2 + C_2^2}}$$

# Check whether the residues 60-69 form an α-helix (4MBN)

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ATOM | 483 | N | ASP | A | 60 | 30.554 | 28.872 | -1.439 | 1.00 | 13.41 |
| ATOM | 484 | CA | ASP | A | 60 | 29.251 | 29.577 | -1.346 | 1.00 | 13.47 |
| ATOM | 485 | C | ASP | A | 60 | 28.082 | 28.593 | -1.208 | 1.00 | 10.85 |
| ATOM | 486 | O | ASP | A | 60 | 27.129 | 28.854 | -0.468 | 1.00 | 10.89 |
| ATOM | 487 | CB | ASP | A | 60 | 29.115 | 30.488 | -2.563 | 1.00 | 11.63 |
| ATOM | 488 | CG | ASP | A | 60 | 27.959 | 31.474 | -2.407 | 1.00 | 11.34 |
| ATOM | 489 | OD1 | ASP | A | 60 | 27.239 | 31.637 | -3.388 | 1.00 | 10.56 |
| ATOM | 490 | OD2 | ASP | A | 60 | 27.926 | 32.207 | -1.407 | 1.00 | 11.89 |
| ATOM | 491 | N | LEU | A | 61 | 28.180 | 27.469 | -1.910 | 1.00 | 11.29 |
| ATOM | 492 | CA | LEU | A | 61 | 27.158 | 26.395 | -1.785 | 1.00 | 13.55 |
| ATOM | 493 | C | LEU | A | 61 | 27.140 | 25.792 | -0.377 | 1.00 | 12.39 |
| ATOM | 494 | O | LEU | A | 61 | 26.076 | 25.569 | 0.204 | 1.00 | 10.30 |
| ATOM | 495 | CB | LEU | A | 61 | 27.403 | 25.337 | -2.873 | 1.00 | 13.67 |
| ATOM | 496 | CG | LEU | A | 61 | 26.455 | 24.132 | -2.929 | 1.00 | 12.05 |
| ATOM | 497 | CD1 | LEU | A | 61 | 26.375 | 23.553 | -4.341 | 1.00 | 12.93 |
| ATOM | 498 | CD2 | LEU | A | 61 | 26.930 | 23.031 | -2.000 | 1.00 | 16.59 |
| ATOM | 499 | N | LYS | A | 62 | 28.333 | 25.566 | 0.144 | 1.00 | 10.86 |
| ATOM | 500 | CA | LYS | A | 62 | 28.491 | 25.175 | 1.544 | 1.00 | 13.89 |
| ATOM | 501 | C | LYS | A | 62 | 27.916 | 26.199 | 2.531 | 1.00 | 13.74 |
| ATOM | 502 | O | LYS | A | 62 | 27.209 | 25.822 | 3.464 | 1.00 | 12.28 |
| ATOM | 503 | CB | LYS | A | 62 | 30.002 | 25.061 | 1.776 | 1.00 | 12.89 |
| ATOM | 504 | CG | LYS | A | 62 | 30.365 | 24.572 | 3.153 | 1.00 | 13.43 |
| ATOM | 505 | CD | LYS | A | 62 | 31.694 | 23.828 | 3.063 | 1.00 | 12.14 |
| ATOM | 506 | CE | LYS | A | 62 | 32.046 | 23.190 | 4.390 | 1.00 | 15.51 |
| ATOM | 507 | NZ | LYS | A | 62 | 33.369 | 22.534 | 4.287 | 1.00 | 17.63 |
| ATOM | 508 | N | LYS | A | 63 | 28.139 | 27.478 | 2.257 | 1.00 | 13.23 |
| ATOM | 509 | CA | LYS | A | 63 | 27.570 | 28.584 | 3.080 | 1.00 | 12.05 |
| ATOM | 510 | C | LYS | A | 63 | 26.048 | 28.526 | 3.104 | 1.00 | 11.13 |
| ATOM | 511 | O | LYS | A | 63 | 25.438 | 28.658 | 4.164 | 1.00 | 12.64 |
| ATOM | 512 | CB | LYS | A | 63 | 27.932 | 29.946 | 2.489 | 1.00 | 11.27 |
| ATOM | 513 | CG | LYS | A | 63 | 28.277 | 30.918 | 3.588 | 1.00 | 13.99 |
| ATOM | 514 | CD | LYS | A | 63 | 28.534 | 32.349 | 3.116 | 1.00 | 14.19 |
| ATOM | 515 | CE | LYS | A | 63 | 29.792 | 32.484 | 2.276 | 1.00 | 16.43 |
| ATOM | 516 | NZ | LYS | A | 63 | 29.892 | 33.917 | 1.896 | 1.00 | 15.02 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ATOM | 517 | N | HIS | A | 64 | 25.479 | 28.309 | 1.940 | 1.00 | 11.61 |
| ATOM | 518 | CA | HIS | A | 64 | 24.016 | 28.237 | 1.840 | 1.00 | 10.72 |
| ATOM | 519 | C | HIS | A | 64 | 23.445 | 26.952 | 2.461 | 1.00 | 10.86 |
| ATOM | 520 | O | HIS | A | 64 | 22.427 | 27.022 | 3.140 | 1.00 | 11.94 |
| ATOM | 521 | CB | HIS | A | 64 | 23.561 | 28.299 | 0.392 | 1.00 | 11.17 |
| ATOM | 522 | CG | HIS | A | 64 | 22.055 | 28.401 | 0.271 | 1.00 | 11.82 |
| ATOM | 523 | ND1 | HIS | A | 64 | 21.332 | 29.442 | 0.778 | 1.00 | 12.13 |
| ATOM | 524 | CD2 | HIS | A | 64 | 21.170 | 27.481 | -0.259 | 1.00 | 16.49 |
| ATOM | 525 | CE1 | HIS | A | 64 | 20.008 | 29.178 | 0.545 | 1.00 | 11.48 |
| ATOM | 526 | NE2 | HIS | A | 64 | 19.911 | 27.978 | -0.083 | 1.00 | 12.54 |
| ATOM | 527 | N | GLY | A | 65 | 24.189 | 25.846 | 2.403 | 1.00 | 9.84 |
| ATOM | 528 | CA | GLY | A | 65 | 23.765 | 24.640 | 3.124 | 1.00 | 10.13 |
| ATOM | 529 | C | GLY | A | 65 | 23.720 | 24.833 | 4.640 | 1.00 | 12.09 |
| ATOM | 530 | O | GLY | A | 65 | 22.785 | 24.356 | 5.310 | 1.00 | 10.90 |
| ATOM | 531 | N | VAL | A | 66 | 24.653 | 25.629 | 5.138 | 1.00 | 11.12 |
| ATOM | 532 | CA | VAL | A | 66 | 24.554 | 26.025 | 6.578 | 1.00 | 11.96 |
| ATOM | 533 | C | VAL | A | 66 | 23.297 | 26.851 | 6.899 | 1.00 | 10.82 |
| ATOM | 534 | O | VAL | A | 66 | 22.631 | 26.593 | 7.910 | 1.00 | 11.04 |
| ATOM | 535 | CB | VAL | A | 66 | 25.799 | 26.777 | 7.027 | 1.00 | 12.19 |
| ATOM | 536 | CG1 | VAL | A | 66 | 25.709 | 27.227 | 8.478 | 1.00 | 11.49 |
| ATOM | 537 | CG2 | VAL | A | 66 | 27.062 | 25.931 | 6.850 | 1.00 | 13.42 |
| ATOM | 538 | N | THR | A | 67 | 22.967 | 27.768 | 5.995 | 1.00 | 11.19 |
| ATOM | 539 | CA | THR | A | 67 | 21.780 | 28.628 | 6.142 | 1.00 | 11.60 |
| ATOM | 540 | C | THR | A | 67 | 20.497 | 27.825 | 6.258 | 1.00 | 9.84 |
| ATOM | 541 | O | THR | A | 67 | 19.740 | 28.004 | 7.218 | 1.00 | 10.87 |
| ATOM | 542 | CB | THR | A | 67 | 21.678 | 29.575 | 4.949 | 1.00 | 11.28 |
| ATOM | 543 | OG1 | THR | A | 67 | 22.640 | 30.627 | 5.157 | 1.00 | 12.09 |
| ATOM | 544 | CG2 | THR | A | 67 | 20.315 | 30.245 | 4.853 | 1.00 | 12.21 |
| ATOM | 545 | N | VAL | A | 68 | 20.397 | 26.839 | 5.373 | 1.00 | 11.34 |
| ATOM | 546 | CA | VAL | A | 68 | 19.193 | 25.996 | 5.309 | 1.00 | 9.46 |
| ATOM | 547 | C | VAL | A | 68 | 19.020 | 25.073 | 6.515 | 1.00 | 8.75 |
| ATOM | 548 | O | VAL | A | 68 | 17.966 | 25.083 | 7.152 | 1.00 | 10.55 |
| ATOM | 549 | CB | VAL | A | 68 | 19.215 | 25.219 | 3.994 | 1.00 | 9.32 |
| ATOM | 550 | CG1 | VAL | A | 68 | 18.113 | 24.174 | 3.924 | 1.00 | 12.56 |
| ATOM | 551 | CG2 | VAL | A | 68 | 19.100 | 26.200 | 2.837 | 1.00 | 10.55 |
| ATOM | 552 | N | LEU | A | 69 | 20.079 | 24.401 | 6.906 | 1.00 | 11.18 |
| ATOM | 553 | CA | LEU | A | 69 | 19.977 | 23.500 | 8.092 | 1.00 | 12.51 |
| ATOM | 554 | C | LEU | A | 69 | 19.814 | 24.216 | 9.418 | 1.00 | 10.58 |
| ATOM | 555 | O | LEU | A | 69 | 19.230 | 23.676 | 10.347 | 1.00 | 10.91 |

# DSSP: Dictionary of Secondary Structures in Proteins

**Sequence and secondary structure for 4MBN chain A**

```
1      VLSEGEWQLV LHVWAKVEAD VAGHGQDILI RLFKSHPETL EKFDRFKHLK
          HHHHHHH HHHHHHHGGG HHHHHHHHHH HHHHH HHHH HT GGGTT

51     TEAEMKASED LKKHGVTVLT ALGAILKKKG HHEAELKPLA QSHATKHKIP
       SHHHHHH HH HHHHHHHHHH HHHHHHTTTT   HHHHHHHH HHHHHTS

101    IKYLEFISEA IIHVLHSRHP GDFGADAQGA MNKALELFRK DIAAKYKELG
       HHHHHHHHHH HHHHHHHH G GGS HHHHHH HHHHHHHHHH HHHHHHHHHT

151    YQG
```

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | A | G | H | > | S+ | 0 | 0 | 35 | 2,-0.2 | 4,-1.6 | 1,-0.2 | -1,-0.2 | 0.823 | 107.4 | 48.1 | -63.[illegible] | [illegible]4.5 |
| 6 | A | E | H | > | S+ | 0 | 0 | 51 | 2,-0.2 | 4,-1.8 | 1,-0.2 | -1,-0.2 | 0.883 | 109.7 | 52.9 | -77.1 | -34.6 |
| 7 | A | W | H | X | S+ | 0 | 0 | 15 | -4,-2.7 | 4,-2.6 | 2,-0.2 | 5,-0.3 | 0.894 | 105.3 | 56.2 | -63.6 | -34.4 |
| 8 | A | Q | H | X | S+ | 0 | 0 | 133 | -4,-2.1 | 4,-2.5 | 1,-0.2 | 5,-0.2 | 0.938 | 107.3 | 48.2 | -56.7 | -47.2 |
| 9 | A | L | H | X | S+ | 0 | 0 | 55 | -4,-1.6 | 4,-1.5 | 1,-0.2 | -1,-0.2 | 0.855 | 112.8 | 50.0 | -60.7 | -40.5 |
| 10 | A | V | H | X | S+ | 0 | 0 | 0 | -4,-1.8 | 4,-2.0 | 2,-0.2 | -1,-0.2 | 0.917 | 114.6 | 40.0 | -65.7 | -50.1 |
| 11 | A | L | H | X | S+ | 0 | 0 | 44 | -4,-2.6 | 4,-2.2 | 2,-0.2 | -2,-0.2 | 0.842 | 107.7 | 61.6 | -78.6 | -27.0 |
| 12 | A | H | H | X | S+ | 0 | 0 | 120 | -4,-2.5 | 4,-0.6 | -5,-0.3 | -1,-0.2 | 0.965 | 109.4 | 43.1 | -64.9 | -40.0 |

# Propensity

**The propensity of an amino acid residue i in any conformation (helix or strand or turn or coil) has been defined as the percentage of residue i in that conformation to the percentage of all residues in the same conformation.**

**propensity$_\alpha$(i) = % of residue i in α-helix/ % of all residues in α-helix.**

**% of residue i in α-helix = $n_\alpha(i)/N(i)$**

**$n_\alpha(i)$ = number of residues of type i in α-helix**

**N(i) = number of residues of type i in the whole dataset**

**% of all residues in α-helix = $n_\alpha/N$**

**$n_a$ = total number of residues in a-helix**

**N = total number of residues in the whole dataset**

# Propensity

```
VLSEGEWQLV LHVWAKVEAD VAGHGQDILI RLFKSHPETL EKFDRFKHLK
   HHHHHHH HHHHHHHGGG HHHHHHHHHH HHHHH HHHH HT GGGTT

TEAEMKASED LKKHGVTVLT ALGAILKKKG HHEAELKPLA QSHATKHKIP
SHHHHHH HH HHHHHHHHHH HHHHHHTTTT   HHHHHHHH HHHHHTS

IKYLEFISEA IIHVLHSRHP GDFGADAQGA MNKALELFRK DIAAKYKELG
HHHHHHHHHH HHHHHHHH G GGS HHHHHH HHHHHHHHHH HHHHHHHHHT

YQG
```

**E.g. Ala: % of Ala in α-helix = $N_\alpha$(Ala)/N(Ala) = 15/16 = 0.94**

**% of all residues in α-helix = $N_\alpha$/N = 115/153 = 0.75**

**Propensity of Ala = 0.94/0.75 = 1.25**

**Propensity of Gly:**

# Algorithm

1. Compute the occurrence of 20 residues in helix

2. Compute the occurrence of 20 residues in whole protein

3. Compute the ratio

4. Compute total number of residues in helix

5. Divide with total number of residues in a protein

6. Divide 3 by 5, to get the propensity of all the 20 amino acid residues in helix.

**Table 5.2** Chou–Fasman parameters

| Residue | $P_\alpha$ | Residue | $P_\beta$ | Residue | $P_t$ |
|---|---|---|---|---|---|
| Glu | Hα 1.53 | Hβ Met | 1.67 | Asn | 1.68 |
| Ala | 1.45 | Val | 1.65 | Gly | 1.68 |
| Leu | 1.34 | Ile | 1.60 | Ser | 1.56 |
| His | hα 1.24 | hβ Cys | 1.30 | Pro | 1.54 |
| Met | 1.20 | Tyr | 1.29 | Asp | 1.26 |
| Gln | 1.17 | Phe | 1.28 | Tyr | 1.25 |
| Trp | 1.14 | Gln | 1.23 | Cys | 1.17 |
| Val | 1.14 | Leu | 1.22 | Trp | 1.11 |
| Phe | 1.12 | Thr | 1.20 | Lys | 1.01 |
| Lys | Iα 1.07 | Trp | 1.19 | Arg | 1.00 |
| Ile | 1.00 | Iβ Ala | 0.97 | Thr | 1.00 |
| Asp | iα 0.98 | iβ Arg | 0.90 | Phe | 0.71 |
| Thr | 0.82 | Gly | 0.81 | His | 0.69 |
| Ser | 0.79 | Asp | 0.80 | Met | 0.67 |
| Arg | 0.79 | bβ Lys | 0.74 | Ile | 0.58 |
| Cys | 0.77 | Ser | 0.72 | Ala | 0.57 |
| Asn | bα 0.73 | His | 0.71 | Gln | 0.56 |
| Tyr | 0.61 | Asn | 0.65 | Leu | 0.53 |
| Pro | Bα 0.59 | Pro | 0.62 | Glu | 0.44 |
| Gly | 0.53 | Bβ Glu | 0.26 | Val | 0.30 |

**Hα β : Strong helix (strand) former**
**hα β : Helix (strand) former**
**Iα β : Weak helix (strand) former**
**iα β : Weak helix (strand) breaker**
**bα β : Helix (strand) breaker**
**Bα β : Strong helix (strand) breaker**

# Rules for identifying helix

**The propensity values and the residues belonging to these 12 classes are shown in Table 5.2.**

**For a protein sequence, assign the appropriate parameters from Table 5.2.**

**Helix:**

**the values of the six parameters are Ha = ha =1; Ia = ia = 0; Ba = ba = -1;**

**scan for window of 6 residues, where score  4, i.e. at least four helix formers and not more than one helix breaker;**

**extend the length in both directions until 4-residue window has the average propensity < 1;**

**continue the search and locate all helical regions in the sequence.**

**Sequence and secondary structure for 4LYZ chain A**

```
1      KVFGRCELAA AMKRHGLDNY RGYSLGNWVC AAKFESNFNT QATNRNTDGS
        B  HHHHHH HHHHTT TTB TTB HHHHHH HHHHHHTTBS S EEE SSS

51     TDYGILQINS RWWCNDGRTP GSRNLCNIPC SALLSSDITA SVNCAKKIVS
       EEETTTTEET TTT B SS T T   SS SBG GGGGSS  HH HHHHHHHHTT

101    DGNGMNAWVA WRNRCKGTDV QAWIRGCRL
       TSSGGGGSHH HHHHTTTS G GGGSTT
```

**KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGS**
**TDYGILQINSRWWCNDGRTPGSRNLCNIPCSALLSSDITASVNCAKKIVS**
**DGNGMNAWVAWRNRCKGTDVQAWIRGCRL**

KVFGRC
VFGRCE
FGRCEL
GRCELA
RCELAA

# Rules for identifying β-strand

**The values of the six parameters are Hb = hb =1; Ib = ib = 0; Bb = bb = -1;**

**scan for window of 5 residues, where score > 3, i.e. at least three strand formers and not more than one strand breaker;**

**extend the length in both directions until 4-residue window has the average propensity < 1;**

**continue the search and locate all strand regions in the sequence.**

**Conflict situation:**

**a region containing overlapping helical and strand assignments is considered as a helix (or strand) if average propensity of a-helix (b-strand) is greater than that of b-strand (a-helix).**

**Choose:**

(A) Program, (B) Protein (sequence/accession) (C) Analyze protein:

(A) Program: Chou-Fasman Secondary Structure prediction

(B) Protein sequence: FASTA format

```
>4LYZ:A|PDBID|CHAIN|SEQUENCE
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQIN
SRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
```

(C) Do Analysis — Submit Sequence

```
 CHOFAS predicts protein secondary structure
 version 2.0u66 September 1998
Please cite:
 Chou and Fasman (1974) Biochem., 13:222-245
```

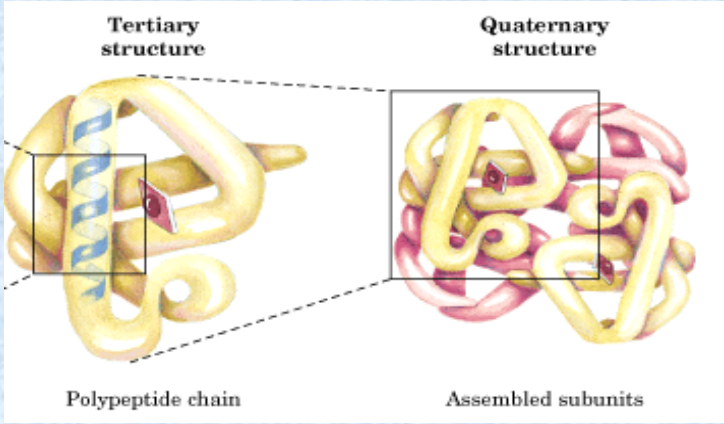
---

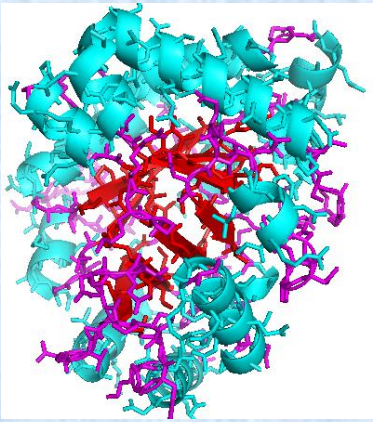
```
 Chou-Fasman plot of @, 129 aa;
4LYZ:A|PDBID|CHAIN|SEQUENCE


                .         .         .         .         .         .
       KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINS
 helix     <-------->             <-------->                  <>
 sheet                        EEEEEEE         EEEEEEE      EEEEEEEE
 turns T                                    T   T    T   T T        T

                .         .         .         .         .         .
       RWWCNDGRTPGSRNLCNIPCSALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDV
 helix                        <--------------->    <---->        <>
 sheet                EEEE                                       EEE
 turns      TT    T T            T         T    T T         T


       QAWIRGCRL
 helix
 sheet EEEE
 turns

 Residue totals: H: 47   E: 33   T: 16
        percent: H: 36.4 E: 25.6 T: 12.4
```

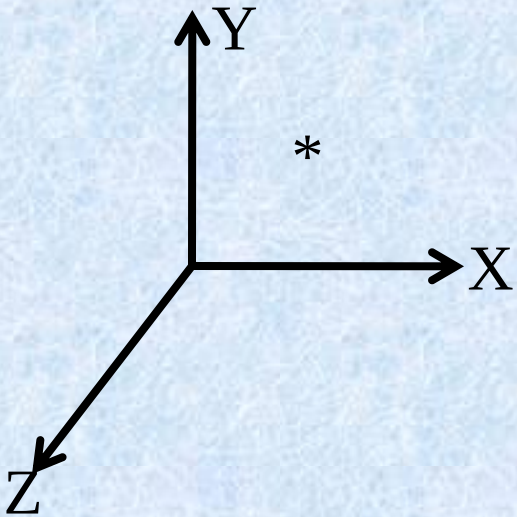
**http://fasta.bioch.virginia.edu/fasta/chofas.htm**

# Tertiary and quaternary structures

**Tertiary structure provides the information about the three-dimensional structure of a protein with atomic details; the positions of each atom in Cartesian coordinate system.**






3D structure of a protein contains the complete information.

# 3D structure determination Experiment

## X-ray crystallography

**Protein purification**

**Protein crystallization**

**Crystal mounting**

**Data collection**

distribution of electrons in the molecule,

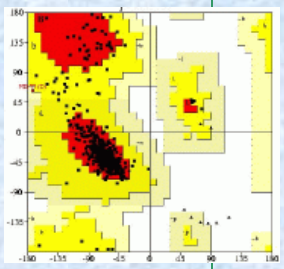
*i.e.* an electron density map.

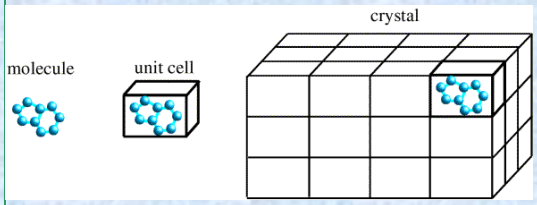
**Data processing**

**Structure solution**

**Structure refinement**

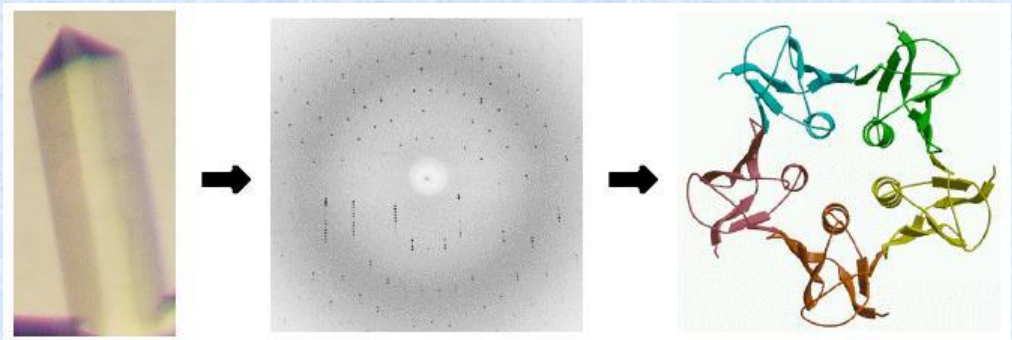
**Structure analysis**






**NMR spectroscopy**
**Electron microscopy**

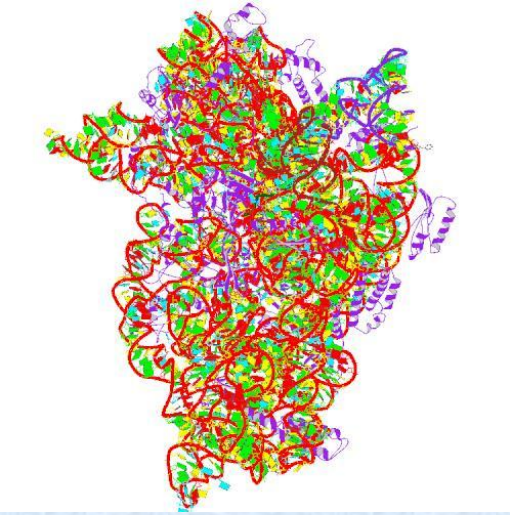
**It is based on the quantum mechanical properties of atoms, particularly spin, and it determines information about atoms from the fact that their local environment influences how they respond to applied magnetic fields.**



**All available structures are stored in Protein Data Bank (PDB).**

# Protein Data Bank

Myoglobin, Kendrew (1958)



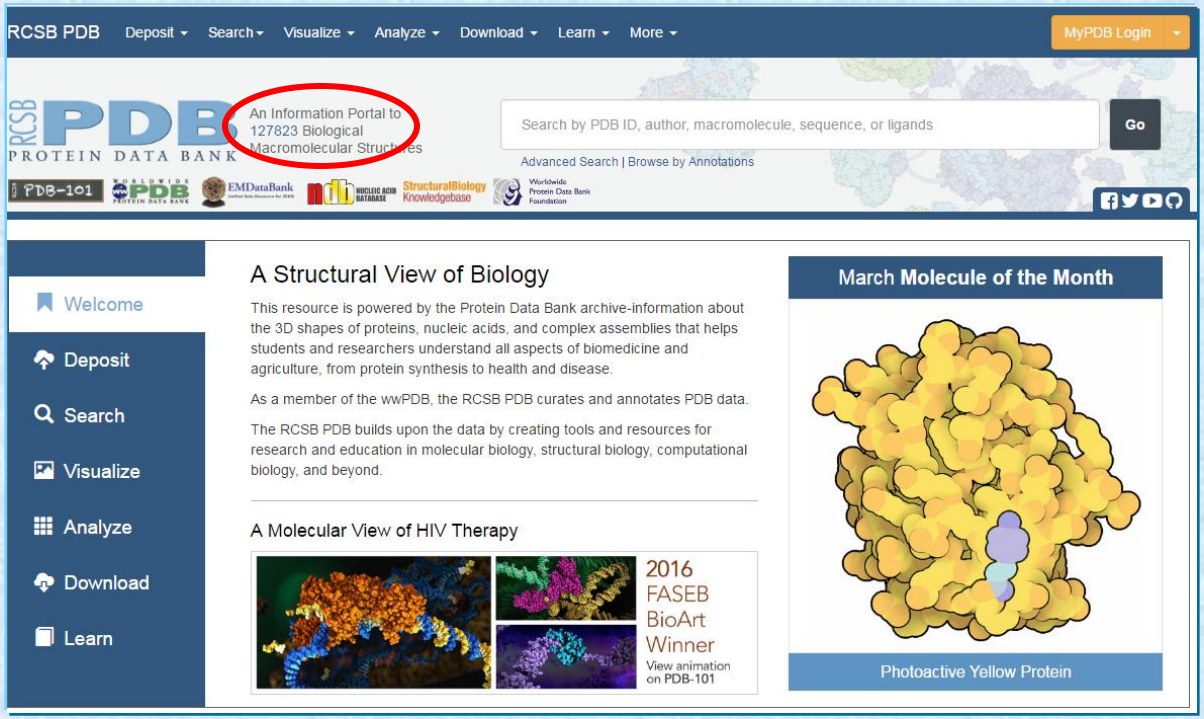
PDB (USA), PDB (Europe) and PDB (**Japan**)

World wide Protein Data Bank: Organizations that act as deposition, data processing and distribution centers for PDB data.

2009 **Nobel Prize for Chemistry** has been awarded to Venki Ramakrishnan (UK), Tom Steitz (USA) and Ada Yonath (Israel). These three crystallographers have contributed enormously to our understanding of how the protein production machinery works at the atomic level by **determining the detailed three-dimensional structure** of so-called ribosomes.

# Protein Data Bank

CODE
**2LZMA**




| Exp.Method | Proteins | Nucleic Acids | Protein/NA Complexes | Other | Total |
|---|---|---|---|---|---|
| X-RAY | 107061 | 1820 | 5471 | 4 | 114356 |
| NMR | 10300 | 1190 | 241 | 8 | 11739 |
| ELECTRON MICROSCOPY | 1022 | 30 | 367 | 0 | 1419 |
| HYBRID | 99 | 3 | 2 | 1 | 105 |
| other | 181 | 4 | 6 | 13 | 204 |
| Total | 118663 | 3047 | 6087 | 26 | 127823 |

Summary | Sequence | Annotations | Seq. Similarity | 3D Similarity | Literature | Biol. & Chem. | Methods | Geometry | Links

## STRUCTURE OF BACTERIOPHAGE T4 LYSOZYME REFINED AT 1.7 ANGSTROMS RESOLUTION

2LZM

Display Files

close

- FASTA Sequence
- PDB File
- PDB File (Header)
- mmCIF File
- mmCIF File (Header)
- PDBML/XML File
- PDBML/XML File (Header)

DOI: 10.2210/pdb2lzm/pdb

## ENTRY 2LZM SUPERSEDES 1LZM

### Primary Citation

**Structure of bacteriophage T4 lysozyme refined at 1.7 A resolution.**

Weaver, L.H., Matthews, B.W.

**Journal:** (1987) J.Mol.Biol. **193:** 189-199

**PubMed:** 3586019
**Search Related Articles in** PubMed

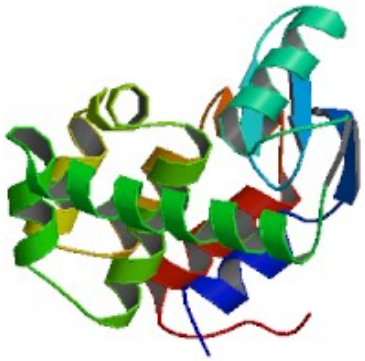
**PubMed Abstract:**

The structure of the lysozyme from bacteriophage T4 has been refined at 1.7 A resolution to a crystallographic residual of 19.3%. The final model has bond lengths and bond angles that differ from "ideal" values by 0.019 A and 2.7... [ Read More & Search PubMed Abstracts ]

### Related Citations in PDB Entry (REMARK 1) — Hide

Temperature-Sensitive Mutations of Bacteriophage T4 Lysozyme Occur at Sites with Low Mobility and Low Solvent Accessibility in the Folded Protein
Alber, T., Dao-Pin, S., Nye, J.A., Muchmore, D.C., Matthews, B.W.
(1987) *Biochemistry* ***26:*** *3754*

Biolo



More Images...

View in Jmol | Simple Viewer
Other Viewers | Protein Workshop

Biological assembly 1 assigned by authors

```
>2LZM:A|PDBID|CHAIN|SEQUENCE
MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRNCNGVITKDEAEKLFNQDVDAAVRGILR
NAKLKPVYDSLDAVRRCALINMVFQMGETGVAGFTNSLRMLQQKRWDEAAVNLAKSRWYNQTPNRAKRVITTFRTGTWDA
YKNL
```

# Protein Data Bank

```
HEADER    HYDROLASE (O-GLYCOSYL)                          18-A
TITLE     STRUCTURE OF BACTERIOPHAGE T4 LYSOZYME REFIN
TITLE    2 RESOLUTION
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: T4 LYSOZYME;
COMPND   3 CHAIN: A;
COMPND   4 EC: 3.2.1.17;
COMPND   5 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: ENTEROBACTERIA
SOURCE   3 ORGANISM_TAXID: 10665;
SOURCE   4 ORGAN: EGG
KEYWDS    HYDROLASE (O-GLYCOSYL)
EXPDTA    X-RAY DIFFRACTION
AUTHOR    L.H.WEAVER,B.W.MATTHEWS
REVDAT   7   13-JUL-11 2LZM    1       VERSN
REVDAT   6   24-FEB-09 2LZM    1       VERSN
REVDAT   5   01-APR-03 2LZM    1       JRNL
REVDAT   4   16-APR-88 2LZM    1       REMARK
REVDAT   3   16-OCT-87 2LZM    1       REMARK
REVDAT   2   16-APR-87 2LZM    1       JRNL
REVDAT   1   24-OCT-86 2LZM    0
SPRSDE     24-OCT-86 2LZM      1LZM
JRNL        AUTH   L.H.WEAVER,B.W.MATTHEWS
JRNL        TITL   STRUCTURE OF BACTERIOPHAGE T
JRNL        TITL 2 RESOLUTION.
JRNL        REF    J.MOL.BIOL.
JRNL        REFN                    ISSN 0022-28
JRNL        PMID   3586019
JRNL        DOI    10.1016/0022-2836(87)90636-X
REMARK   1
REMARK   1 REFERENCE 1
REMARK   1  AUTH   T.ALBER,S.DAO-PIN,J.A.NYE,D.
REMARK   1  TITL   TEMPERATURE-SENSITIVE MUTATI
REMARK   1  TITL 2 OCCUR AT SITES WITH LOW MOBI
REMARK   1  TITL 3 ACCESSIBILITY IN THE FOLDED
REMARK   1  REF    BIOCHEMISTRY
REMARK   1  REFN                    ISSN 0006-29
```

```
REMARK   2
REMARK   2 RESOLUTION.    1.70 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3   PROGRAM     : TNT
REMARK   3   AUTHORS     : TRONRUD,TEN EYCK,MATTHEWS
REMARK   3
```

```
SEQRES   1 A  164  MET ASN ILE PHE GLU MET LEU ARG ILE ASP GLU GLY LEU
SEQRES   2 A  164  ARG LEU LYS ILE TYR LYS ASP THR GLU GLY TYR TYR THR
SEQRES   3 A  164  ILE GLY ILE GLY HIS LEU LEU THR LYS SER PRO SER LEU
SEQRES   4 A  164  ASN ALA ALA LYS SER GLU LEU ASP LYS ALA ILE GLY ARG
SEQRES   5 A  164  ASN CYS ASN GLY VAL ILE THR LYS ASP GLU ALA GLU LYS
SEQRES   6 A  164  LEU PHE ASN GLN ASP VAL ASP ALA ALA VAL ARG GLY ILE
SEQRES   7 A  164  LEU ARG ASN ALA LYS LEU LYS PRO VAL TYR ASP SER LEU
SEQRES   8 A  164  ASP ALA VAL ARG ARG CYS ALA LEU ILE ASN MET VAL PHE
SEQRES   9 A  164  GLN MET GLY GLU THR GLY VAL ALA GLY PHE THR ASN SER
SEQRES  10 A  164  LEU ARG MET LEU GLN GLN LYS ARG TRP ASP GLU ALA ALA
SEQRES  11 A  164  VAL ASN LEU ALA LYS SER ARG TRP TYR ASN GLN THR PRO
SEQRES  12 A  164  ASN ARG ALA LYS ARG VAL ILE THR THR PHE ARG THR GLY
SEQRES  13 A  164  THR TRP ASP ALA TYR LYS ASN LEU
FORMUL   2  HOH    *118(H2 O)
HELIX    1  H1 ILE A    3  GLU A   11  1                                  9
HELIX    2  H2 LEU A   39  ILE A   50  1                                 12
HELIX    3  H3 LYS A   60  ARG A   80  1                                 21
HELIX    4  H4 ALA A   82  SER A   90  1                                  9
HELIX    5  H5 ALA A   93  MET A  106  1                                 14
HELIX    6  H6 GLU A  108  GLY A  113  5                                  6
HELIX    7  H7 THR A  115  GLN A  123  1                                  9
HELIX    8  H8 TRP A  126  ALA A  134  1                                  9
HELIX    9  H9 ARG A  137  GLN A  141  1                                  5
HELIX   10 H10 PRO A  143  THR A  155  1                                 13
SHEET    1  S1 4 GLY A  56  ILE A  58  0
SHEET    2  S1 4 ARG A  14  ASP A  20 -1  O  LEU A  15   N  ILE A  58
SHEET    3  S1 4 TYR A  24  ILE A  27 -1  O  TYR A  24   N  ASP A  20
SHEET    4  S1 4 HIS A  31  THR A  34 -1  N  HIS A  31   O  ILE A  27
```

# Protein Data Bank

Sample entry in PDB

**X Y Z**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| HETATM | 1311 | O | HOH | A | 166 | 39.473 | -8.320 | 18.352 | 1.00 | 12.66 | O |
| HETATM | 1312 | O | HOH | A | 167 | 40.661 | -1.625 | 3.146 | 1.00 | 26.18 | O |
| HETATM | 1313 | O | HOH | A | 168 | 45.808 | -7.180 | 1.965 | 1.00 | 6.69 | O |
| HETATM | 1314 | O | HOH | A | 169 | 48.014 | -0.841 | 0.963 | 1.00 | 20.51 | O |
| HETATM | 1315 | O | HOH | A | 170 | 37.070 | -18.992 | -0.228 | 1.00 | 24.62 | O |
| HETATM | 1316 | O | HOH | A | 171 | 47.674 | 0.600 | 28.073 | 1.00 | 24.92 | O |
| HETATM | 1317 | O | HOH | A | 172 | 24.914 | -23.325 | 4.060 | 1.00 | 27.79 | O |
| HETATM | 1318 | O | HOH | A | 173 | 37.436 | 1.541 | -8.076 | 1.00 | 25.60 | O |
| HETATM | 1319 | O | HOH | A | 175 | 21.473 | 0.044 | 12.578 | 1.00 | 49.22 | O |
| HETATM | 1320 | O | HOH | A | 176 | 41.618 | -8.847 | 20.042 | 1.00 | 12.50 | O |
| HETATM | 1321 | O | HOH | A | 177 | 55.648 | -1.163 | 27.564 | 1.00 | 50.85 | O |
| HETATM | 1322 | O | HOH | A | 178 | 44.310 | -13.062 | -1.772 | 1.00 | 28.63 | O |
| HETATM | 1323 | O | HOH | A | 179 | 21.105 | -20.149 | 4.804 | 1.00 | 22.47 | O |
| HETATM | 1324 | O | HOH | A | 180 | 35.498 | -5.517 | 22.505 | 1.00 | 22.52 | O |
| HETATM | 1325 | O | HOH | A | 181 | 35.182 | -7.197 | 29.056 | 1.00 | 36.84 | O |
| HETATM | 1326 | O | HOH | A | 182 | 47.378 | -15.737 | 28.584 | 1.00 | 43.98 | O |
| HETATM | 1327 | O | HOH | A | 184 | 29.866 | -10.719 | 23.430 | 1.00 | 39.76 | O |
| HETATM | 1328 | O | HOH | A | 185 | 38.214 | -15.827 | 20.622 | 1.00 | 33.83 | O |
| HETATM | 1329 | O | HOH | A | 186 | 40.794 | -19.490 | 20.666 | 1.00 | 29.74 | O |
| HETATM | 1330 | O | HOH | A | 187 | 45.407 | 0.094 | -6.590 | 1.00 | 28.07 | O |
| HETATM | 1331 | O | HOH | A | 188 | 53.213 | -9.478 | 19.700 | 1.00 | 37.14 | O |
| HETATM | 1332 | O | HOH | A | 190 | 45.016 | 5.850 | 15.793 | 1.00 | 29.76 | O |
| HETATM | 1333 | O | HOH | A | 191 | 46.234 | 7.935 | 13.619 | 1.00 | 47.62 | O |
| HETATM | 1334 | O | HOH | A | 192 | 50.975 | 0.435 | 9.753 | 1.00 | 42.44 | O |
| HETATM | 1335 | O | HOH | A | 194 | 50.670 | -10.434 | -3.124 | 1.00 | 21.98 | O |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ATOM | 32308 | P | | A | A1531 | -35.808 | 111.810 | 71.195 | 1.00149.38 | P |
| ATOM | 32309 | OP1 | | A | A1531 | -36.543 | 110.930 | 70.262 | 1.00 80.60 | O |
| ATOM | 32310 | OP2 | | A | A1531 | -34.421 | 112.237 | 70.863 | 1.00 79.54 | O |
| ATOM | 32311 | O5' | | A | A1531 | -36.706 | 113.117 | 71.371 | 1.00137.78 | O |
| ATOM | 32312 | C5' | | A | A1531 | -36.782 | 114.102 | 70.307 | 1.00140.15 | C |
| ATOM | 32313 | C4' | | A | A1531 | -37.439 | 115.378 | 70.797 | 1.00142.28 | C |
| OM | 32314 | O4' | | A | A1531 | -36.965 | 115.684 | 72.133 | 1.00142.54 | O |
| OM | 32315 | C3' | | A | A1531 | -37.122 | 116.620 | 69.973 | 1.00143.71 | C |
| OM | 32316 | O3' | | A | A1531 | -37.665 | 116.753 | 68.649 | 1.00146.81 | O |
| OM | 32317 | C2' | | A | A1531 | -37.276 | 117.753 | 70.980 | 1.00143.32 | C |
| OM | 32318 | O2' | | A | A1531 | -38.623 | 118.161 | 71.132 | 1.00142.77 | O |
| OM | 32319 | C1' | | A | A1531 | -36.782 | 117.083 | 72.272 | 1.00142.89 | C |
| OM | 32320 | N9 | | A | A1531 | -35.360 | 117.309 | 72.554 | 1.00 95.76 | N |
| OM | 32321 | C8 | | A | A1531 | -34.295 | 116.513 | 72.186 | 1.00 95.21 | C |
| OM | 32322 | N7 | | A | A1531 | -33.129 | 116.984 | 72.551 | 1.00 95.76 | N |
| OM | 32323 | C5 | | A | A1531 | -33.440 | 118.165 | 73.210 | 1.00 97.00 | C |
| OM | 32324 | C6 | | A | A1531 | -32.639 | 119.138 | 73.829 | 1.00 97.05 | C |
| OM | 32325 | N6 | | A | A1531 | -31.306 | 119.075 | 73.874 | 1.00 97.96 | N |
| OM | 32326 | N1 | | A | A1531 | -33.260 | 120.195 | 74.404 | 1.00 97.56 | N |
| OM | 32327 | C2 | | A | A1531 | -34.600 | 120.258 | 74.349 | 1.00 98.19 | C |
| OM | 32328 | N3 | | A | A1531 | -35.463 | 119.409 | 73.792 | 1.00 97.78 | N |
| OM | 32329 | C4 | | A | A1531 | -34.812 | 118.372 | 73.231 | 1.00 96.66 | C |
| R | 32330 | | | A | A1531 | | | | | |
| OM | 32331 | N | VAL | B | 7 | -19.397 | 100.525 | 10.485 | 1.00200.65 | N |
| OM | 32332 | CA | VAL | B | 7 | -20.293 | 99.359 | 10.728 | 1.00200.65 | C |
| OM | 32333 | C | VAL | B | 7 | -20.301 | 98.977 | 12.206 | 1.00200.65 | C |
| OM | 32334 | O | VAL | B | 7 | -21.119 | 99.472 | 12.983 | 1.00200.65 | O |
| OM | 32335 | CB | VAL | B | 7 | -19.848 | 98.133 | 9.890 | 1.00200.65 | C |
| OM | 32336 | CG1 | VAL | B | 7 | -20.768 | 96.948 | 10.160 | 1.00200.65 | C |
| OM | 32337 | CG2 | VAL | B | 7 | -19.859 | 98.485 | 8.409 | 1.00200.65 | C |
| OM | 32338 | N | LYS | B | 8 | -19.385 | 98.097 | 12.593 | 1.00150.41 | N |
| OM | 32339 | CA | LYS | B | 8 | -19.304 | 97.656 | 13.974 | 1.00150.41 | C |
| OM | 32340 | C | LYS | B | 8 | -18.666 | 98.735 | 14.847 | 1.00150.41 | C |
| OM | 32341 | O | LYS | B | 8 | -19.366 | 99.574 | 15.413 | 1.00150.41 | O |
| OM | 32342 | CB | LYS | B | 8 | -18.499 | 96.360 | 14.059 | 1.00 97.02 | C |
| OM | 32343 | CG | LYS | B | 8 | -18.979 | 95.269 | 13.102 | 1.00 97.02 | C |
| OM | 32344 | CD | LYS | B | 8 | -20.427 | 94.863 | 13.355 | 1.00 97.02 | C |
| OM | 32345 | CE | LYS | B | 8 | -20.836 | 93.721 | 12.432 | 1.00 97.02 | C |
| OM | 32346 | NZ | LYS | B | 8 | -22.236 | 93.250 | 12.662 | 1.00 97.02 | N |
| ATOM | 32347 | N | GLU | B | 9 | -17.339 | 98.717 | 14.942 | 1.00169.14 | N |
| ATOM | 32348 | CA | GLU | B | 9 | -16.600 | 99.684 | 15.754 | 1.00169.14 | C |
| ATOM | 32349 | C | GLU | B | 9 | -16.730 | 101.138 | 15.292 | 1.00169.14 | C |
| ATOM | 32350 | O | GLU | B | 9 | -16.709 | 102.060 | 16.113 | 1.00169.14 | O |
| ATOM | 32351 | CB | GLU | B | 9 | -15.118 | 99.288 | 15.808 | 1.00197.29 | C |

# Occupancy

Macromolecular crystals are composed of many individual molecules packed into a symmetrical arrangement.

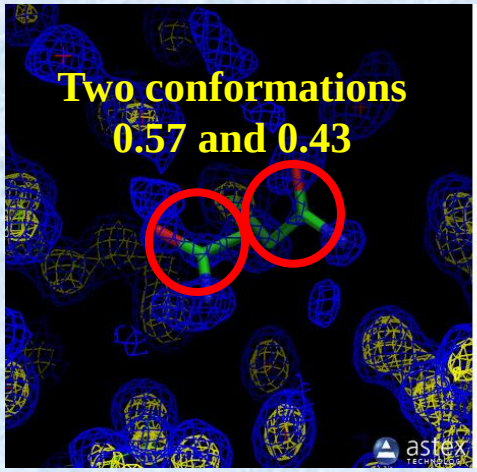
A sidechain on the surface may wag back and forth between several conformations, or a substrate may bind in two orientations in an active site, or a metal ion may be bound to only a few of the molecules.

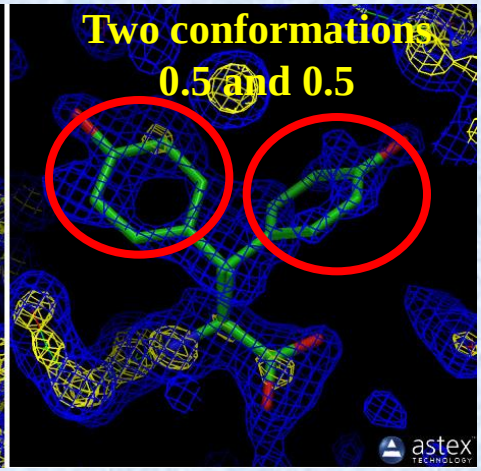
When building the atomic model of these portions, the occupancy can be used to estimate the amount of each conformation that is observed in the crystal.

For most atoms, the occupancy is given a value of 1, indicating that the atom is found in all of the molecules in the same place in the crystal.

However, if a metal ion binds to only half of the molecules in the crystal, the researcher will see a weak image of the ion in the electron density map, and can assign an occupancy of 0.5 in the PDB structure file for this atom.

The occupancy value is used to indicate the fraction of molecules that have each of the conformations. Two (or more) atom records are included for each atom, with occupancies like 0.5 and 0.5, or 0.4 and 0.6, or other fractional occupancies that sum to a total of 1.




Gln8 (0.57 and 0.43) and Tyr151 in 1A6M

```
ATOM     63  N   GLN A   8       5.404  13.203  22.532  1.00  8.42           N
ANISOU   63  N   GLN A   8     991   1372    837    -55    112    -21        N
ATOM     64  CA  GLN A   8       6.475  12.812  23.418  1.00  8.84           C
ANISOU   64  CA  GLN A   8    1086   1404    868     36    143     60        C
ATOM     65  C   GLN A   8       7.602  12.149  22.631  1.00  8.08           C
ANISOU   65  C   GLN A   8    1159   1168    744    -7[illegible]    175        C
ATOM     66  O   GLN A   8       8.769  12.399  22.918  1.00  [illegible].39           O
ANISOU   66  O   GLN A   8    1114   1262    812    [illegible]     7     -3        O
ATOM     67  CB AGLN A   8       5.987  11.822  24.520  0.57 1[illegible].03           C
ANISOU   67  CB AGLN A   8    1886   2182    884    [illegible]    38[illegible]    376        C
ATOM     68  CB BGLN A   8       5.948  11.968  24.580  0.43  [illegible].68           C
ANISOU   68  CB BGLN A   8     765   1824   1090    [illegible]    188    255        C
ATOM     69  CG AGLN A   8       7.030  11.303  25.500  0.57 16.30           C
ANISOU   69  CG AGLN A   8    2914   1636   1644   -[illegible]    -429    642        C
ATOM     70  CG BGLN A   8       6.967  12.094  25.68[illegible]  0.43 12.07           C
ANISOU   70  CG BGLN A   8    1548   1875   1164   -[illegible]    -193    395        C
ATOM     71  CD AGLN A   8       7.981  10.227  25.06[illegible]  0.57 15.[illegible]1           C
ANISOU   71  CD AGLN A   8    2461   1432   2038   -[illegible]    -802    629        C
ATOM     72  CD BGLN A   8       6.439  11.470  26.95[illegible]  0.43 14.43           C
ANISOU   72  CD BGLN A   8    2356   2084   1045   -[illegible]    -11     87        C
ATOM     73  OE1AGLN A   8       7.688   9.392  24.21[illegible]  0.57 19.[illegible]4           O
ANISOU   73  OE1AGLN A   8    3567   1731   2125   -[illegible]    -831    300        O
ATOM     74  OE1BGLN A   8       5.419  10.767  26.918  0.43 17.46           O
ANISOU   74  OE1BGLN A   8    2786   2908    939  -1[illegible]3    -341    822        O
ATOM     75  NE2AGLN A   8       9.219  10.114  25.607  0.57 2[illegible].38           N
ANISOU   75  NE2AGLN A   8    2518   2791   2813   -5[illegible]8    -105[illegible]   1739        N
ATOM     76  NE2BGLN A   8       7.067  11.762  28.084  0.43 1[illegible].03           N
ANISOU   76  NE2BGLN A   8    2245   1968   1117   -5[illegible]    -7[illegible]    225        N
ATOM     77  N   LEU A   9       7.271  11.269  21.697  1.00  [illegible].05           N
ANISOU   77  N   LEU A   9    1062   1121    877    [illegible]    [illegible]     81        N
ATOM     78  CA  LEU A   9       8.296  10.604  20.915  1.00  8.31           C
ANISOU   78  CA  LEU A   9    1085   1164    907    41[illegible]    [illegible]4    206        C
ATOM     79  C   LEU A   9       9.080  11.630  20.080  1[illegible]  7.19           C
ANISOU   79  C   LEU A   9    1008   1038    685     70     41      2        C
ATOM     80  O   LEU A   9      10.311  11.526  19.962  1.00  7.98           O
ANISOU   80  O   LEU A   9    1019   1097    915    232    143     83        O
ATOM     81  CB  LEU A   9       7.677   9.576  19.969  1.00  9.04           C
ANISOU   81  CB  LEU A   9    1440    983   1012    -45    222     79        C
ATOM     82  CG  LEU A   9       7.090   8.338  20.698  1.00 10.56           C
ANISOU   82  CG  LEU A   9    1658    982   1370    -49    291     80        C
ATOM     83  CD1 LEU A   9       6.290   7.498  19.721  1.00 12.91           C
ANISOU   83  CD1 LEU A   9    2111   1030   1763   -292    327   -155        C
ATOM     84  CD2 LEU A   9       8.189   7.510  21.315  1.00 16.75           C
ANISOU   84  CD2 LEU A   9    2279   1349   2735     -8   -215    792        C
ATOM     85  N   VAL A  10       8.391  12.589  19.483  1.00  6.70           N
ANISOU   85  N   VAL A  10     931    934    681     29     49     12        N
ATOM     86  CA  VAL A  10       9.019  13.628  18.670  1.00  6.25           C
ANISOU   86  CA  VAL A  10     822    859    692     15    113      1        C
```

# Temperature factor

If we were able to hold an atom rigidly fixed in one place, we could observe its distribution of electrons in an ideal situation.

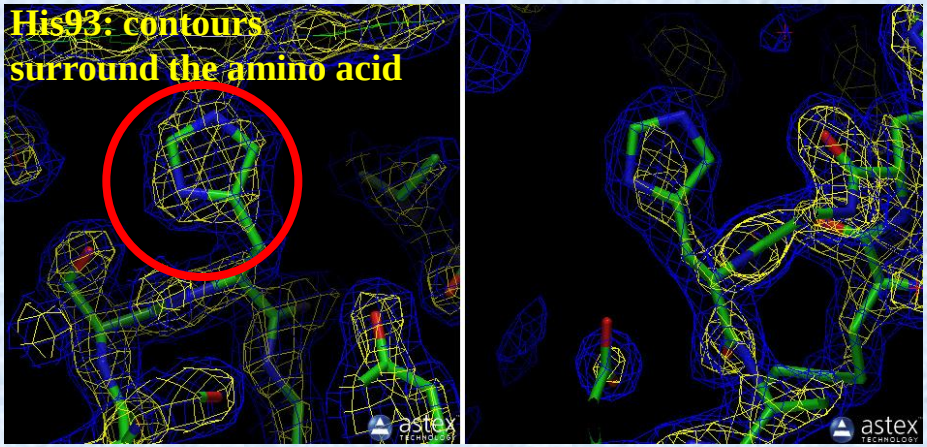
The image would be dense towards the center with the density falling off further from the nucleus.

The electrons usually have a wider distribution than this ideal. This may be due to vibration of the atoms, or differences between the many different molecules in the crystal lattice. The observed electron density will include an average of all these small motions, yielding a slightly smeared image of the molecule.

These motions, and the resultant smearing of the electron density, are incorporated into the atomic model by a B-value or temperature factor.

Values under 10 create a model of the atom that is very sharp, indicating that the atom is not moving much and is in the same position in all of the molecules in the crystal.

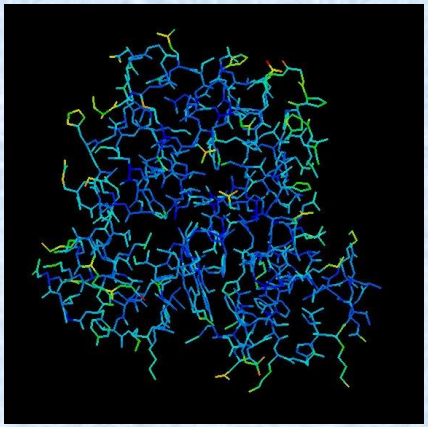
Values greater than 50 or so indicate that the atom is moving so much that it can barely been seen. This is often the case for atoms at the surface of proteins, where long sidechains are free to wag in the surrounding water.




```
ATOM   737  ND1 HIS A  93     11.243  30.390   4.397  1.00 12.73          N
ATOM   738  CD2 HIS A  93     12.005  28.593   3.361  1.00 20.40          C
ATOM   739  CE1 HIS A  93     12.419  30.059   4.926  1.00 15.17          C
ATOM   740  NE2 HIS A  93     12.897  28.963   4.332  1.00 16.09
```

```
ATOM   642  ND1 HIS A  81     -2.541  24.541  20.570  1.00 72.63          N
ATOM   643  CD2 HIS A  81     -0.874  23.158  20.970  1.00 54.37          C
ATOM   644  CE1 HIS A  81     -3.006  23.416  21.159  1.00 53.46          C
ATOM   645  NE2 HIS A  81     -2.022  22.561  21.396  1.00 73.87          N
```

High values, indicating lots of motion, are in red and yellow, and low values are in blue.



Summary | **Sequence** | Annotations | Seq. Similarity | 3D Similarity | Literature | Biol. & Chem. | Methods | Geometry | Links

# STRUCTURE OF BACTERIOPHAGE T4 LYSOZYME REFINED AT 1.7 ANGSTROMS RESOLUTION

# 2LZM

**FASTA** | **Sequence & DSSP** | **Image**

Polymer 1
Length: 164 residues
Chain Type: polypeptide(L)
Reference: **UniProtKB** **P00720**

**Display Parameters**

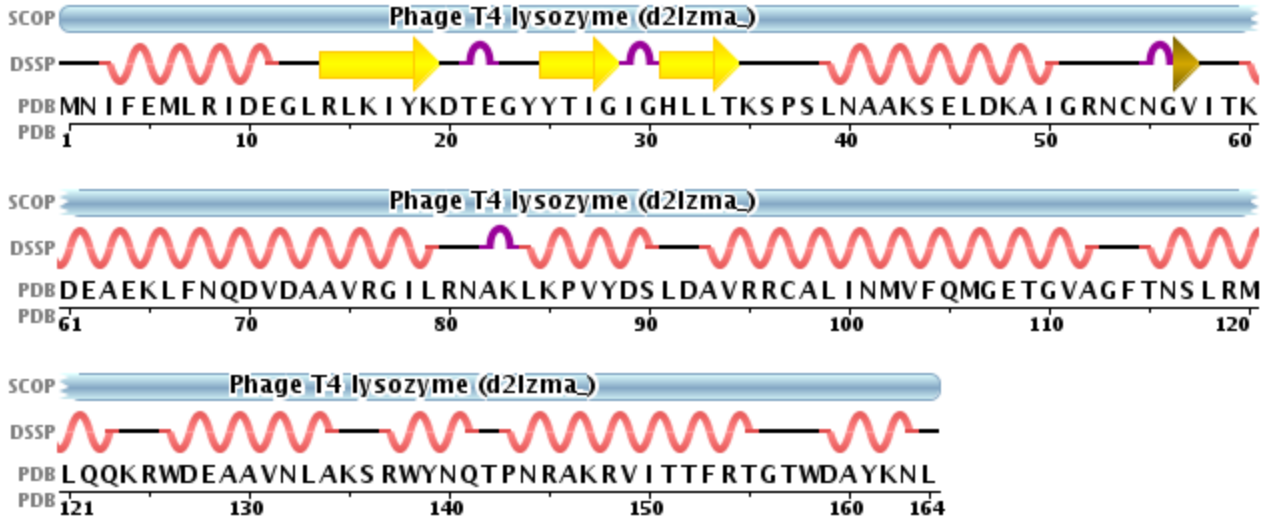
Currently displayed: **SEQRES sequence.**
**Display external (UniProtKB) sequence**
Mouse over an annotation to see details.

**Annotations** *Add Annotations*

Select

Domain Assignment:**SCOP** d2lzma_ Phage T4 lysozyme: 164 residues
**[hide] [reference]**

Secondary Structure:**DSSP** 66% helical (10 helices; 109 residues)
**[hide] [reference]** 9% beta sheet (4 strands; 15 residues)



Summary | Sequence | **Annotations** | Seq. Similarity | 3D Similarity | Literature | Biol. & Chem. | Methods | Geometry | Links

# STRUCTURE OF BACTERIOPHAGE T4 LYSOZYME REFINED AT 1.7 ANGSTROMS RESOLUTION

# 2LZM

## Domain Annotation: SCOP Classification (version 1.75)

Hide

| Domain Info | Class | Fold | Superfamily | Family | Domain | Species |
|---|---|---|---|---|---|---|
| d2lzma_ | Alpha and beta proteins (a+b) | Lysozyme-like | Lysozyme-like | Phage lysozyme | Phage T4 lysozyme | Bacteriophage T4 [TaxId: 10665] |

## Domain Annotation: CATH Classification (version 3.4.0)

Hide

| Domain | Class | Architecture | Topology | Homology |
|---|---|---|---|---|
| 2lzmA00 | Mainly Alpha | Orthogonal Bundle | Lysozyme | |

## Protein Family Annotation: PFAM Classification

Hide

| Chain | PFAM Accession | PFAM ID | Description | Type | Clan ID |
|---|---|---|---|---|---|
| A | PF00959 | Phage_lysozyme | Phage lysozyme | Domain | |

## Gene Product Annotation: GO Terms

Hide

| Polymer | Molecular Function | Biological Process | Cellular Component |
|---|---|---|---|
| T4 LYSOZYME (2LZM:A) | • lysozyme activity<br>• catalytic activity<br>• hydrolase activity<br>• hydrolase activity, acting on glycosyl bonds | • metabolic process<br>• peptidoglycan catabolic process<br>• cell wall macromolecule catabolic process<br>• cytolysis<br>• defense response to bacterium | • none |

## Structural Biology Knowledgebase Data

Hide

Information from the Structural Biology Knowledgebase

- Protein Targets from TargetDB: 0 targets
- Related Biological Annotations: >21 annotations
- Related Clones from PSI:Biology Materials Repository: 0 clones
- Related Protocols from PepcDB: 0 protocols

Summary | Sequence | Annotations | **Seq. Similarity** | 3D Similarity | Literature | Biol. & Chem. | Methods | Geometry | Links

# STRUCTURE OF BACTERIOPHAGE T4 LYSOZYME REFINED AT 1.7 ANGSTROMS RESOLUTION

# 2LZM

**Cluster # 3648 Members. Sequence cutoff: 100% Cluster size: 9 sequences**

Select for download / view details OR: Select two chains for comparison

Download | Details | --- Select Comparison Method --- | Submit

| Rank | PDB id | Entity id | Chains | Description | Details | Taxonomy | EC Number |
|---|---|---|---|---|---|---|---|
| 1 | 3FA0 | 1 | A | Lysozyme | | 10665 | 3.2.1.17 |
| 2 | 3LZM | 1 | A | T4 LYSOZYME | | 10665 | 3.2.1.17 |
| 3 | 4LZM | 1 | A | T4 LYSOZYME | | 10665 | 3.2.1.17 |
| 4 | 7LZM | 1 | A | T4 LYSOZYME | | 10665 | 3.2.1.17 |
| 5 | 5LZM | 1 | A | T4 LYSOZYME | | 10665 | 3.2.1.17 |
| 6 | 6LZM | 1 | A | T4 LYSOZYME | | 10665 | 3.2.1.17 |
| 7 | 2LZM | 1 | A | T4 LYSOZYME | | 10665 | 3.2.1.17 |
| 8 | 1T6H | 1 | A | Lysozyme | | 10665 | 3.2.1.17 |
| 9 | 1LYD | 1 | A | T4 LYSOZYME | | 10665 | |

**Sequence Clusters for the Sequence Entities in PDB 2LZM**

Entity #1: Chains: A - T4 LYSOZYME protein, length: 164 [Blast]

| Cluster Sequence Similarity Cutoff | Rank | Nr. of chains in Cluster | Cluster Nr. |
|---|---|---|---|
| 100% | 7 | 9 | 3648 |
| 95% | 199 | 522 | 2 |
| 90% | 200 | 530 | 2 |
| 70% | 200 | 530 | 6 |
| 50% | 200 | 530 | 8 |
| 40% | 200 | 530 | 9 |
| 30% | 200 | 530 | 14 |

Summary | Sequence | Annotations | Seq. Similarity | **3D Similarity** | Literature | Biol. & Chem. | Methods | Geometry | Links

**STRUCTURE OF BACTERIOPHAGE T4 LYSOZYME REFINED AT 1.7 ANGSTROMS RESOLUTION**

**2LZM**

Display Files
Download Files
Share this Page

## Structural Similarities for the Entities in PDB 2LZM

The following structural similarities have been found using the jFATCAT-rigid algorithm. In order to reduce the number of hits, a 40% sequence identity clustering has been applied and a representative chain is taken from each cluster.

### Entity #1: Chains: A

**Description**: T4 LYSOZYME protein
**Length**: 164

**No structure alignment results are available for 2LZM.A explicitly.**

It is represented by chain **1SX7.A** which has more than 95% sequence similarity.

**Click here to hide the structure comparison results for representative 1SX7.A**

### Results options — Hide

Currently viewing only **significant results** (P-value < 0.001).
**Show all available results**

### Update Status — Hide

The data on this page are up to date as of: Oct 04 2011

### 1SX7.A (chain 1) vs. representatives of other sequence clusters (chain 2)

| Rank | Results | Chain 2 | Title | P-value | Score | Rmsd | Len1 | Len2 | %ID | %Cov1 | %Cov2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | view | 2QB0.B | TELSAM domain - Lysozym | 0.0 | 477.58 | 1.29 | 164 | 241 | 96 | 98 | 67 |
| 2 | view | 2O7A.A | Lysozyme | 0.0 | 301.62 | 0.47 | 164 | 116 | 97 | 59 | 83 |
| 3 | view | 3SN6.R | Lysozyme, Beta-2 adrene | 1.14E-10 | 457.36 | 1.62 | 443 | 164 | 97 | 36 | 97 |
| 4 | view | 1WTH.A | Tail-associated lysozyme | 5.0E-10 | 442.54 | 1.16 | 164 | 571 | 43 | 99 | 28 |
| 5 | view | 2RH1.A | beta-2-adrenergic recept | 1.55E-9 | 471.90 | 2.22 | 164 | 443 | 96 | 100 | 37 |
| 6 | view | 3HDE.A | Lysozyme | 2.48E-8 | 270.98 | 2.99 | 164 | 165 | 17 | 83 | 82 |

### Table Legend — Hide

- **Rank**: current row position. Changes with different sorting orders and filter rules
- **Chain 2**: PDB chain name
- **Title**: Protein chain description
- **P-value**: P-value of this alignment (FATCAT) (default sorted by this)
- **Score**: Raw alignment score (FATCAT)
- **RMSD**: RMSD value of the

Summary | Sequence | Annotations | Seq. Similarity | 3D Similarity | **Literature** | Biol. & Chem. | Methods | Geometry | Links

Display Files

**Related articles from PubMedCentral (provided by BioLit)** Hide

**FOLD-RATE: prediction of protein folding rates from amino acid sequence**

M. Michael Gromiha, A. Mary Thangakani, S. Selvaraj

**Other PDB IDs found in the above articles**

| Image | PDB ID | Structure Title | Sequence Similarity |
|---|---|---|---|
| | 1A6N | DEOXY-MYOGLOBIN, ATOMIC RESOLUTION | - |
| | 1AON | CRYSTAL STRUCTURE OF THE ASYMMETRIC CHAPERONIN COMPLEX GROEL/GROES/(ADP)7 | - |
| | 1BNI | BARNASE WILDTYPE STRUCTURE AT PH 6.0 | - |
| | 1CBI | APO-CELLULAR RETINOIC ACID BINDING PROTEIN I | - |
| | 1CEI | STRUCTURE DETERMINATION OF THE COLICIN E7 IMMUNITY PROTEIN (IMME7) THAT BINDS SPECIFICALLY TO THE DNASE-TYPE COLICIN E7 AND INHIBITS ITS BACTERIOCIDAL ACTIVITY | - |

Summary | Sequence | Annotations | Seq. Similarity | 3D Similarity | Literature | **Biol. & Chem.** | Methods | Geometry | Links

STRUCTURE OF BACTERIOPHAGE T4 LYSOZYME REFINED AT 1.7 ANGSTROMS RESOLUTION

**2LZM**

Display Files
Download Files

Biology a

Structu

Structur

Keywords
Text

Polymeri

Chain A

Descriptior
Nonstanda
Nonstanda
Polymer Ty
Formula W
Source Me

## CSA entry for 2lzm

| | | | |
|---|---|---|---|
| **Title:** | Hydrolase (o-glycosyl) | | |
| **Compound:** | Lysozyme (e.c.3.2.1.17) | | |
| **Mutant:** | No | | |
| **UniProt/Swiss-Prot:** | P00720-LYCV_BPT4 | **EC Class:** | 3.2.1.17 |
| **Other CSA Entries:** | Homologues of 2lzm<br>Entries for UniProt/Swiss-Prot: P00720<br>Entries for EC: 3.2.1.17 | **Other Databases:** | PDB entry: 2lzm<br>PDBsum entry: 2lzm<br>UniProt/Swiss-Prot: P00720<br>IntEnz entry: 3.2.1.17<br>KEGG entry: 3.2.1.17<br>EzCatDB entry: S00021 |

**Sites:**

Catalytic Site (Get help with this section)

Found by: PsiBLAST alignment on 206l

| Residue | Chain | Number | UniProt number | Functional part |
|---|---|---|---|---|
| GLU | A | 11 | 11 | Sidechain |
| ASP | A | 20 | 20 | Sidechain |

Use the check-boxes to select site(s) to view on the 3D structure in RasMol, and press Display Protein

**Sequence:**

```
Catalytic residues are indicated in red.

Chain: A
        10        20        30        40        50        60        70        80        90       100
         |         |         |         |         |         |         |         |         |         |
MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRNCNGVITKDEAEKLFNQDVDAAVRGILRNAKLKPVYDSLDAVRRCALI

       110       120       130       140       150       160
         |         |         |         |         |         |
NMVFQMGETGVAGFTNSLRMLQQKRWDEAAVNLAKSRWYNQTPNRAKRVITTFRTGTWDAYKNL
```

| | | | |
|---|---|---|---|
| 42742 | P | Defense Response to Bacterium | Reactions Triggered in Response to the Presence of a Bacterium That Act to Protect the Cell or Organism. |

Summary | Sequence | Annotations | Seq. Similarity | 3D Similarity | Literature | Biol. & Chem. | **Methods** | Geometry | Links

# X-RAY DIFFRACTION

# 2LZM

**Materials and Methods page**

## Crystal Data

Unit Cell

| Length | (Å) | Angle | (°) |
|---|---|---|---|
| a = | 61.2 | | |
| b = | 61.2 | | |
| c = | 96.8 | | |

Space Group

**Space Group Name:**

## Diffraction

## Refinement

Refinement Statistics

**reflnsShellList**

**R-Factor(Observed)**

RMS Deviations

| Parameter Type | Deviation from Ideal |
|---|---|
| t_angle_deg | 2.7 |
| t_bond_d | 0.019 |

Number of Non-Hydrogen Atoms Used in Refinement

| | |
|---|---|
| Protein Atoms | 1309 |
| Nucleic Acid Atoms | 0 |
| Heterogen Atoms | 0 |
| Solvent Atoms | 118 |

## Software and Computing

Computing

| | |
|---|---|
| Structure Refinement | TNT |

Software

| | |
|---|---|
| refinement | TNT |

Summary | Sequence | Annotations | Seq. Similarity | 3D Similarity | Literature | Biol. & Chem. | Methods | **Geometry** | Links

## Bond Length

| Bond Type | Chain Id | Tot Num | Cal Ave | Cal StdDev | Std Val | Std StdDev | Minimum | Maximum |
|---|---|---|---|---|---|---|---|---|
| C-N | A | 160 | 1.33 | 0.020 | 1.329 | 0.014 | 1.27 | 1.38 |
| C-N(P) | A | 3 | 1.33 | 0.014 | 1.341 | 0.016 | 1.31 | 1.34 |
| C-O | A | 164 | 1.23 | 0.017 | 1.231 | 0.02 | 1.19 | 1.27 |
| CA-C | A | 153 | 1.53 | 0.020 | 1.525 | 0.021 | 1.47 | 1.58 |
| CA-C(G) | A | 11 | 1.53 | 0.014 | 1.516 | 0.018 | 1.51 | 1.56 |
| CA-CB | A | 108 | 1.53 | 0.025 | 1.53 | 0.02 | 1.43 | 1.61 |
| CA-CB(A) | A | 15 | 1.53 | 0.010 | 1.521 | 0.033 | 1.51 | 1.55 |
| CA-CB(I,T,V) | A | 30 | 1.55 | 0.021 | 1.54 | 0.027 | 1.51 | 1.61 |
| N-CA | A | 150 | 1.45 | 0.018 | 1.458 | 0.019 | 1.40 | 1.50 |

## Dihedral Angle

| Dihedral Angle | Chain Id | Tot Num | Cal Ave | Cal StdDev | Std Val | Std StdDev | Minimum | Maximum |
|---|---|---|---|---|---|---|---|---|
| Chi1 g(+) | A | 78 | -71.67 | 18.724 | -66.7 | 15.0 | -115.50 | -3.60 |
| Chi1 g(-) | A | 18 | 60.64 | 29.323 | 64.1 | 15.7 | 0.70 | 116.10 |
| Chi1 trans | A | 42 | 184.84 | 16.233 | 183.6 | 16.8 | 144.40 | 237.90 |
| Omega | A | 163 | 179.68 | 3.595 | 180 | 5.8 | 169.30 | 189.80 |
| Phi | A | 53 | -71.51 | 72.347 | -65.3 | 11.9 | -157.70 | 88.50 |
| Phi helix | A | 109 | -63.55 | 17.101 | -65.3 | 11.9 | -106.90 | 77.20 |
| Phi(P) | A | 1 | -69.50 | 0.000 | -65.4 | 11.2 | -69.50 | -69.50 |
| Psi | A | 46 | 94.20 | 72.308 | -39.4 | 11.3 | -48.60 | 175.00 |
| Psi helix | A | 109 | -38.07 | 21.279 | -39.4 | 11.3 | -61.30 | 157.70 |
| Psi(G) | A | 8 | -26.81 | 76.181 | -39.4 | 11.3 | -177.40 | 31.70 |

Save Dihedral Angle Summary in: ◉ CSV (Excel) Format

Save Report

Summary | Sequence | Annotations | Seq. Similarity | 3D Similarity | Literature | Biol. & Chem. | Methods | Geometry | Links

STRUCTURE OF BACTERIOP
AT 1.7 ANGSTROMS RESOL

External Links

## STRUCTURE SUMMARY

- Protein Databank in Europe (PDBe)
- Protein Data Bank Japan (wwPDB Partn
- PSI Structural Biology Knowledgebase (
- Protein Interfaces, Surfaces and Assemb
- Molecular Modeling DataBase (NCBI/Ent
- PDBsum
- Jena Library
- PDBWiki
- Proteopedia
- OCA Browser (OCA)

## STRUCTURE FEATURES

- Homology derived Secondary Structure
- Analysis of Ligand-Protein Contacts (LP
- Analysis of interatomic Contacts of Struc
- Computed Atlas of Surface Topography o
- Guassian Network Model (GNM)
- HIV Sequence/Structure Function Analyzer

## LIGAND FEATURES

- BindingDB : *No external link available*
- Ligand-Expo
- Chem-BLAST
- PubChem
- DrugBank

## STRUCTURE CLASSIFICATION AND COMPARISON

- Structural Classification of Proteins (SCOP)
- Protein Structure Classification (CATH)
- Vector Alignment Search Tool (VAST)
- Flexible structure AlignmenT by Chaining Aligned fragment pairs allowing Twists (FATCAT)
- DALI
- SUPERFAMILY

## SECONDARY STRUCTURE

- Secondary Structure Assignments (DSSP)

## EXPERIMENTAL DATA

*NO EXTERNAL LINK IN EXPERIMENTAL DATA*

## BIOLOGICAL DETAILS

- CSA
- IEDB : *No external link available*

## PATHWAYS

- METACYC : *No external link available*

## PROTEIN MOTIONS

- Molecular Movements Database (MMD)

## STEREOCHEMICAL QUALITY

- WHAT_CHECK (WHAT IF)
- PROCHECK

Adva
ID a
Stru
D
M
S
C
M

**Advanced Search Interface**

Macromolecule Type

*Search based on whether the structure contains chains of certain molecule types (e.g. protein vs. DNA)*

Contains Protein Yes

Result Count

**2064 Structure Hits** | 1038 Citations | 409 Ligand Hits

**Query Parameters:** Query Details | Save Query to MyPDB

Chain Type: there is a Protein and a DNA chain but not any RNA or Hybrid and Homologue Removal - 40% Identity Cutoff

**Query Refinements** Hide

**Organism**
- Homo sapiens (538)
- Escherichia coli (248)
- Mus musculus (117)
- Geobacillus stearothermophilus (90)
- Saccharomyces cerevisiae (70)
- Escherichia coli K-12 (67)
- Sulfolobus solfataricus P2 (57)
- Other (874)

**Taxonomy**
- Eukaryota (942)
- Bacteria (715)
- Viruses (262)
- Archaea (152)
- Other (37)
- Unassigned (21)

**Experimental Method**
- X-RAY (1951)
- Solution NMR (111)
- Electron Microscopy (2)

**X-Ray Resolution**
- less than 1.5 Å (18)
- 1.5 - 2.0 Å (340)
- 2.0 - 2.5 Å (634)
- 2.5 - 3.0 Å (621)
- 3.0 and more Å (338)
- more choices...

**Release Date**
- before 2000 (395)
- 2000 - 2005 (505)
- 2005 - 2010 (743)
- 2010 - today (421)
- this year (154)
- this month (2)
- more choices...

**Polymer Type**
- Protein/DNA (2064)

**Enzyme Classification**
- 2: Transferases (485)
- 3: Hydrolases (308)
- 4: Lyases (83)
- 5: Isomerases (40)
- 1: Oxidoreductases (23)
- 6: Ligases (13)

**SCOP Classification**
- All alpha proteins (564)
- Alpha and beta proteins (a+b) (445)
- Alpha and beta proteins (a/b) (292)
- Multi-domain proteins (alpha an ... (207)
- All beta proteins (180)
- Small proteins (86)
- Coiled coil proteins (31)
- Other (11)

# Protein visualization

**Pymol**

**Rasmol**

**Jmol**

**KING**

**Webmol**

**SWISS-PDB viewer**

# 3D structure prediction

PDB Status (22 March 2017)

Proteins: **127,823**

Amino acid sequences are stored in Protein sequence database (SWISS-PROT /EMBL)

Number of available sequences: **~78 million**

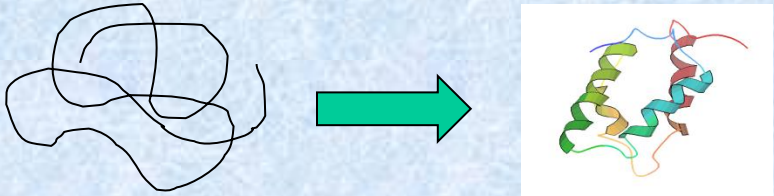
(increasing exponentially)

**Approximate cost**

**1 structure: $10,000-15,000**

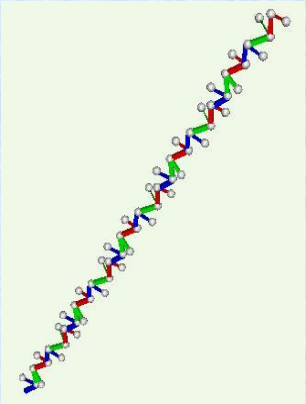
**Duration: 2 months – 1 year**

**Bioinformatics**

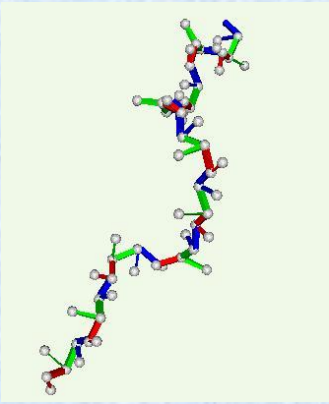


Theme: Deciphering the native conformation of a protein (3D structure) from its amino acid sequence → Protein Folding Problem.

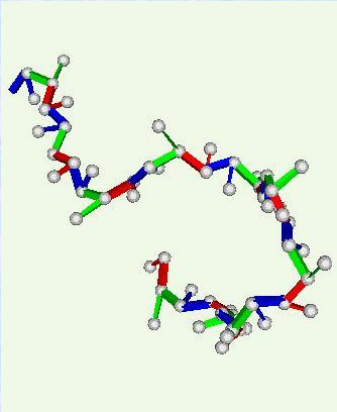
# Protein folding problem



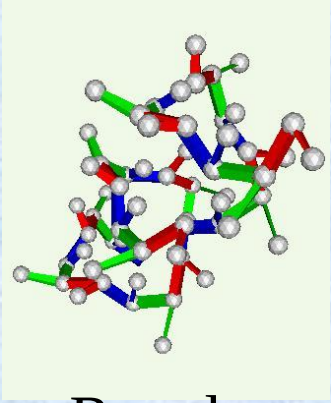
Elongated chain



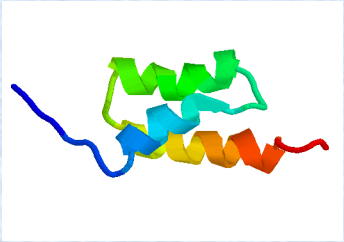
Distortion



Bends



Regular shape

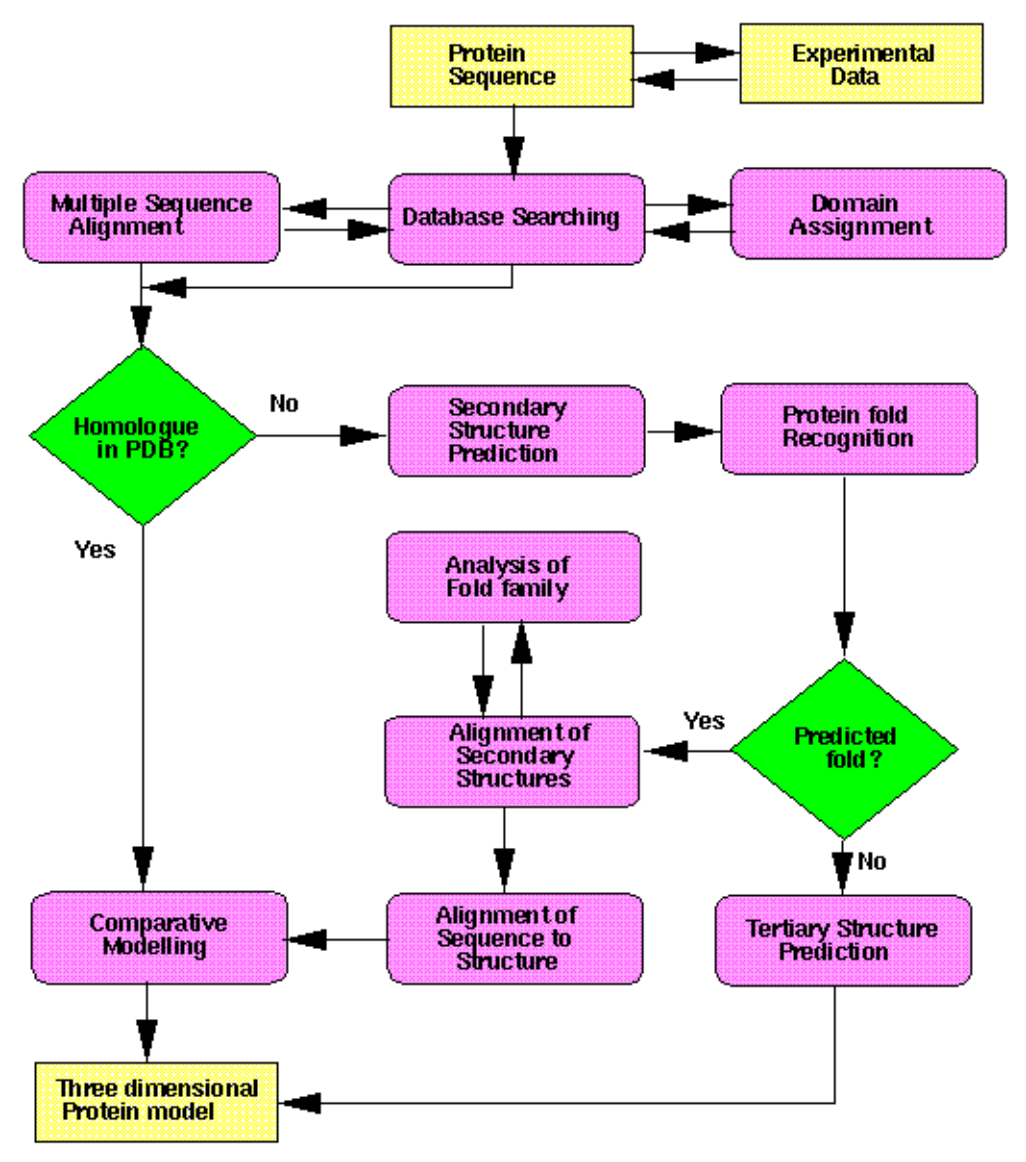


3D structure

# 3D structure prediction

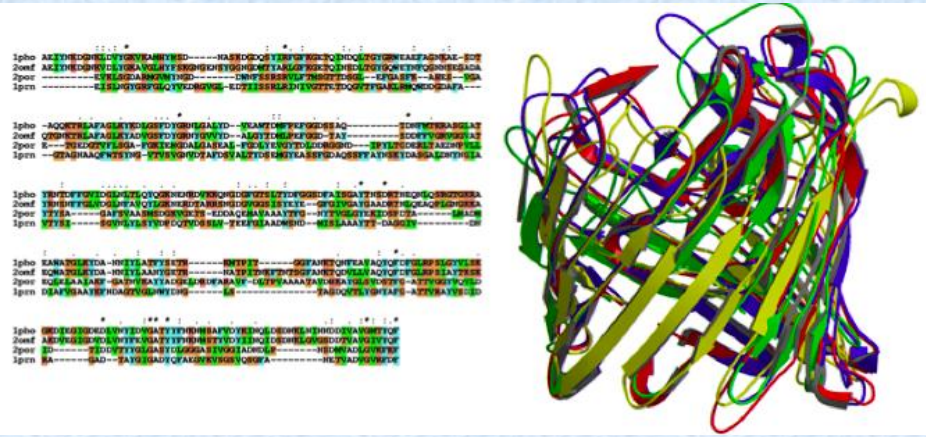
**Critical Assessment of Structure Prediction of Proteins (CASP)**

- Homology modeling
- ab initio method (energetic approach)
- Fold recognition



# Protein Homology Modelling

**Prediction of a 3D structure of a protein from its sequence with an accuracy that is comparable to the best results achieved experimentally.**



**Allow users to safely use rapidly generated *in silico protein models in all the contexts***

- **Structure-based drug design**
- **Analysis of protein function**
- **Interactions**
- **Antigenic behavior and**
- **Rational design of proteins with increased stability or novel functions.**

**Protein modeling is the only way to obtain structural information if experimental techniques fail.**

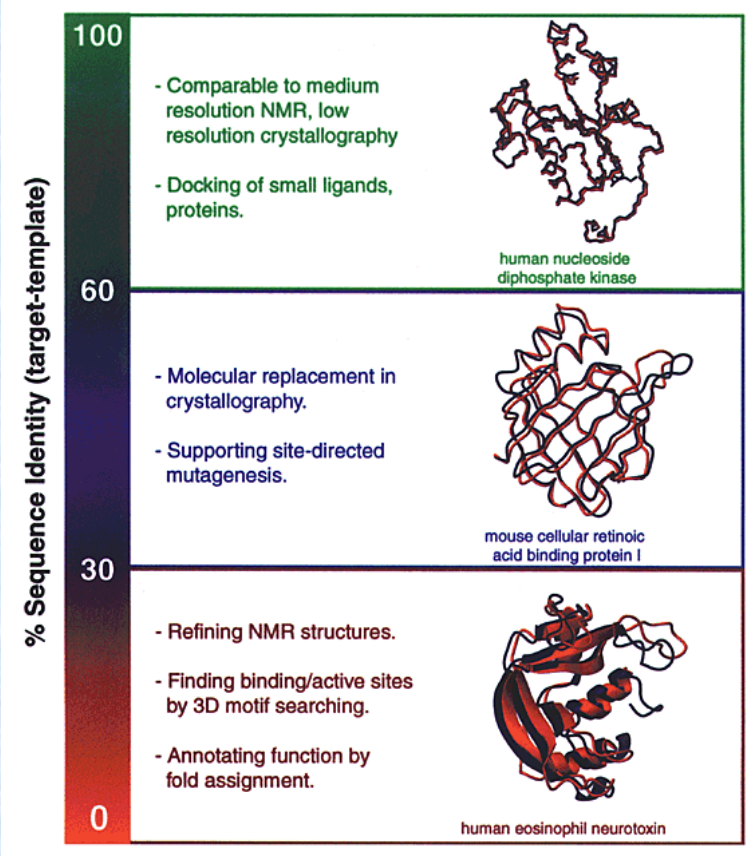
# Protein Sequence ➡ Structure

**The structure of a protein is uniquely determined by its amino acid sequence (Epstein, Goldberger, and Anfinsen, 1963).**

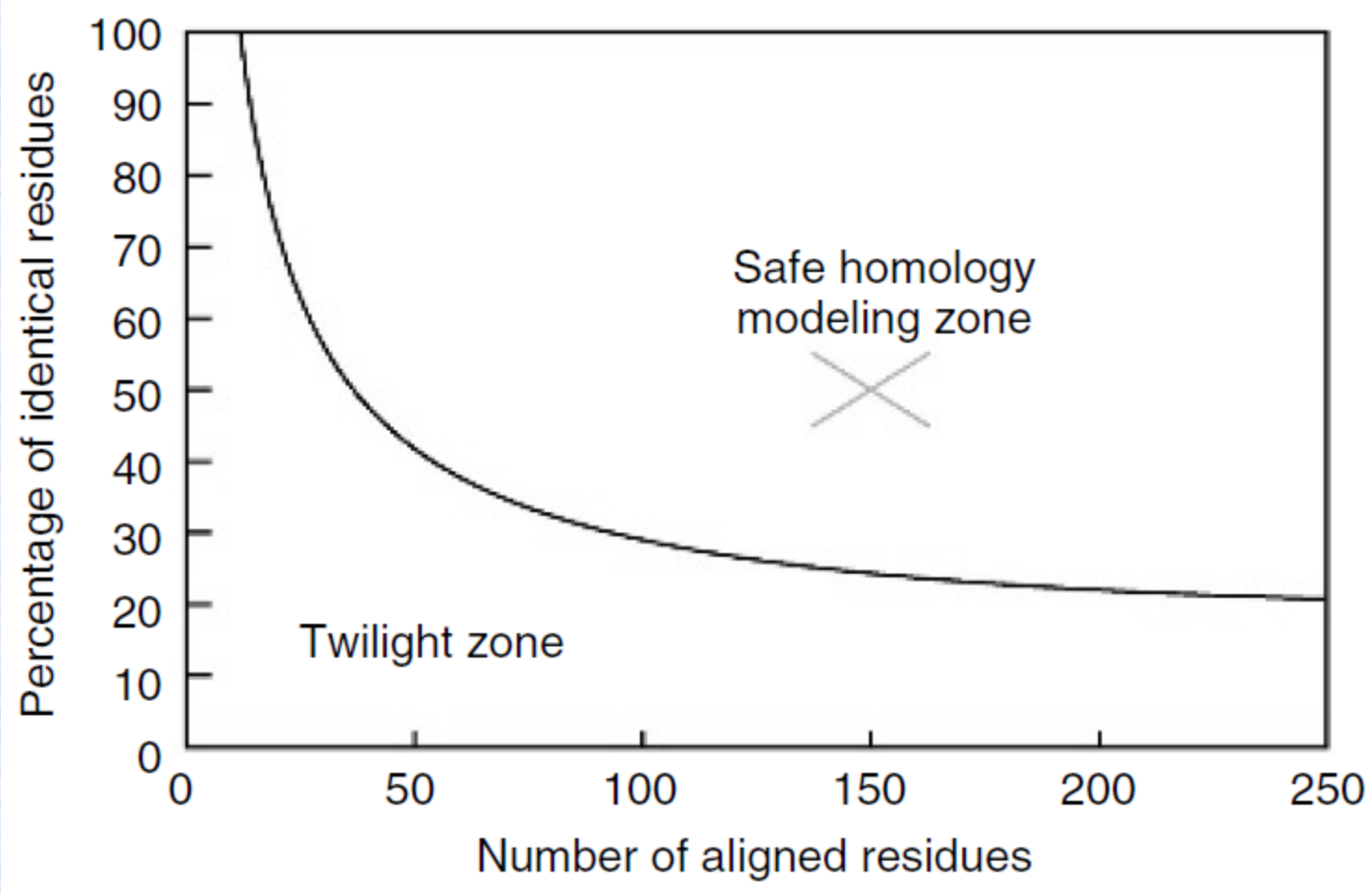
**Knowing the sequence should, at least in theory, suffice to obtain the structure.**

**During evolution, the structure is more stable and changes much slower than the associated sequence, so that similar sequences adopt practically identical structures, and distantly related sequences still fold into similar structures.**

**As long as the length of two sequences and the percentage of identical residues fall in the region marked as "safe," the two sequences are practically guaranteed to adopt a similar structure.**

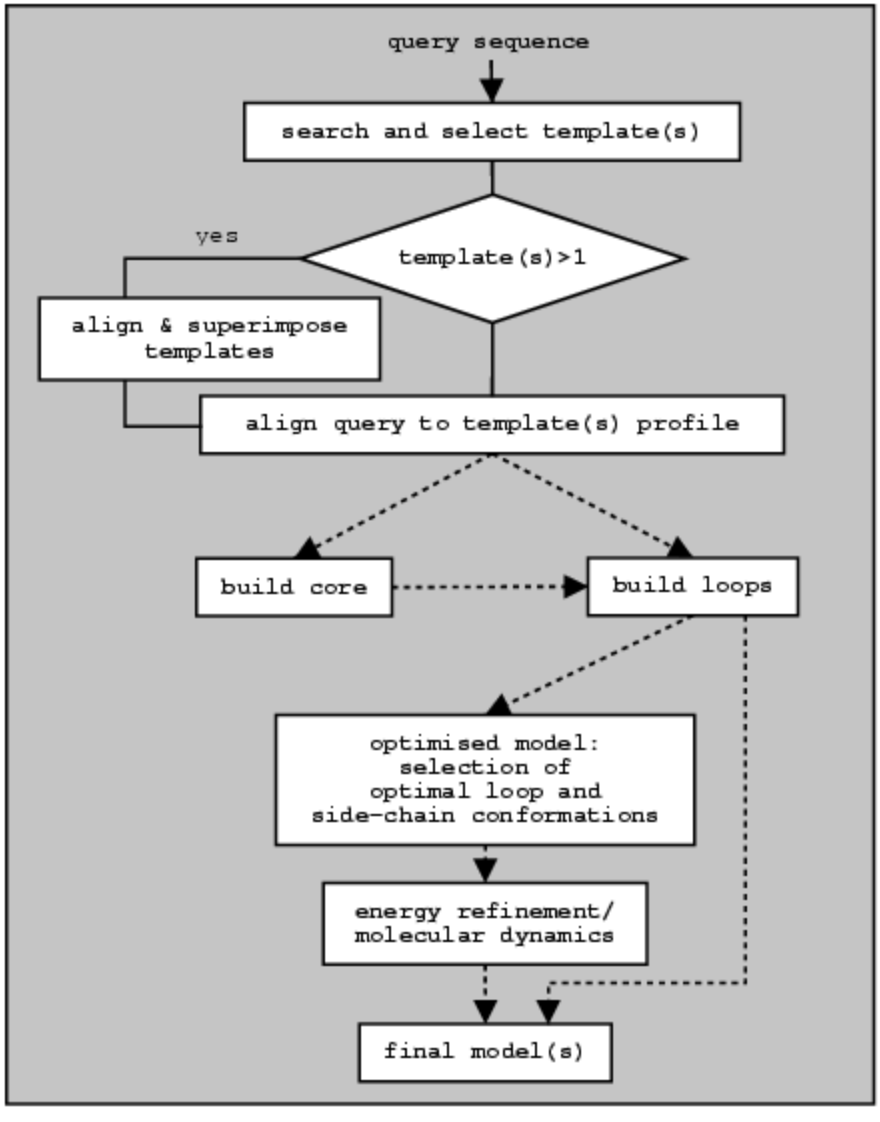



Sali, A. & Kuriyan, J. *Trends Biochem. Sci.* **22**, M20–M24 (1999)

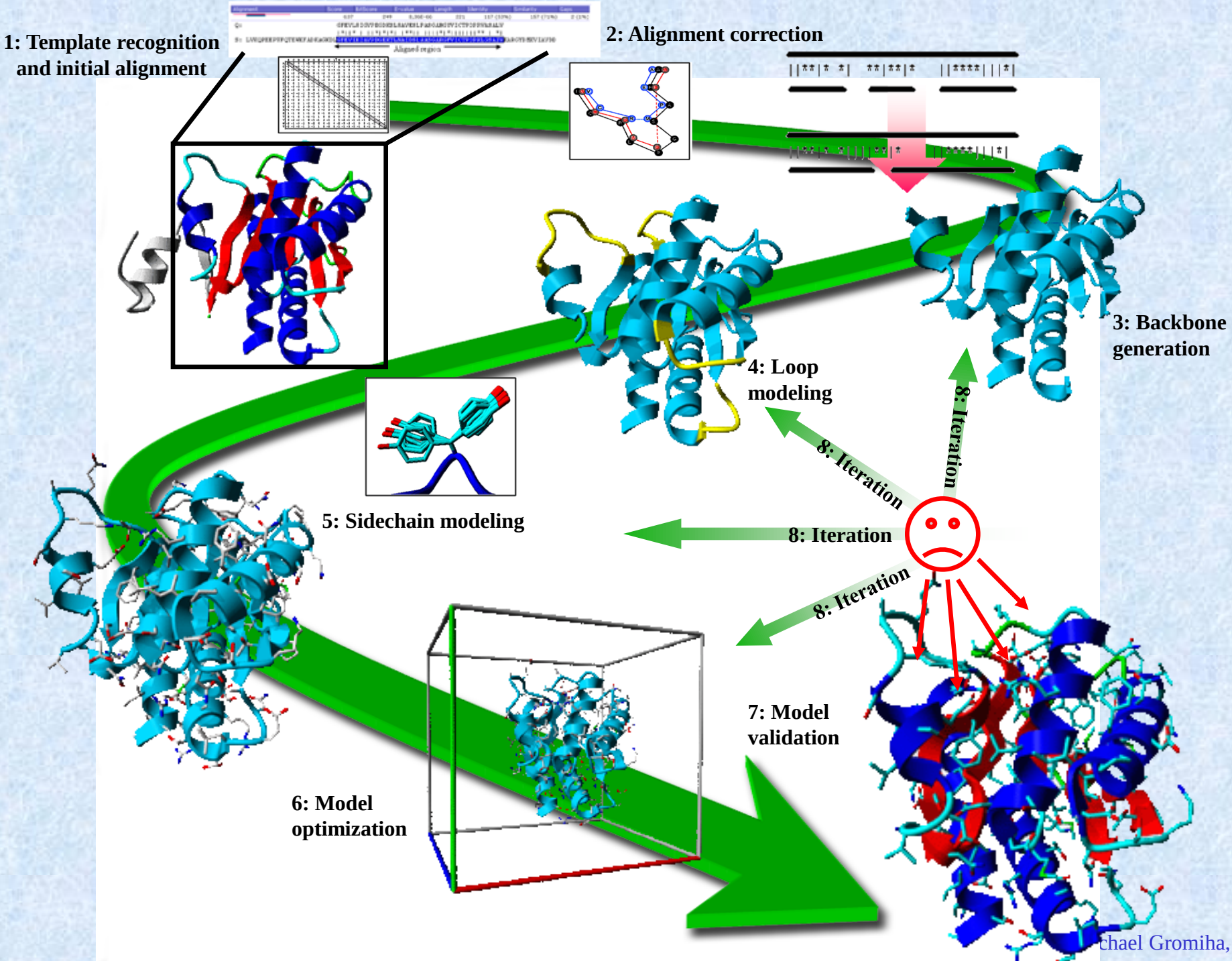



**The two zones of sequence alignments. Two sequences are practically guaranteed to fold into the same structure if their length and percentage sequence identity fall into the region marked as "safe." An example of two sequences with 150 amino acids, 50% of which are identical, is shown (gray cross).**

# Homology Modeling flowchart




1. Template recognition and initial alignment
2. Alignment correction
3. Backbone generation
4. Loop modeling
5. Side-chain modeling
6. Model optimization
7. Model validation


1: Template recognition and initial alignment
2: Alignment correction
3: Backbone generation
4: Loop modeling
5: Sidechain modeling
6: Model optimization
7: Model validation
8: Iteration
8: Iteration
8: Iteration
8: Iteration

# 1. Template recognition and initial alignment

- Search the related protein sequences (templates) to the target sequence in any structural database of proteins
- The accuracy of model depends on the selection of proper template
- FASTA and BLAST from EMBL-EBI and NCBI can be used
- Gives a probable set of templates but the final one is not yet decided
- After intial aligments and finding structurally conserved regions among templates, we choose the final template

- Search with sequence
  - Blast
  - Psi-Blast
  - Fold recognition methods
- Use biological information
- Functional annotation in databases
- Active site/motifs

# 1. Template recognition and initial alignment

- When two or more reference protein structures are available
- Establish structural guidelines for the family of proteins under consideration
- Regions, which are structurally conserved or constant among all the reference proteins
- Target protein is supposed to assume the same conformation in conserved regions

**Structurally Conserved Regions**

- **SCRs are region in all proteins of a particular family that are nearly identical in structures.**
- **Tend to be at inner cores of the proteins**
- **Usually contains alpha-helices and beta sheets**
- **No SCR can span more than one secondary structure**

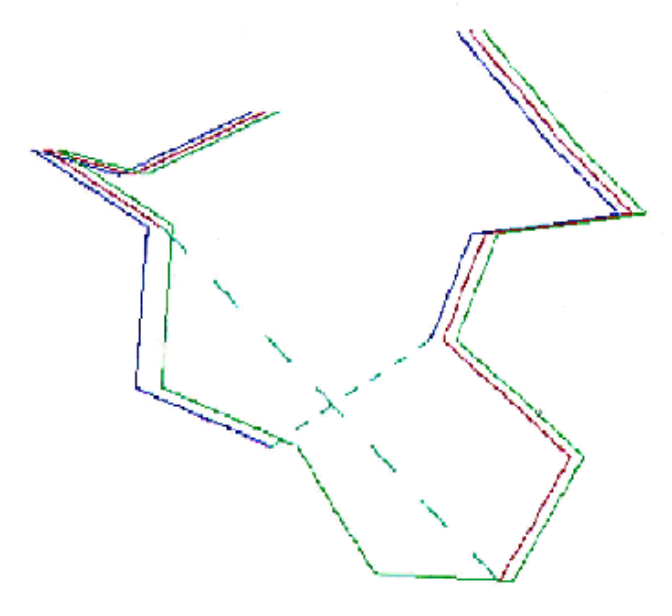
**There are generally two main approaches**

Constructing c-alpha distance matrix

Aligning vectors of secondary structure units

# 2. Alignment correction

- It is possible that the alignment has to be corrected.
- A change of Ala to Glu is possible but unlikely to happen in a hydrophobic core, so this Ala and Glu cannot be aligned.
- Using a multiple sequence alignment program (ClustalW) the residues and properties that have to be conserved can be found.
- By looking at the template structure it will become clear which residues are in the core and are less likely to be changed than the residues at the outside.
- Insertions and deletions can be made in widely divergent parts of the molecule and a multiple sequence alignment can be helpful to find these places.
- Gaps have to be shifted around until they are as small as possible.



*Template structure (green) with the best aligned target (red) with a large gap, and the target after shifting several residues (blue). The gap is much smaller now.*

## How to find an appropriate template Structure for homology modelling…

- **B**asic **L**ocal **A**lignment **S**earch **T**ool

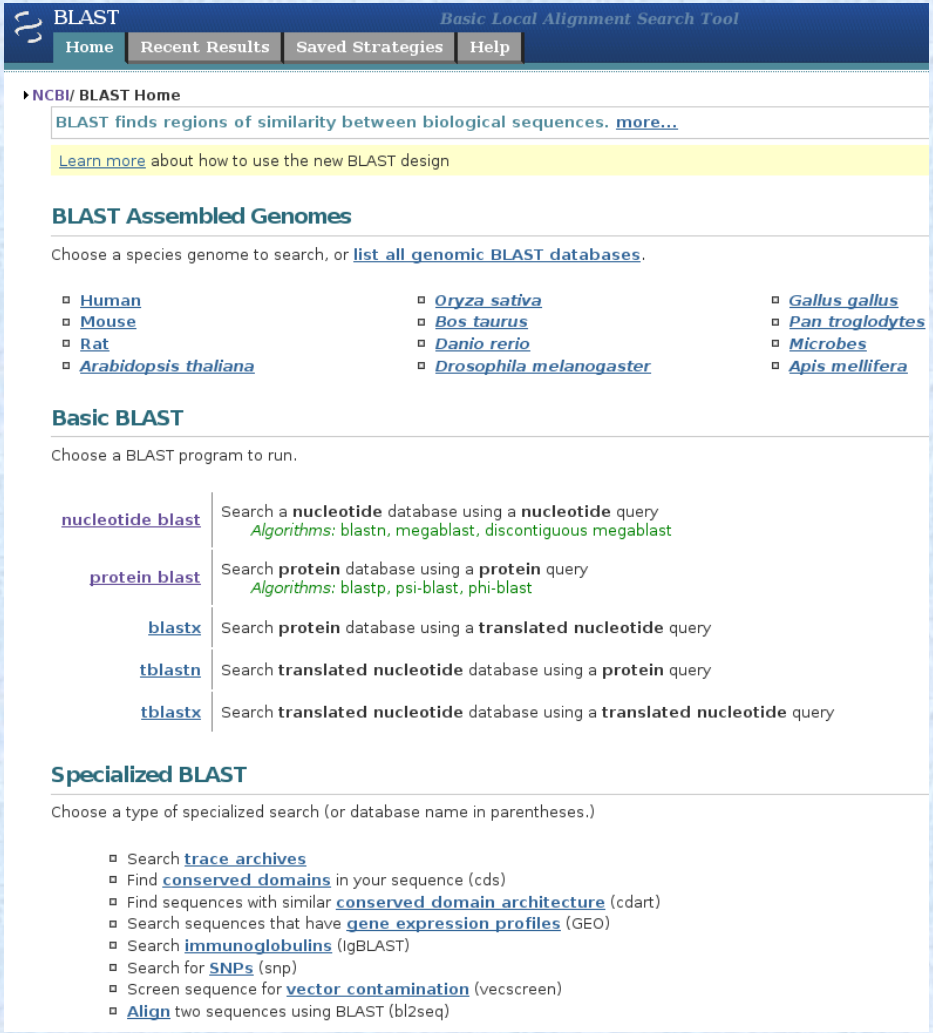
- Used to search protein databases:

- E.g. Non-redundant (nr) & SwissProt to find similar **sequences**.

- Protein Data Bank (PDB) to find **structures** with similar sequences.

- PSI- & PHI-blast are more advanced Blast methods.

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi



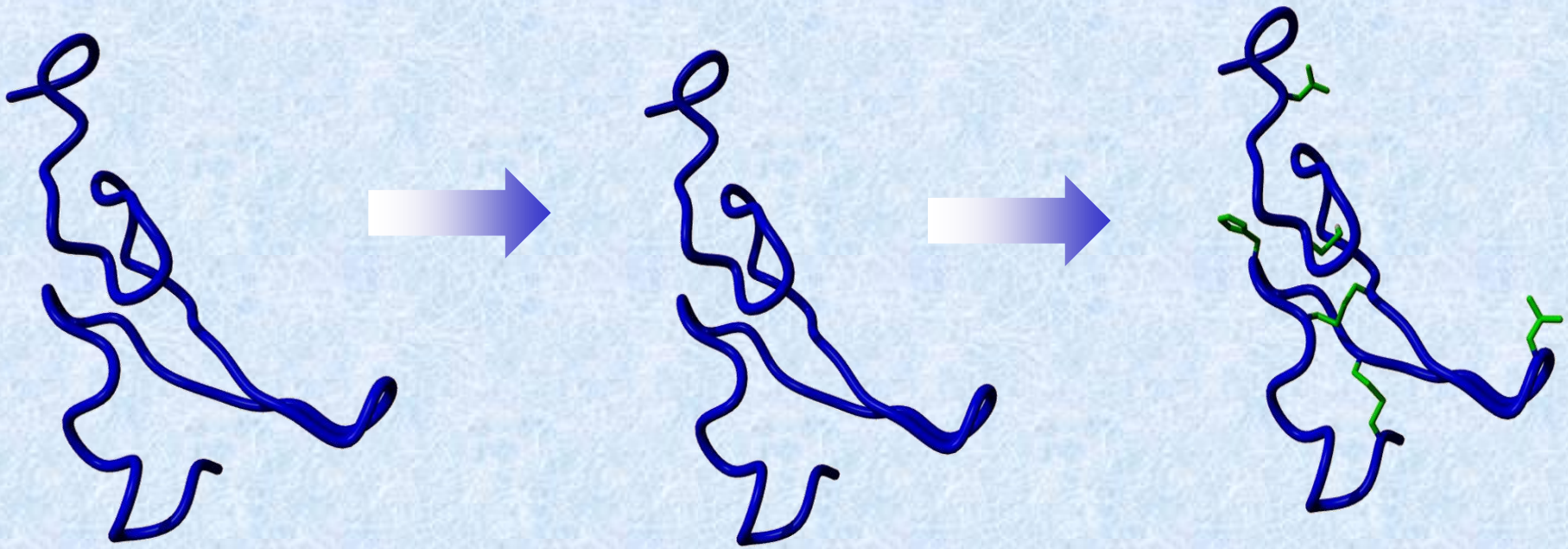
# 3. Backbone generation

When the alignment is correct, the backbone of the target can be created.

The coordinates of the template-backbone are copied to the target.

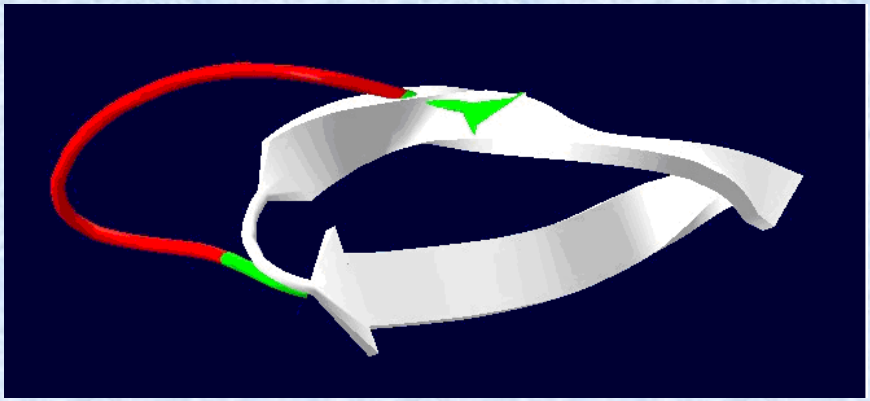
When the residues are identical, the side-chain coordinates are also copied.

Because a PDB-file can always contain some errors, it can be useful to make use of multiple templates.



# 4. Loop modeling

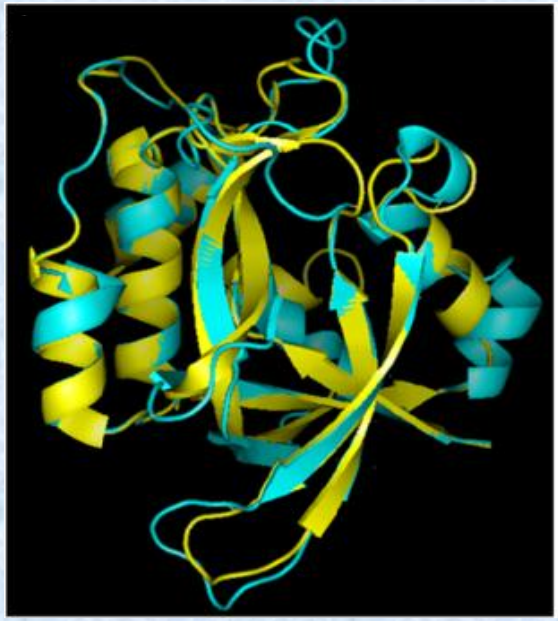
- Check the loops on the basis of steric overlaps
- A specified degree of overlap can be tolerated
- Check the atoms within the loop agains each other
- Then check loop atoms against rest of the protein's atoms

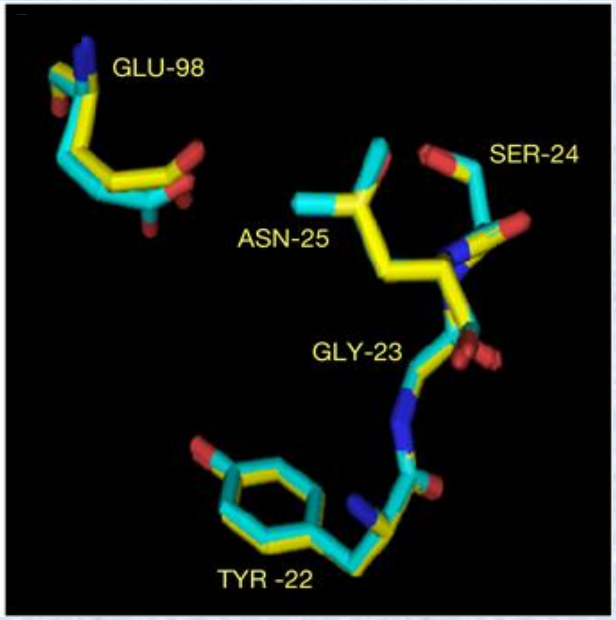


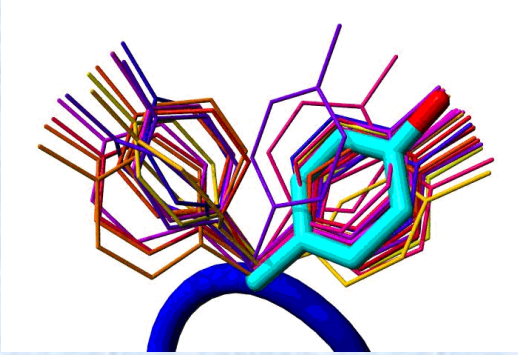
```
TARGET   SLTINNDITLLKLSTAASFSQTVSAVCLPSASDDFAAGTTCVTTGWGLTRYTNANTPDRLQQASLPLLS
         |  |:|||||.|:||.:||||:::.|:::.||:::..   :||||.|:.:|||  |:  :.::  ||  |.   .  |:||
1AZ8     SNTLNNDIMLIKLKSAASLNSRVASISLPTSCA--SAGTQCLISGWGNTKSSGTSYPDVLKCLKAPILS
```

# 5. Side-chain modeling

- With bond lengths, bond angles and two rotable backbone bonds per residue φ and ψ, its very difficult to find the best conformation of a side chain
- In addition, side chains of many residues have one or more degree of freedom.
- Side chain conformational search in loop regions is must
- Side chain residues replaced during coordinate transformations should also be checked





# Rotamer

- Statstical studies show side chain adopt only a small number of many possible conformations
- The correct rotamer of a particular residue is mainly determined by local environment
- Side chain generally adopt conformations where they are closely packed

**Observations:**

- In homologous proteins, corresponding residues virtually retain the same rotameric state
- Within a range of $\chi$ values, 80% of the identical residues and 75% of the mutated residues have the same conformations
- Certain rotamers are almost always associated with certain secondary structure

# 6. Model optimization

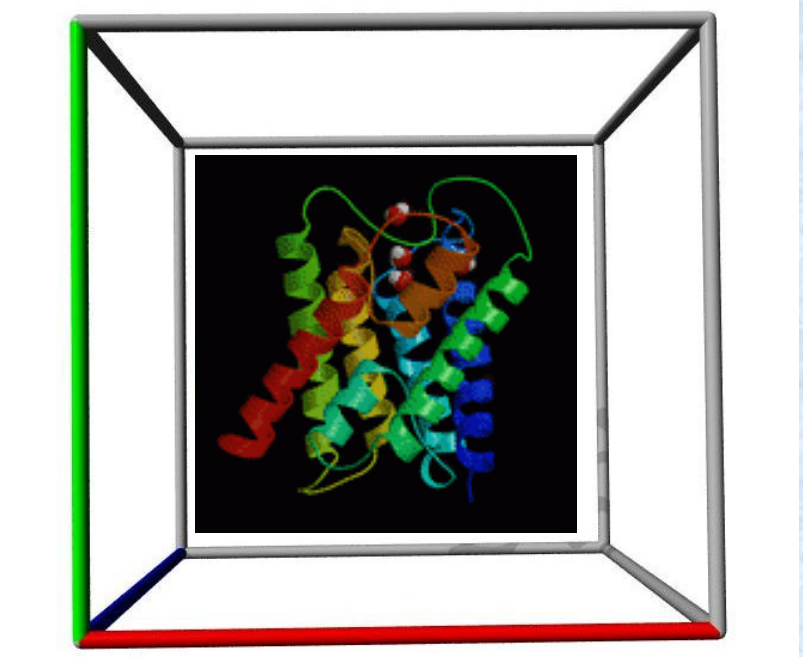
Many structural artifacts can be introduced while the model protein is being built

- ❑ Substitution of large side chains for small ones
- ❑ Strained peptide bonds between segments taken from difference reference proteins
- ❑ Non optimum conformation of loops

Molecular dynamics simulation

Remove big errors

Structure moves to lowest energy conformation



❑Energy Minimisation is used to produce a chemically and conformationally reasonable model protein structure

❑Two mainly used optimisation algorithms are

- ➢ Steepest Descent
- ➢ Conjugate Gradients

❑ Molecular Dynamics is used to explore the conformational space a molecule could visit

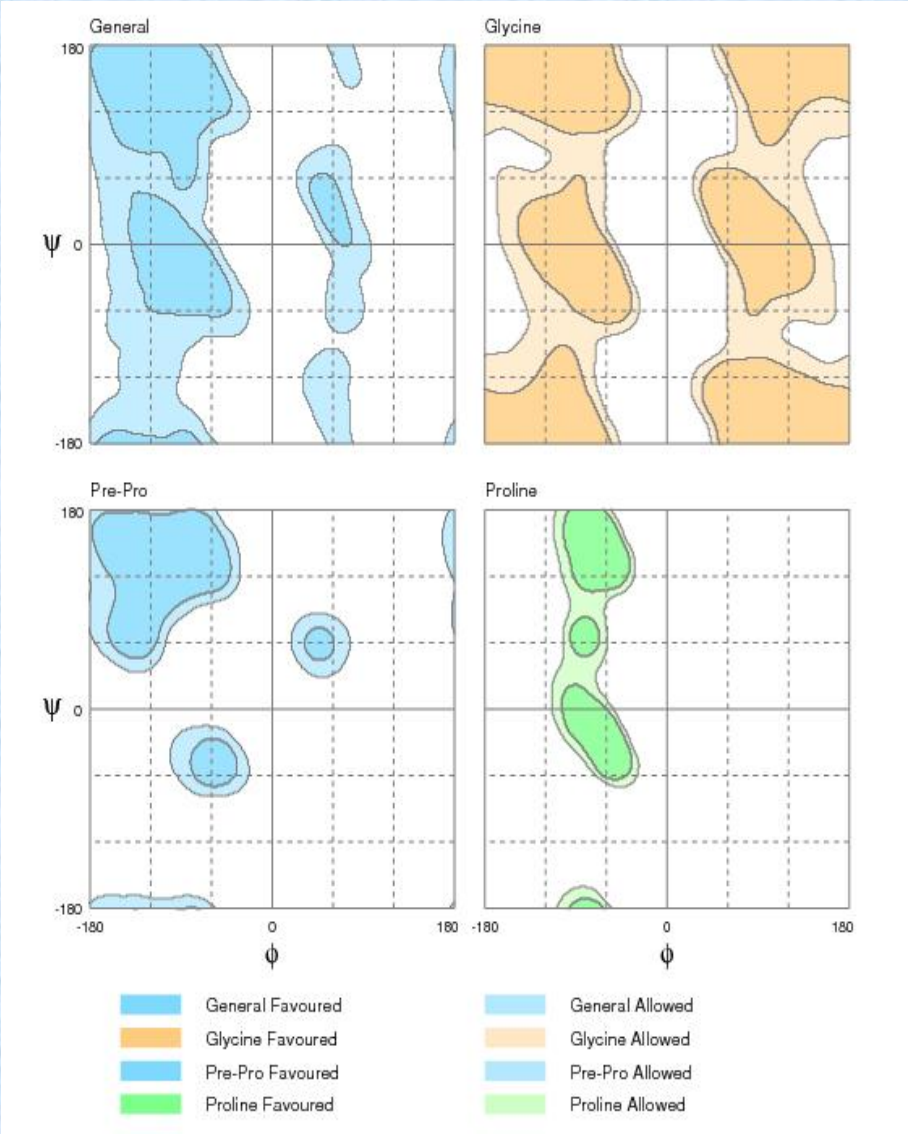
# 7. Model validation

- Every homology model contains errors.Two main reasons
  - % sequence identity between reference and model
  - The number of errors in templates
- Hence it is essential to check the correctness of overall fold/ structure, errors of localized regions and stereochemical parameters: bond lengths, angles, geometries

  - WHAT IF http://www.cmbi.kun.nl/gv/servers/WIWWWI/
  - SOV http://predictioncenter.llnl.gov/local/sov/sov.html
  - PROVE http://www.ucmb.ulb.ac.be/UCMB/PROVE/
  - ANOLEA http://www.fundp.ac.be/pub/ANOLEA.html
  - ERRAT http://www.doe-mbi.ucla.edu/Services/ERRATv2/
  - VERIFY3D http://shannon.mbi.ucla.edu/DOE/Services/Verify_3D/
  - BIOTECH http://biotech.embl-ebi.ac.uk:8400/
  - ProsaII http://www.came.sbg.ac.at
  - WHATCHECK http://www.sander.embl-heidelberg.de/whatcheck/

# Ramachandran Plot

- The results of the Ramachandran plot will be very similar to that of the template.
- A Good template is therefore key!
- Most residues are mainly found on the left-hand side of the plot.
- Glycine is found more randomly within plot (orange), due to its small sidechain (H) preventing clashes with its backbone.
- Proline can only adopt a Phi angle of ~-60° (green) due to its sidechain.
- This also restricts the conformational space of the pre-proline residue.



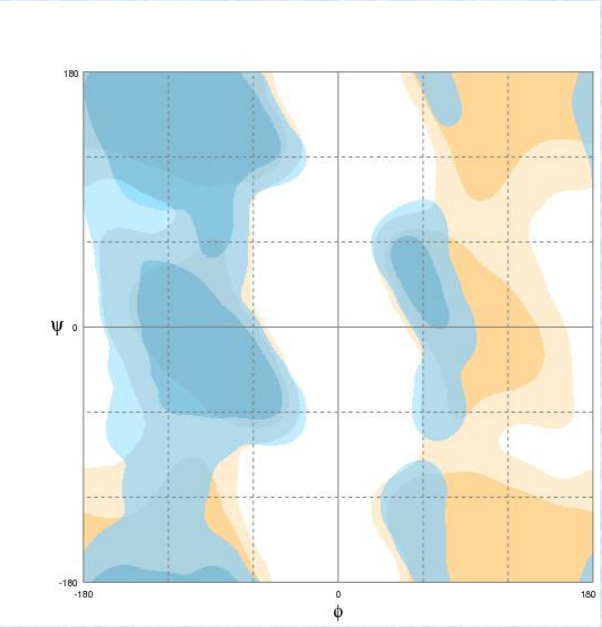
# More Advanced Options

- **Procheck, PROVE, WhatIf:**
  Stereochemical checks on bond lengths, angles and atomic contacts.

- **Ramachandran Plot** is a major component of the evaluation.

- Ensures that the backbone conformation of the model is normal.

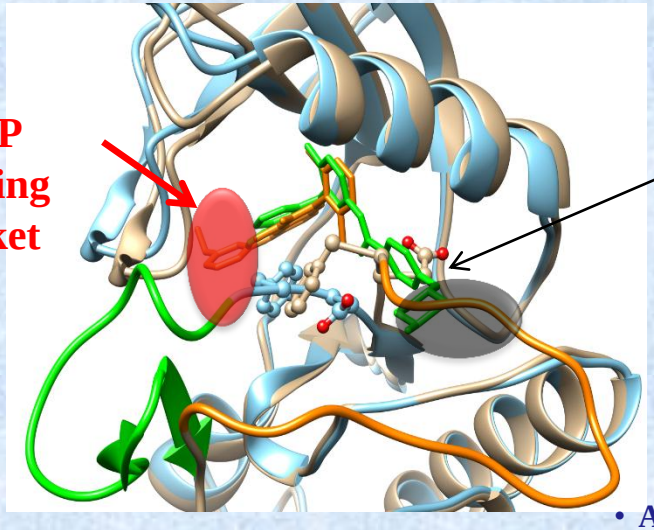
- Modeller is good on the whole, but sometimes struggles with residues found in loops.



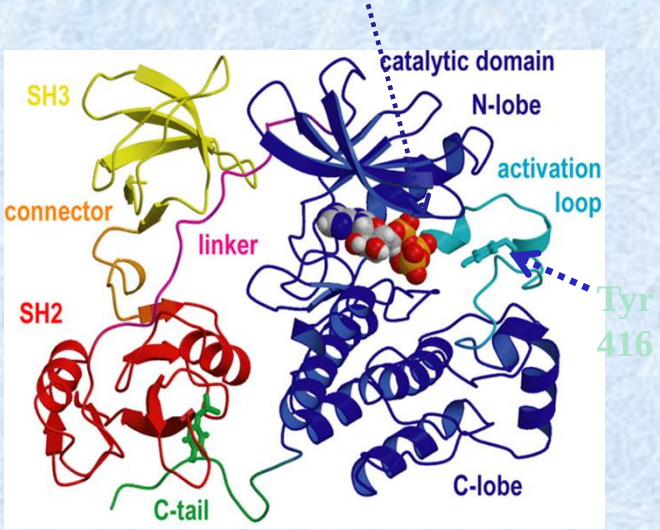

```
 +----------<<<  P  R  O  C  H  E  C  K    S  U  M  M  A  R  Y  >>>----------+
 |                                                                            |
 | mgirk .pdb   2.5                                              104 residues |
 |                                                                            |
*| Ramachandran plot:   91.7% core    7.6% allow    0.3% gener    0.4% disall |
 |                                                                            |
*| All Ramachandrans:   15 labelled residues  Backbone                        |
*| Chi1-chi2 plots:      6 labelled residues  Sidechain                       |
 | Main-chain params:    6 better     0 inside      0 worse                  |
 | Side-chain params:    5 better     0 inside      0 worse                  |
 |                                                                            |
*| Residue properties: Max.deviation:    16.1              Bad contacts:   10 |
*|                     Bond len/angle:    8.0    Morris et al class:  1  1  3 |
 |                                                                            |
 | G-factors           Dihedrals:   0.10  Covalent:   0.29    Overall:   0.16 |
 |                                                                            |
 | M/c bond lengths: 99.1% within limits   0.9% highlighted                   |
*| M/c bond angles:  98.1% within limits   1.9% highlighted                   |
 | Planar groups:   100.0% within limits   0.0% highlighted                   |
 |                                                                            |
 +----------------------------------------------------------------------------+
   + May be worth investigating further.  * Worth investigating further.
```

# Case Study

```
UniProtKB - P07947 (YES_HUMAN)
>sp|P07947|YES_HUMAN Tyrosine-protein kinase Yes OS=Homo sapiens GN=YES1 PE=1 SV=3
MGCIKSKENKSPAIKYRPENTPEPVSTSVSHYGAEPTTVSPCPSSSAKGTAVNFSSLSMT
PFGGSSGVTPFGGASSSFSVVPSSYPAGLTGGVTIFVALYDYEARTTEDLSFKKGERFQI
INNTEGDWWEARSIATGKNGYIPSNYVAPADSIQAEEWYFGKMGRKDAERLLLNPGNQRG
IFLVRESETTKGAYSLSIRDWDEIRGDNVKHYKIRKLDNGGYYITTRAQFDTLQKLVKHY
TEHADGLCHKLTTVCPTVKPQTQGLAKDAWEIPRESLRLEVKLGQGCFGEVWMGTWNGTT
KVAIKTLKPGTMMPEAFLQEAQIMKKLRHDKLVPLYAVVSEEPIYIVTEFMSKGSLLDFL
KEGDGKYLKLPQLVDMAAQIADGMAYIERMNYIHRDLRAANILVGENLVCKIADFGLARL
IEDNEYTARQGAKFPIKWTAPEAALYGRFTIKSDVWSFGILQTELVTKGRVPYPGMVNRE
VLEQVERGYRMPCPQGCPESLHELMNLCWKKDPDERPTFEYIQSFLEDYFTATEPQYQPG
ENL
```




ABL1 kinase in complex
with Imatinib
Type 2 inhibitor
**DFG-OUT**
Template PDB id: 1IEP




- ABL1 kinase in complex with PD166326
- Type 1 inhibitors
- ATP Competitive
- **DFG-IN**
- Template PDB id: 1OPK

# Searching templates for DFG-IN and DFG-OUT conformations

| Description | Max score | Total score | Query cover | E value | Ident | Accession |
|---|---|---|---|---|---|---|
| Chain A, Src Kinase In Complex With A Quinazoline Inhibitor | 830 | 830 | 87% | 0.0 | 82% | 2H8H_A |
| Chain A, Chicken Src Tyrosine Kinase | 815 | 815 | 83% | 0.0 | 84% | 2PTK_A |
| Chain A, Structure Of Unphosphorylated C-Src In Complex With An Inhibitor | 818 | 818 | 83% | 0.0 | 84% | 1Y57_A |
| Chain A, Crystal Structure Of Human Tyrosine-Protein Kinase C-Src | 815 | 815 | 83% | 0.0 | 84% | 1FMK_A |
| Chain A, Structure Of Human C-Src Tyrosine Kinase (Thr338gly Mutant) In Complex With N6-Benzyl Adp | 812 | 812 | 83% | 0.0 | 84% | 1KSW_A |
| Chain A, Crystal Structure Analysis Of Full-Length Carboxyl-Terminal Src Kinase At 2.5 A Resolution | 296 | 296 | 82% | 2e-93 | 38% | 1K9A_A |
| Chain A, Crystal Structure Of Hck In Complex With A Src Family- Selective Tyrosine Kinase Inhibitor | 625 | 625 | 82% | 0.0 | 66% | 1QCF_A |
| Chain A, The Structure Of 1na In Complex With Src T338g | 807 | 807 | 82% | 0.0 | 84% | 4K11_A |
| Chain A, Src Family Kinase Hck-Amp-Pnp Complex | 599 | 599 | 81% | 0.0 | 65% | 1AD5_A |
| Chain A, Structural Basis For The Auto-Inhibition Of C-Abl Tyrosine Kinase | 376 | 376 | 80% | 1e-123 | 44% | 1OPK_A |
| Chain A, Organization Of The Sh3-Sh2 Unit In Active And Inactive Forms Of The C-Abl Tyrosine Kinase | 375 | 375 | 80% | 3e-123 | 44% | 2FO0_A |
| Chain A, Structural Basis For The Auto-Inhibition Of C-Abl Tyrosine Kinase | 376 | 376 | 80% | 5e-123 | 44% | 1OPL_A |
| Chain A, Crystal Structure Of An Auto-inhibited Form Of Bruton's Tyrosine Kinase | 319 | 319 | 79% | 3e-102 | 38% | 4XI2_A |
| Chain B, Crystal Structure Of An Sh2-kinase Domain Construct Of C-abl Tyrosine Kinase | 337 | 337 | 70% | 1e-109 | 45% | 4XEY_B |
| Chain A, Crystal Structure Of Human Feline Sarcoma Viral Oncogene Homologue (V- Fes) | 233 | 233 | 69% | 4e-70 | 39% | 3BKB_A |
| Chain A, Crystal Structure Of Chicken C-Src Kinase Domain In Complex With The Cancer Drug Imatinib | 543 | 543 | 52% | 0.0 | 90% | 2OIQ_A |
| Chain A, Crystal Structure Of The L317i Mutant Of The Chicken C-Src Tyrosine Kinase Domain Complexed With Imatinib | 542 | 542 | 52% | 0.0 | 90% | 3OEZ_A |
| Chain A, Human Src Kinase Bound To Kinase Inhibitor Bosutinib | 542 | 542 | 52% | 0.0 | 90% | 4MXO_A |

- **1OPK is selected as a template to model cYES kinase structure with DFG-IN conformation. It is due to 80% query coverage and well defined active site for binding of type 1 kinase inhibitors.**
- **Moreover, crystal structure of corresponding DFG-OUT conformation (1IEP) is also available and it is used as a template to model cYES kinase structure with DFG-OUT conformation.**

# Target – Template Sequence Alignment (BLASTp)

1OPK:A

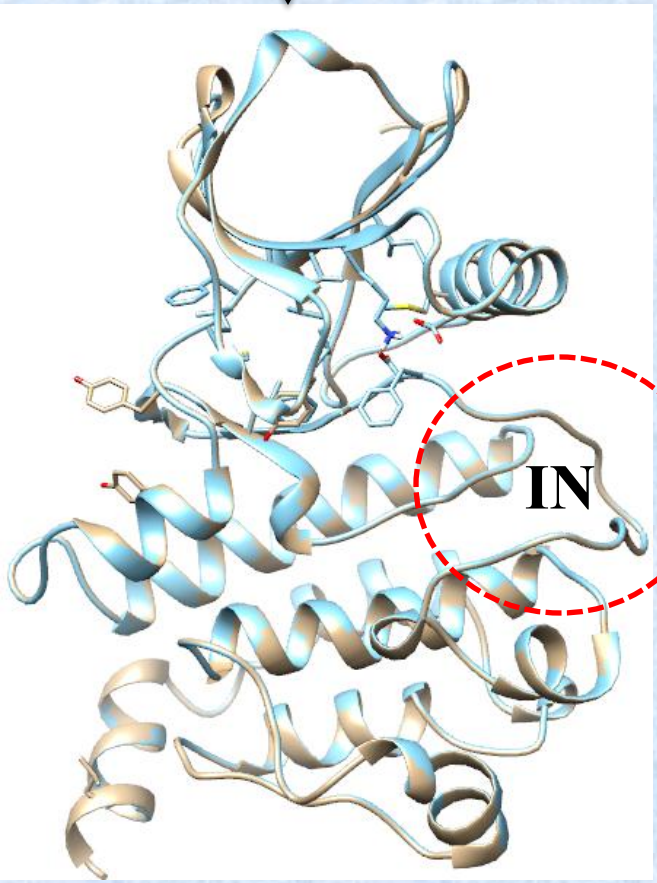
Sequence ID: lcl|Query_34595 Length: 495 Number of Matches: 2

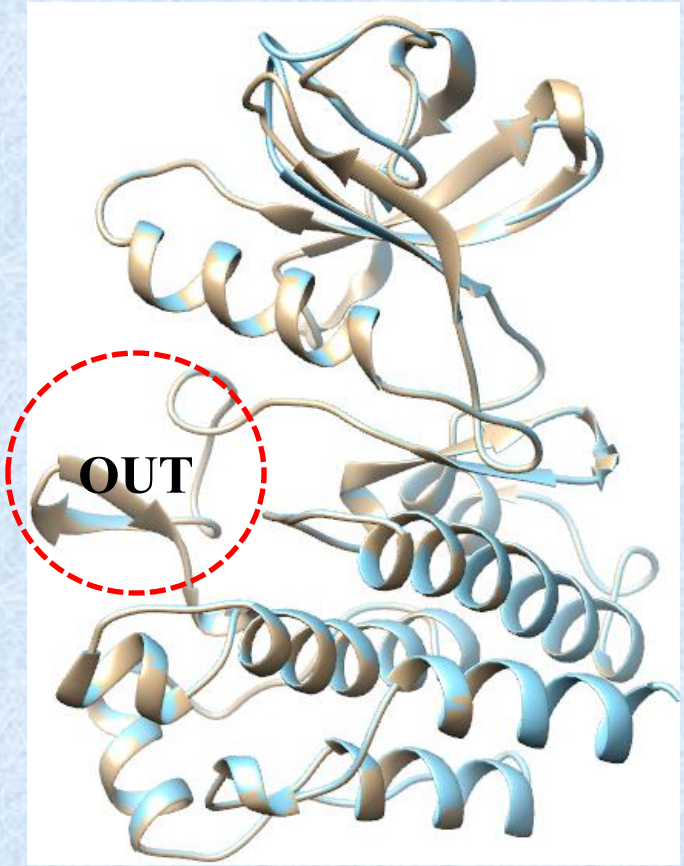
Range 1: 44 to 477 Graphics ▼ Next Match ▲ Previous

| Score | Expect | Method | Identities | Positives | Gaps |
|---|---|---|---|---|---|
| 376 bits(965) | 4e-128 | Compositional matrix adjust. | 197/443(44%) | 281/443(63%) | 15/443(3%) |

```
Query  94   TIFVALYDYEARTTEDLSFKKGERFQIIN-NTEGDWWEARSIATGKNGYIPSNYVAPADS  152
            +FVALYD+ A     LS  KGE+ +++  N  G+W EA++      G++PSNY+ P +S
Sbjct  44   NLFVALYDFVASGDNTLSITKGEKLRVLGYNHNGEWCEAQT--KNGQGWVPSNYITPVNS  101

Query  153  IQAEEWYFGKMGRKDAERLLLNPGNQRGIFLVRESETTKGAYSLSIRDWDEIRGDNVKHY  212
            ++   WY G + R  AE LL +  N  G FLVRESE++ G  S+S+R         V HY
Sbjct  102  LEKHSWYHGPVSRNAAEYLLSSGIN--GSFLVRESESSPGQRSISLR-----YEGRVYHY  154

Query  213  KIRKLDNGGYYITTRAQFDTLQKLVKHYTEHADGLCHKLTTVCPTV-KPQTQGLAK--DA  269
            +I    +G  Y+++ ++F+TL +LV H++  ADGL   L    P   KP   G++   D
Sbjct  155  RINTASDGKLYVSSESRFNTLAELVHHHSTVADGLITTLHYPAPKRNKPTIYGVSPNYDK  214

Query  270  WEIPRESLRLEVKLGQGCFGEVWMGTWNG-TTKVAIKTLKPGTMMPEAFLQEAQIMKKLR  328
            WE+ R  + ++ KLG G +GEV+ G W   +  VA+KTLK  TM  E FL+EA +MK+++
Sbjct  215  WEMERTDITMKHKLGGGQYGEVYEGVWKKYSLTVAVKTLKEDTMEVEEFLKEAAVMKEIK  274

Query  329  HDKLVPLYAVVSEEP-IYIVTEFMSKGSLLDFLKEGDGKYLKLPQLVDMAAQIADGMAYI  387
            H  LV L  V + EP  YI+TEFM+ G+LLD+L+E + + +    L+ MA QI+  M Y+
Sbjct  275  HPNLVQLLGVCTREPPFYIITEFMTYGNLLDYLRECNRQEVSAVVLLYMATQISSAMEYL  334

Query  388  ERMNYIHRDLRAANILVGENLVCKIADFGLARLIEDNEYTARQGAKFPIKWTAPEAALYG  447
            E+ N+IHR+L A N LVGEN + K+ADFGL+RL+  + YTA  GAKFPIKWTAPE+  Y
Sbjct  335  EKKNFIHRNLAARNCLVGENHLVKVADFGLSRLMTGDTYTAHAGAKFPIKWTAPESLAYN  394

Query  448  RFTIKSDVWSFGILQTELVTKGRVPYPGMVNREVLEQVERGYRMPCPQGCPESLHELMNL  507
            +F+IKSDVW+FG+L  E+ T G  PYPG+   +V E +E+ YRM  P+GCPE ++ELM
Sbjct  395  KFSIKSDVWAFGVLLWEIATYGMSPYPGIDLSQVYELLEKDYRMERPEGCPEKVYELMRA  454

Query  508  CWKKDPDERPTFEYIQSFLEDYF  530
            CW+ +P +RP+F  I    E  F
Sbjct  455  CWQWNPSDRPSFAEIHQAFETMF  477
```

| Kinases | RMSD | |
|---|---|---|
| | Back Bone | All |
| 1OPK vs cYes (DFG-IN) | 0.18 | 0.47 |
| 1IEP vs cYes (DFG-OUT) | 0.39 | 0.67 |

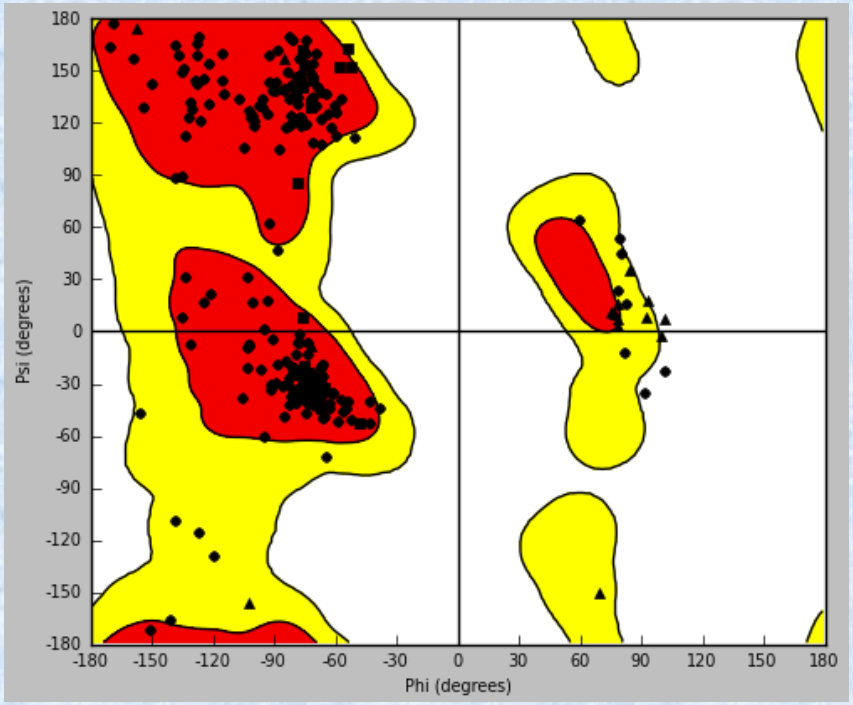


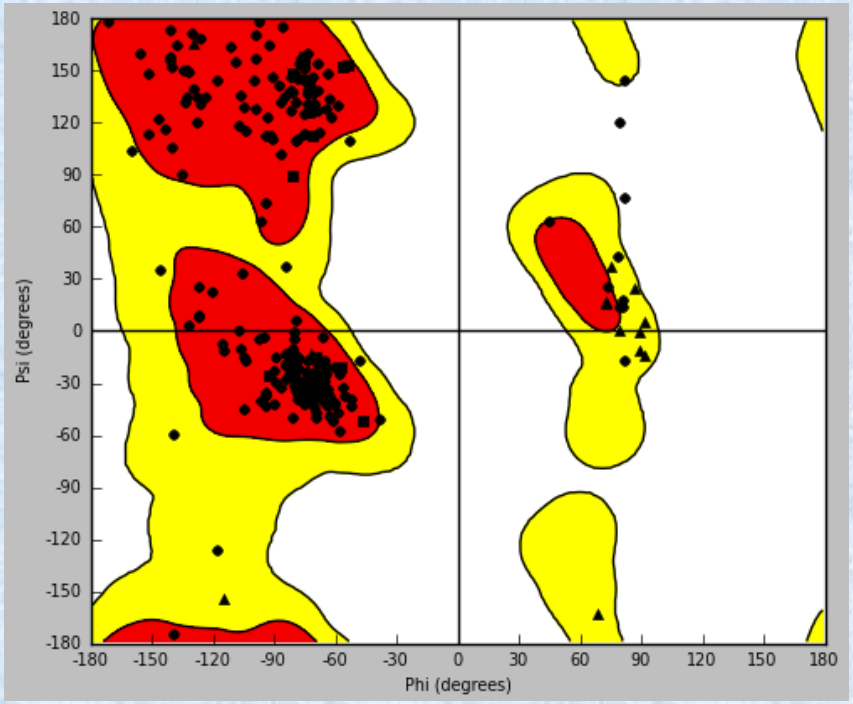



# Model validation

Template protein : c-Abl tyrosine kinase
Target protein : c-Yes tyrosine kinase

| Identities | Positives | Gaps |
|---|---|---|
| 197/443 (44%) | 281/443 (63%) | 15/443 (3%) |




| | | DFG-OUT | DFG-IN |
|---|---|---|---|
| Template PDB id | : | 1IEP | 1OPK |
| Resolution | : | 2.1 Å | 1.80 Å |
| # of residues in favoured region | : | 249 ( 94.3%) | 247 ( 93.9%) |
| # of residues in allowed region | : | 11 ( 04.2%) | 12 ( 04.6%) |
| # of residues in outlier region | : | 4 ( 01.5%) | 4 ( 01.5%) |

# Web-servers for Homology Modeling

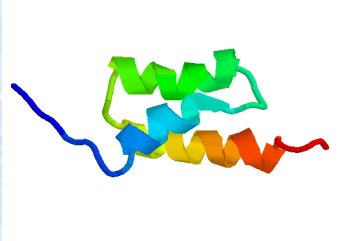
- SWISS Model : http://www.expasy.org/swissmod/SWISS-MODEL.html
- WHAT IF : http://www.cmbi.kun.nl/swift/servers/
- The CPHModels Server : http://www.cbs.dtu.dk/services/CPHmodels/
- 3D Jigsaw : http://www.bmm.icnet.uk/~3djigsaw/
- SDSC1 : http://cl.sdsc.edu/hm.html
- EsyPred3D : http://www.fundp.ac.be/urbm/bioinfo/esypred/

# Tools

- COMPOSER http://www.tripos.com/sciTech/inSilicoDisc/bioInformatics/matchmaker.html
- MODELER http://salilab.org/modeler
- InsightII http://www.msi.com/
- SYBYL http://www.tripos.com/

# Ab-initio Prediction

## Prediction from sequence using first principles

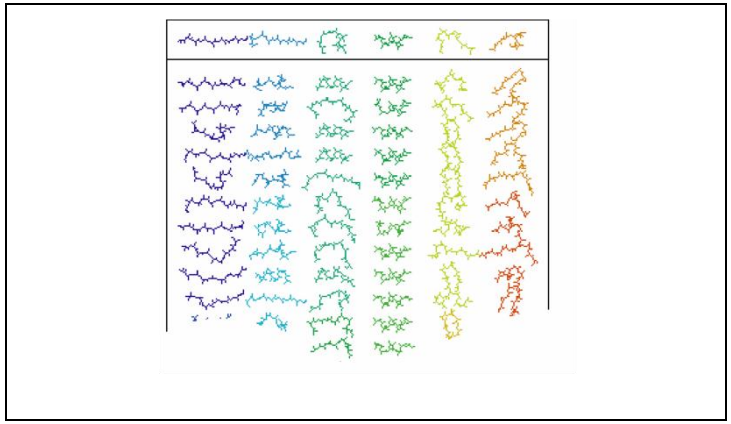
AVVTW...GTTWVR ⟶ 

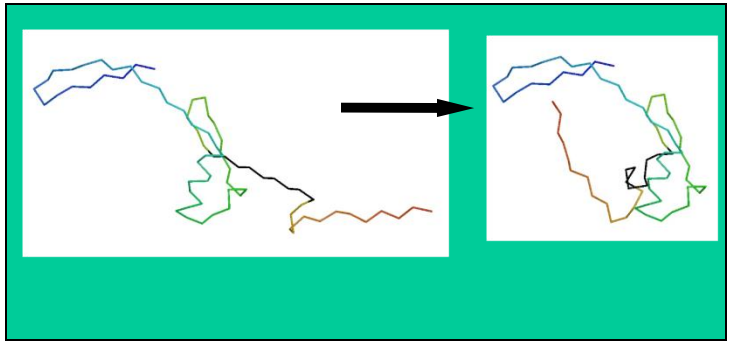
- ***Ab initio*** **protein structure prediction methods build protein 3D structures from sequence based on physical principles.**
- **Importance**
  - **The *ab initio* methods are important even though they are computationally demanding**
  - ***Ab initio* methods predict protein structure based on physical models, they are indispensable complementary methods to Knowledge-based approach**

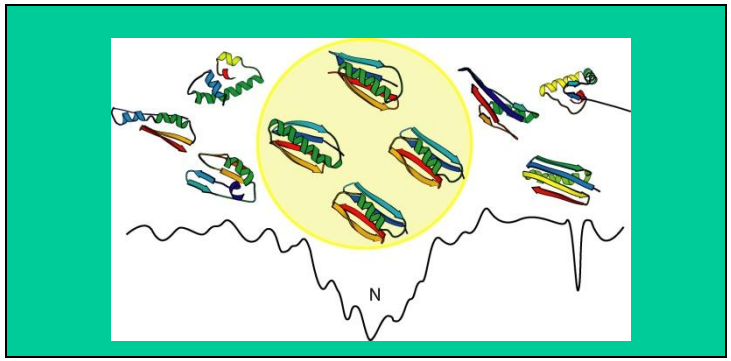
**Knowledge-based approach would fail in following conditions:**

- **Structure homologues are not available**
- **Possible undiscovered new fold exists**

# Structure Prediction with Rosetta








- **Select fragments consistent with local sequence preferences**
- **Assemble fragments into models with native-like global properties**
- **Identify the best model from the population of decoys**

$$P(\text{structure}|\text{sequence}) \propto P(\text{sequence}|\text{structure}) \times P(\text{structure})$$

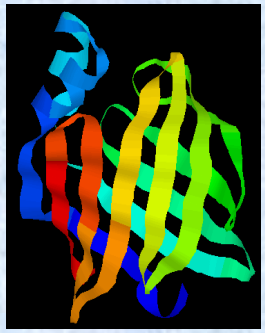
Sequence dependent:
- hydrophobic burial
- residue pair interaction

Sequence independent:
- helix-strand packing
- strand-strand packing
- sheet configurations
- vdW interactions

# Threading - Fold Recognition

*Identify "best" fit between target sequence & template structure*

1. Develop energy function
2. Develop template library
3. Align target sequence with each template & score
4. Identify best scoring template (1D to 3D alignment)
5. Refine structure as in homology modeling



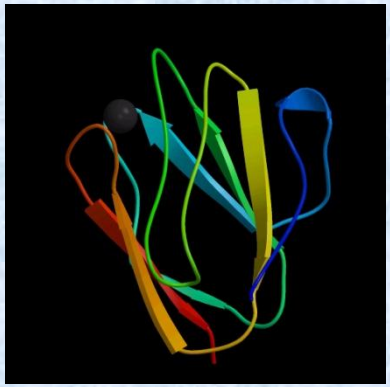
- **Threading** - works "sometimes"
  - Computationally? Can be expensive or cheap, depends on energy function & whether "all atom" or "backbone only" threading
  - Accuracy? in theory, should not depend on sequence identity (should depend on quality of template library & "luck")
  - But, usually higher sequence identity better model

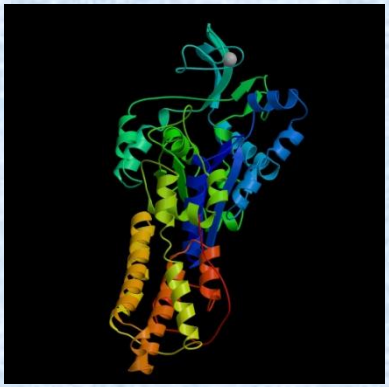
# Threading

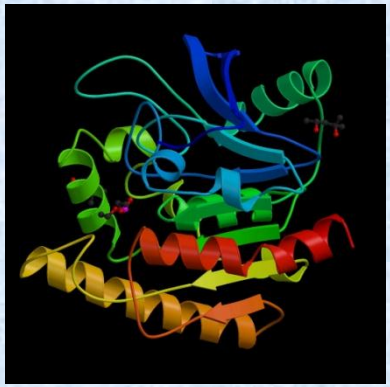
Target Sequence

ALKKGF…HFDTSE

Structure Templates







1. **Align target sequence with template structures (fold library) from the Protein Data Bank (PDB)**
2. **Calculate energy score to evaluate goodness of fit between target sequence & template structure**
3. **Rank models based on energy scores**

# Threading Goals & Issues

**Find "correct" sequence-structure alignment of a target sequence with its native-like fold in PDB**

- ***Structure database*** **- must be complete: no decent model if no good template in library!**
- **Sequence-structure alignment algorithm:**

  **Bad alignment  Bad score!**

- **Energy function (scoring scheme):**
  - **must distinguish correct sequence-fold alignment from incorrect sequence-fold alignments**
  - **must distinguish "correct" fold from close decoys**
- **Prediction reliability assessment - how determine whether predicted structure is correct (or even close?)**

**Two main methods (and combinations of these)**

**Structural profile (environmental) physicochemical properties of aa's**

**Contact potential (statistical)**

**based on contact statistics from PDB**

# Protein Threading: typical energy function

MTYKLILNGKTKGETTTEAVDAATAEKVFQYANDNGVDGEWTYTE

What is "**p**robability" that two specific residues are in contact?

How well does a specific residue fit **s**tructural environment?

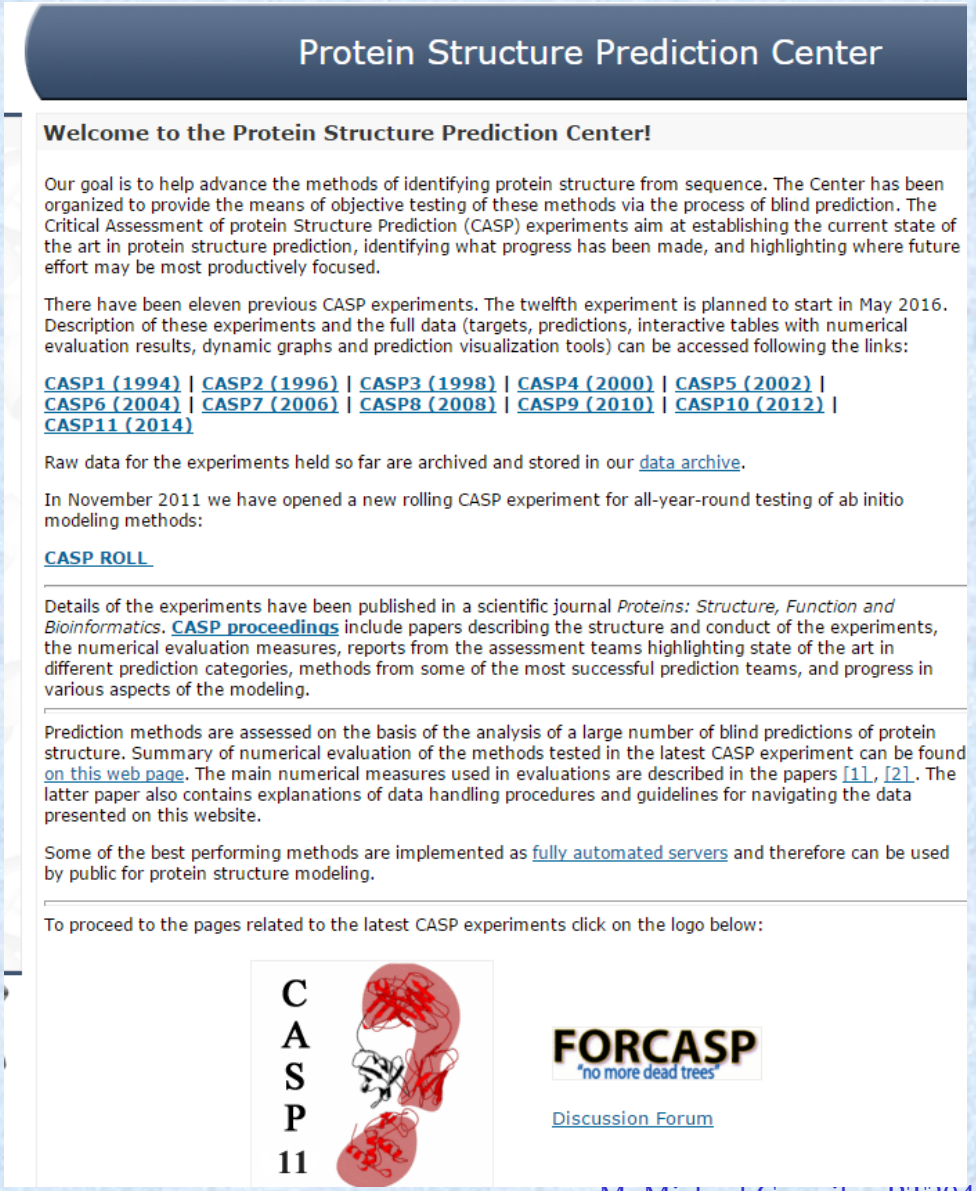
Alignment **g**ap penalty?

**Total energy: $E_p + E_s + E_g$**

Find a sequence-structure alignment that minimizes the energy function

# Critical Assessment of Structure Prediction (CASP)

- A Biennial *competition* that has run since 1994.
- The next competition will be in 2016 (CASP12)
- http://predictioncenter.org/
- The goal is to advance the methods for predicting protein structure from sequence.
- Protein structures yet to be published are used as blind targets for the prediction methods, with only sequence information released.
- Competitors may use Homology Modelling, Fold recognition or Ab Initio structural prediction methods to propose the structure of the protein.



Protein Structure Prediction Center

**Welcome to the Protein Structure Prediction Center!**

Our goal is to help advance the methods of identifying protein structure from sequence. The Center has been organized to provide the means of objective testing of these methods via the process of blind prediction. The Critical Assessment of protein Structure Prediction (CASP) experiments aim at establishing the current state of the art in protein structure prediction, identifying what progress has been made, and highlighting where future effort may be most productively focused.

There have been eleven previous CASP experiments. The twelfth experiment is planned to start in May 2016. Description of these experiments and the full data (targets, predictions, interactive tables with numerical evaluation results, dynamic graphs and prediction visualization tools) can be accessed following the links:

CASP1 (1994) | CASP2 (1996) | CASP3 (1998) | CASP4 (2000) | CASP5 (2002) | CASP6 (2004) | CASP7 (2006) | CASP8 (2008) | CASP9 (2010) | CASP10 (2012) | CASP11 (2014)

Raw data for the experiments held so far are archived and stored in our data archive.

In November 2011 we have opened a new rolling CASP experiment for all-year-round testing of ab initio modeling methods:

CASP ROLL

Details of the experiments have been published in a scientific journal *Proteins: Structure, Function and Bioinformatics*. CASP proceedings include papers describing the structure and conduct of the experiments, the numerical evaluation measures, reports from the assessment teams highlighting state of the art in different prediction categories, methods from some of the most successful prediction teams, and progress in various aspects of the modeling.

Prediction methods are assessed on the basis of the analysis of a large number of blind predictions of protein structure. Summary of numerical evaluation of the methods tested in the latest CASP experiment can be found on this web page. The main numerical measures used in evaluations are described in the papers [1], [2]. The latter paper also contains explanations of data handling procedures and guidelines for navigating the data presented on this website.

Some of the best performing methods are implemented as fully automated servers and therefore can be used by public for protein structure modeling.

To proceed to the pages related to the latest CASP experiments click on the logo below:

CASP 11

FORCASP "no more dead trees"

Discussion Forum

# Course contents: Review

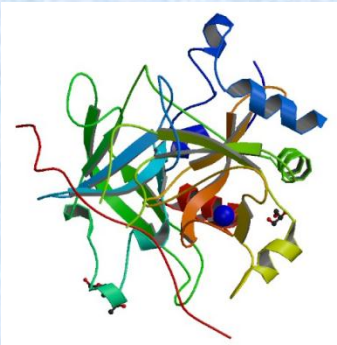
| | |
|---|---|
| **Bioinformatics: introduction** | **DONE** |
| **SSH, Unix commands, running programs in unix** | **DONE** |
| **Bioinformatics Databases** | **DONE** |
| **Sequence alignment: algorithms and online resources** | **DONE** |
| **Protein sequence analysis** | **DONE** |
| **Protein secondary and tertiary structure** | **DONE** |
| **Protein structure prediction** | **DONE** |
| **Phylogeny: methods and applications** | **DONE** |
| **Online resources in Bioinformatics/programming** | **DONE** |
| **Machine learning techniques/Genome analysis** | **Ongoing** |

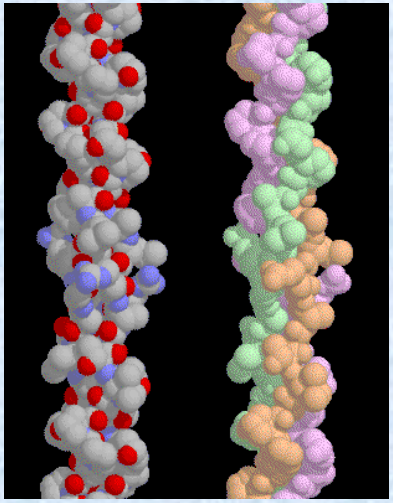
# Proteins: secondary databases

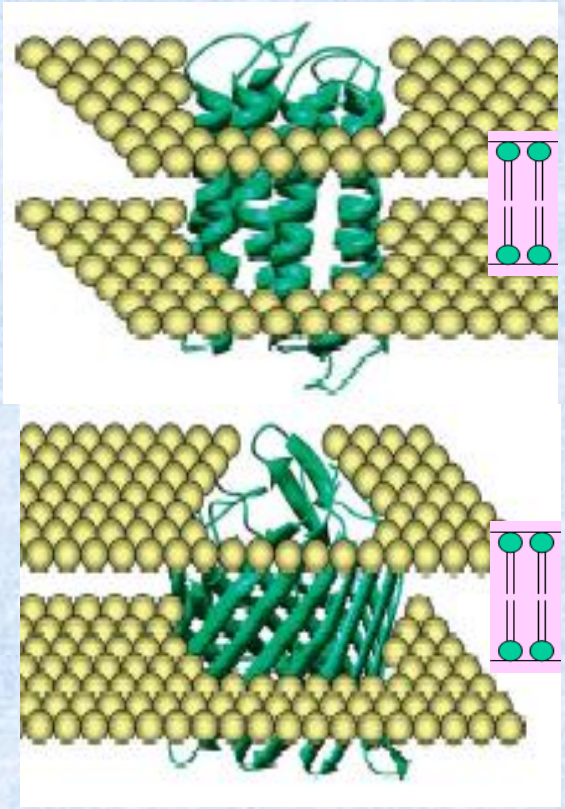
Globular proteins

Fibrous proteins

Membrane proteins







Cytoplasm

Inner membrane

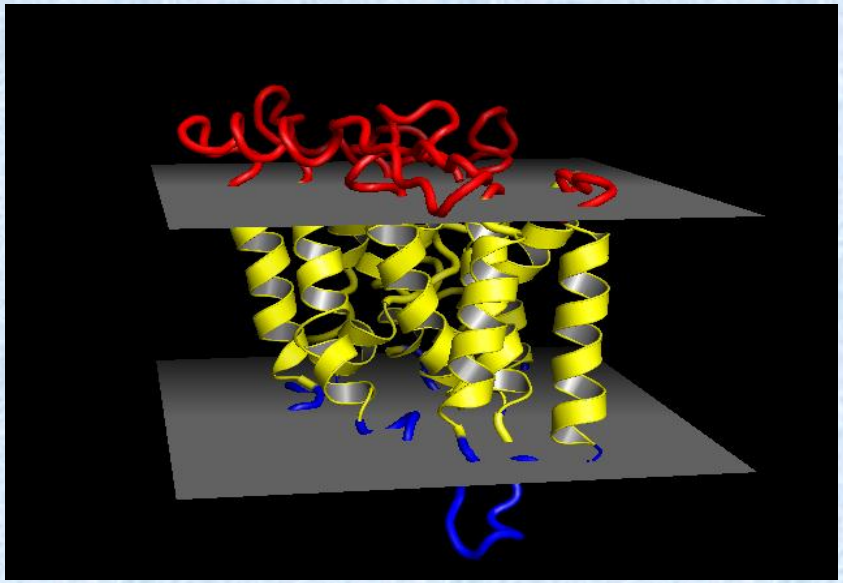
α-helical

Periplasm

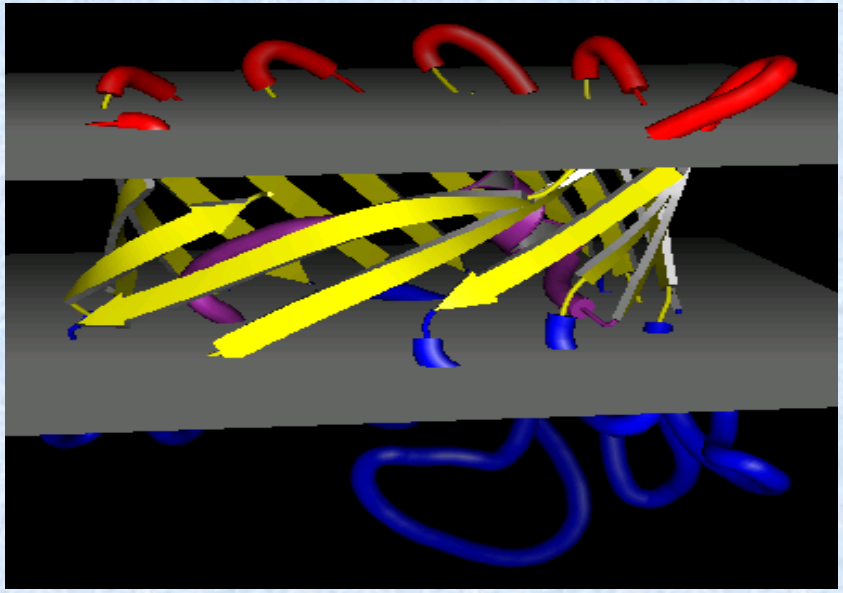
Outer membrane

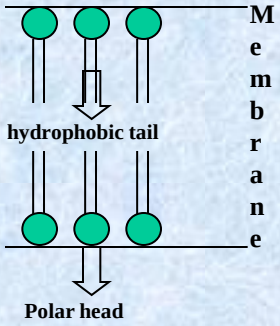
β-barrel (TMβ)

Outer space



(a)






(b)

Figure 1.9

# PDBTM

**Database for membrane proteins**

**Algorithm to (TMDET) to identify the membrane spanning regions from 3D structure**

- **Protein is cut into 1Å slices and compute the membrane exposed area of hydrophobic (F, G, I, L, M, V, W and Y) and hydrophilic residues**
- **Hydrophobic factor is the hydrophobic area divided by total surface area.**
- **Structure factor is defined as the product of three factors, the straightness, turn and end-chain factor. These factors depend on the projection of Ca atoms of the residues i-3, i and i+3 onto a normal vector of membrane planes.**
- **The objective function (Q-value) is the average of the products of the hydrophobic factor and the structure factor in each slice.**

# PDBTM: Protein Data Bank of Transmembrane Proteins

PDBTM version: 2017-02-10 Number of transmembrane proteins: 3099 (alpha: 2723 , beta: 364 )

all 1a0s

- Home
- Search
- Download
- Statistics
- Documents
- Help

## Welcome to the PDBTM home page

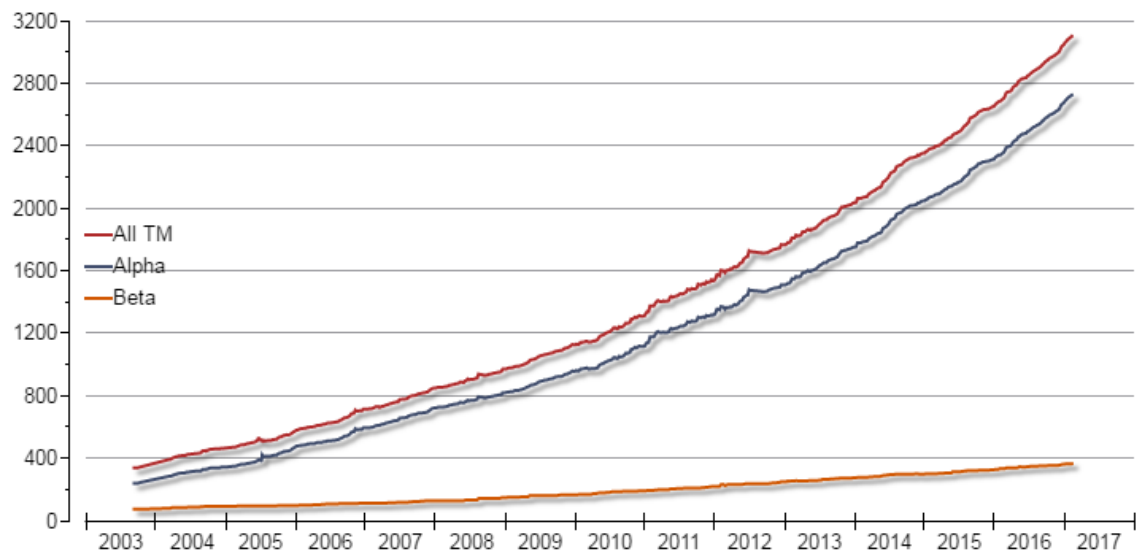
PDBTM is the first comprehensive and up-to-date transmembrane protein selection of the Protein Data Bank (PDB). PDBTM database is maintained at the Institute of Enzymology by the Membrane Protein Bioinformatics Research Group. The PDBTM database was created by scanning all PDB entries with the TMDET algorithm. You can get more information about PDBTM in our articles and in the PDBTM manual. If you find PDBTM useful in your research, please cite our articles (Bioinformatics 20, 2964-2972; Nucleic Acids Research 33 Database Issue, D275-D278; Nucleic Acids Research 41 Database Issue, D524-D529 ).

**5jjj**

5jjj

PDBTM type: Tm_Alpha
Chain(s): A[7], B[2], C[7], D[2]

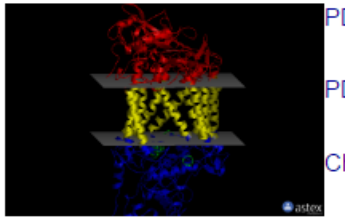
**PDBTM growth statistics**



search pdb_id: 1prc Submit

number of match: 1 Download XML files as zip Get it!

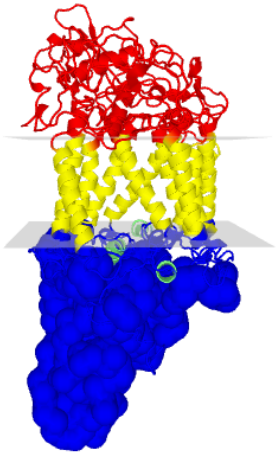
**1. 1prc**



PDB classification: PHOTOSYNTHETIC REACTION CENTER 04-FEB-88 1PRC 1PRC 3

PDBTM type: Tm_Alpha

Chain(s): C[0], L[5], M[5], H[1]



Sequence

Chain: L

C
H
L
M

ALLSFE GGTLIGGDLFDFWVGPYFVGF
GVSAIF SLIGYAASQGPTWDPFAISINP
PDLKYG ytosolic QAITVCALGAFISWML
REVEIS FCVPIFMFCVLQVFRP
LLLGSWGHAFPYGIL SHLDWVNNFGYQYLNWHYN
PGHMSSVSFLFVNAMALGLHGGLILSVANPGDGD
KVKTAEHENQYFRDVVGYSIGALSIHRLGLFLAS
NIFLTGAFGTIASGPFWTRGWPEWWGWWLDIPFW
S

Downloads

Cross references

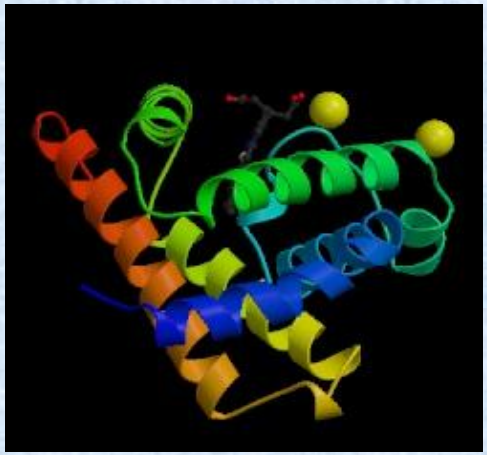
```
    <TMTYPE>Tm_Alpha</TMTYPE>
    <SPRES>Tm_Alpha</SPRES>
    <PDBKWRES>yes</PDBKWRES>
    <PDBKWORD>Membrane</PDBKWORD>
  </RAWRES>
▼<MEMBRANE>
    <NORMAL X="0.00000000" Y="0.00000000" Z="17.75000000"/>
  ▼<TMATRIX>
      <ROWX X="0.83854312" Y="-0.09631874" Z="-0.53625381" T="-52.36920929"/>
      <ROWY X="0.00000000" Y="0.98424947" Z="-0.17678508" T="-14.93281200"/>
      <ROWZ X="0.54483527" Y="0.14824191" Z="0.82533562" T="-33.65978378"/>
    </TMATRIX>
  </MEMBRANE>
▼<CHAIN CHAINID="A" NUM_TM="7" TYPE="alpha">
  ▼<SEQ>
      MESPIQIFRG EPGPTCAPSA CLPPNSSAWF PGWAEPDSNG SAGSEDAQLE PAHISPAIPV IITAVYSVVF
      MNSWPFGDVL CKIVISIDYY NMFTSIFTLT MMSVDRYIAV CHPVKALDFR TPLKAKIINI CIWLLSSSVG
      IIIVCYTLMI LRLKSVRLLS GSREKDRNLR RITRLVLVVV AVFVVCWTPI HIFILVEALG STSHSTAALS
      RNTVQDPAYL RDIDGMNKPV
    </SEQ>
    <REGION seq_beg="1" pdb_beg="1" seq_end="59" pdb_end="59" type="1"/>
    <REGION seq_beg="60" pdb_beg="60" seq_end="87" pdb_end="87" type="H"/>
    <REGION seq_beg="88" pdb_beg="88" seq_end="90" pdb_end="90" type="2"/>
    <REGION seq_beg="91" pdb_beg="91" seq_end="118" pdb_end="118" type="H"/>
    <REGION seq_beg="119" pdb_beg="119" seq_end="132" pdb_end="132" type="1"/>
    <REGION seq_beg="133" pdb_beg="133" seq_end="158" pdb_end="158" type="H"/>
    <REGION seq_beg="159" pdb_beg="159" seq_end="172" pdb_end="172" type="2"/>
    <REGION seq_beg="173" pdb_beg="173" seq_end="195" pdb_end="195" type="H"/>
    <REGION seq_beg="196" pdb_beg="196" seq_end="225" pdb_end="225" type="1"/>
    <REGION seq_beg="226" pdb_beg="226" seq_end="250" pdb_end="250" type="H"/>
    <REGION seq_beg="251" pdb_beg="251" seq_end="272" pdb_end="272" type="2"/>
    <REGION seq_beg="273" pdb_beg="273" seq_end="294" pdb_end="294" type="H"/>
    <REGION seq_beg="295" pdb_beg="295" seq_end="307" pdb_end="307" type="1"/>
    <REGION seq_beg="308" pdb_beg="308" seq_end="334" pdb_end="334" type="H"/>
    <REGION seq_beg="335" pdb_beg="335" seq_end="380" pdb_end="380" type="2"/>
  </CHAIN>
</pdbtm>
```

# Different structural classes

4MBN

all-α

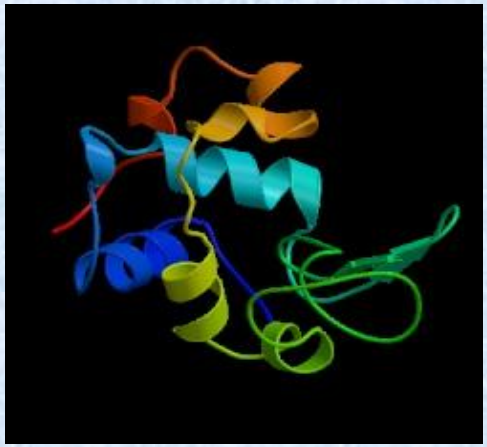
Dominated by α-helices
α>40; β<5



4LYZ

α β

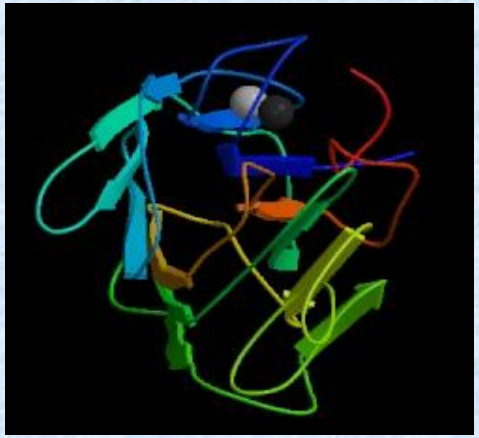
Helices and strands tend to Segregate
α>15; β>10



3CNA

all-β

Dominated by β-strands
β>40; α<20



1TIM

α β

Helices and strands mix each other
α>15; β>10

# SCOP
# Structural Classification of Proteins

**Provides a detailed and comprehensive description of the structural and evolutionary relationships of the proteins of known structure.**

**For each protein, the classification has the hierarchical levels, family, superfamily, fold and structural class.**

**Family: C-type lysozyme**

**Fold: Lysozyme-like** *common alpha+beta motif for the active site region*

**Structural class: α β**

**The structure can be identified with a six letter code (1lz1__; 5th Chain; 6th Domain**

**Protein: Lysozyme from Human (*Homo sapiens*)**

**Lineage:**

1. Root: scop
2. Class: Alpha and beta proteins (a+b) [53931]
   *Mainly antiparallel beta sheets (segregated alpha and beta regions)*
3. Fold: Lysozyme-like [53954]
   *common alpha+beta motif for the active site region*
4. Superfamily: Lysozyme-like [53955]
5. Family: C-type lysozyme [53960]
6. Protein: Lysozyme [53961]
   *ubiquitous in a variety of tissues and secretions*
7. Species: Human (*Homo sapiens*) [53969]

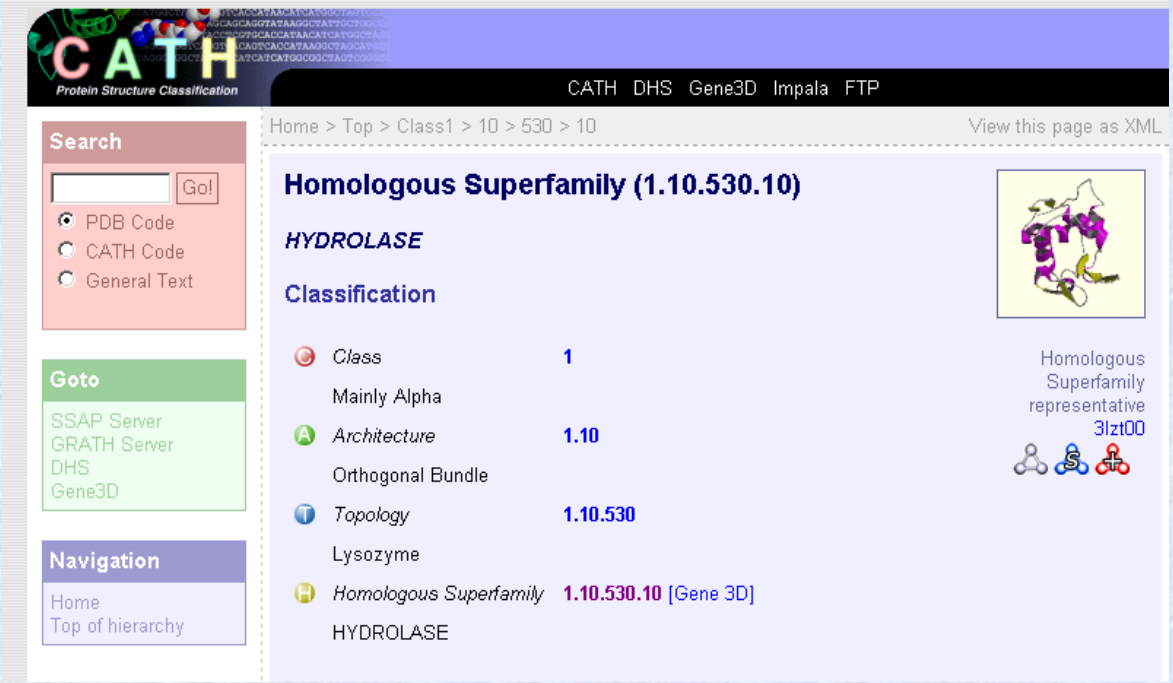
**PDB Entry Domains:**

1. 1jsf [36410] L
   *complexed with no3*
2. 1iwu
   *complexed with cl*
   1. chain a [76891] L
3. 1iwt
   *complexed with cl*
   1. chain a [76890] L
4. 1iwv
   *complexed with cl*
   1. chain a [76892] L
5. 1iww
   *complexed with cl*
   1. chain a [76893] L
6. 1jwr
   1. chain a [63311] L

# CATH: Hierarchical classification of protein domain structures

**Provides class information for all the structures in PDB.**

**The four main levels of CATH classification are protein class (C), architecture (A), topology (T) and homologous superfamily (H).**



# CATH: Hierarchical classification of protein domain structures

**Class is the simplest level, and it essentially describes the secondary structure composition of each domain.**

**Architecture summarizes the shape revealed by the orientations of the secondary structure units, such as barrels and sandwiches.**

**At the topology level, sequential connectivity is considered, such that members of the same architecture might have quite different topologies.**

**The homologous superfamilies cluster proteins with highly similar structures and functions.**

**The CATH classification for human lysozyme: Architecture of orthogonal bundle and it is assigned as all-α class proteins in CATH.**

# SCOP and CATH: Comparison

**Usually CATH and SCOP have similar classification for structural class.**

**It may be possible to have different assignments in SCOP and CATH.**

**Proteins: percentage of helical (or strand) content is high and strand (helix) content is less or vice versa.**

**In human lysozyme, the content of α-helical structures is 31% and that of β-strands is 8%.**

**Based on the high content of α-helix it was classified as all-α proteins in CATH**

**Due to the presence of α-helices and β-strands SCOP classified it as α β protein.**

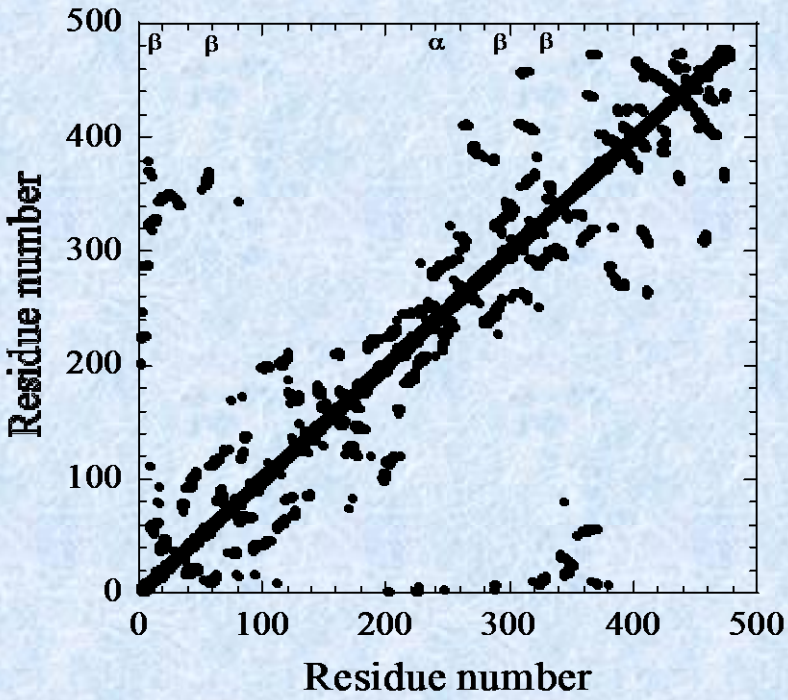
# Protein structure analysis

## Contact maps

A protein contact map represents the distance between all possible residue pairs of a three-dimensional protein structure using a binary two-dimensional matrix.

For two residues *i* and *j*, the *ij* element of the matrix is 1 if the two residues are closer than a predetermined threshold, and 0 otherwise.

Various contact definitions have been proposed: The distance between the $C_\alpha$-$C_\alpha$ atom with threshold 6-12 Å; distance between $C_\beta$-$C_\beta$ atoms with threshold 6-12 Å ($C_\alpha$ is used for Glycine); distance between the side-chain centers of mass; or consideration of any atom and lower threshold 4.5-6 Å.



# Construction of contact maps

## Cα atoms (limit of 8Å)

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ATOM | 1 | N | MET | A | 1 | 36.644 | -24.949 | 8.853 | 1.00 | 29.12 | N |
| ATOM | 2 | CA | MET | A | 1 | 36.942 | -23.581 | 8.984 | 1.00 | 19.55 | C |
| ATOM | 3 | C | MET | A | 1 | 35.712 | -22.887 | 9.526 | 1.00 | 22.27 | C |
| ATOM | 4 | O | MET | A | 1 | 34.626 | -23.375 | 9.258 | 1.00 | 18.31 | O |
| ATOM | 5 | CB | MET | A | 1 | 37.365 | -23.090 | 7.599 | 1.00 | 8.40 | C |
| ATOM | 6 | CG | MET | A | 1 | 37.639 | -21.603 | 7.644 | 1.00 | 30.36 | C |
| ATOM | 7 | SD | MET | A | 1 | 39.309 | -21.106 | 7.226 | 1.00 | 39.80 | S |
| ATOM | 8 | CE | MET | A | 1 | 40.241 | -22.126 | 8.356 | 1.00 | 44.83 | C |
| ATOM | 9 | N | ASN | A | 2 | 35.890 | -21.796 | 10.310 | 1.00 | 17.74 | N |
| ATOM | 10 | CA | ASN | A | 2 | 34.809 | -21.015 | 10.918 | 1.00 | 5.91 | C |
| ATOM | 11 | C | ASN | A | 2 | 35.236 | -19.557 | 10.931 | 1.00 | 11.34 | C |
| ATOM | 12 | O | ASN | A | 2 | 36.390 | -19.244 | 10.620 | 1.00 | 9.58 | O |
| ATOM | 13 | CB | ASN | A | 2 | 34.487 | -21.602 | 12.355 | 1.00 | 9.68 | C |
| ATOM | 14 | CG | ASN | A | 2 | 35.645 | -21.566 | 13.309 | 1.00 | 11.82 | C |
| ATOM | 15 | OD1 | ASN | A | 2 | 36.176 | -20.496 | 13.515 | 1.00 | 18.42 | O |
| ATOM | 16 | ND2 | ASN | A | 2 | 36.013 | -22.685 | 13.919 | 1.00 | 9.43 | N |
| ATOM | 17 | N | ILE | A | 3 | 34.315 | -18.689 | 11.287 | 1.00 | 6.65 | N |
| ATOM | 18 | CA | ILE | A | 3 | 34.543 | -17.262 | 11.353 | 1.00 | 10.61 | C |
| ATOM | 19 | C | ILE | A | 3 | 35.794 | -16.849 | 12.198 | 1.00 | 14.73 | C |
| ATOM | 20 | O | ILE | A | 3 | 36.530 | -15.891 | 11.865 | 1.00 | 13.47 | O |
| ATOM | 21 | CB | ILE | A | 3 | 33.235 | -16.515 | 11.748 | 1.00 | 6.98 | C |
| ATOM | 22 | CG1 | ILE | A | 3 | 33.352 | -14.965 | 11.537 | 1.00 | 7.31 | C |
| ATOM | 23 | CG2 | ILE | A | 3 | 32.955 | -16.849 | 13.220 | 1.00 | 12.46 | C |
| ATOM | 24 | CD1 | ILE | A | 3 | 33.729 | -14.545 | 10.094 | 1.00 | 7.03 | C |
| ATOM | 25 | N | PHE | A | 4 | 36.027 | -17.510 | 13.335 | 1.00 | 6.09 | N |
| ATOM | 26 | CA | PHE | A | 4 | 37.178 | -17.109 | 14.110 | 1.00 | 6.45 | C |
| ATOM | 27 | C | PHE | A | 4 | 38.472 | -17.460 | 13.428 | 1.00 | 7.23 | C |
| ATOM | 28 | O | PHE | A | 4 | 39.418 | -16.667 | 13.450 | 1.00 | 11.23 | O |
| ATOM | 29 | CB | PHE | A | 4 | 37.105 | -17.747 | 15.496 | 1.00 | 13.74 | C |
| ATOM | 30 | CG | PHE | A | 4 | 35.931 | -17.216 | 16.290 | 1.00 | 22.31 | C |
| ATOM | 31 | CD1 | PHE | A | 4 | 34.698 | -17.879 | 16.239 | 1.00 | 14.28 | C |
| ATOM | 32 | CD2 | PHE | A | 4 | 36.093 | -16.097 | 17.122 | 1.00 | 12.02 | C |
| ATOM | 33 | CE1 | PHE | A | 4 | 33.635 | -17.415 | 17.016 | 1.00 | 13.79 | C |
| ATOM | 34 | CE2 | PHE | A | 4 | 35.038 | -15.603 | 17.909 | 1.00 | 15.06 | C |
| ATOM | 35 | CZ | PHE | A | 4 | 33.840 | -16.322 | 17.873 | 1.00 | 15.21 | C |
| ATOM | 36 | N | GLU | A | 5 | 38.539 | -18.674 | 12.852 | 1.00 | 9.52 | N |
| ATOM | 37 | CA | GLU | A | 5 | 39.746 | -19.063 | 12.152 | 1.00 | 16.61 | C |
| ATOM | 38 | C | GLU | A | 5 | 39.972 | -18.177 | 10.940 | 1.00 | 14.11 | C |
| ATOM | 39 | O | GLU | A | 5 | 41.077 | -17.723 | 10.662 | 1.00 | 10.00 | O |

ATOM 2 CA MET A 1 36.942 -23.581 8.984 1.00 19.55 C

ATOM 10 CA ASN A 2 34.809 -21.015 10.918 1.00 5.91 C

ATOM 18 CA ILE A 3 34.543 -17.262 11.353 1.00 10.61 C

ATOM 26 CA PHE A 4 37.178 -17.109 14.110 1.00 6.45 C

ATOM 37 CA GLU A 5 39.746 -19.063 12.152 1.00 16.61 C

ATOM 46 CA MET A 6 39.005 -17.022 9.081 1.00 8.98 C

ATOM 54 CA LEU A 7 39.394 -13.637 10.850 1.00 9.64 C

ATOM 62 CA ARG A 8 42.437 -14.708 12.525 1.00 19.95 C

ATOM 73 CA ILE A 9 44.096 -15.463 9.190 1.00 11.21 C

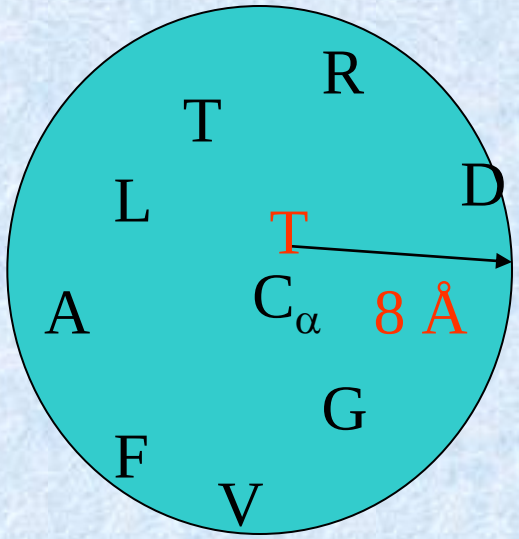
ATOM 81 CA ASP A 10 43.052 -12.038 7.999 1.00 14.11 C

Construct contact map?

# Inter-residue contacts: Definition

Basic concepts:

- The amino acid residues, which are in contact with each other in the *native structure* (8Å)
- Their respective locations in the *amino acid sequence*




DFA…LTRTDAH…STA

# Classification

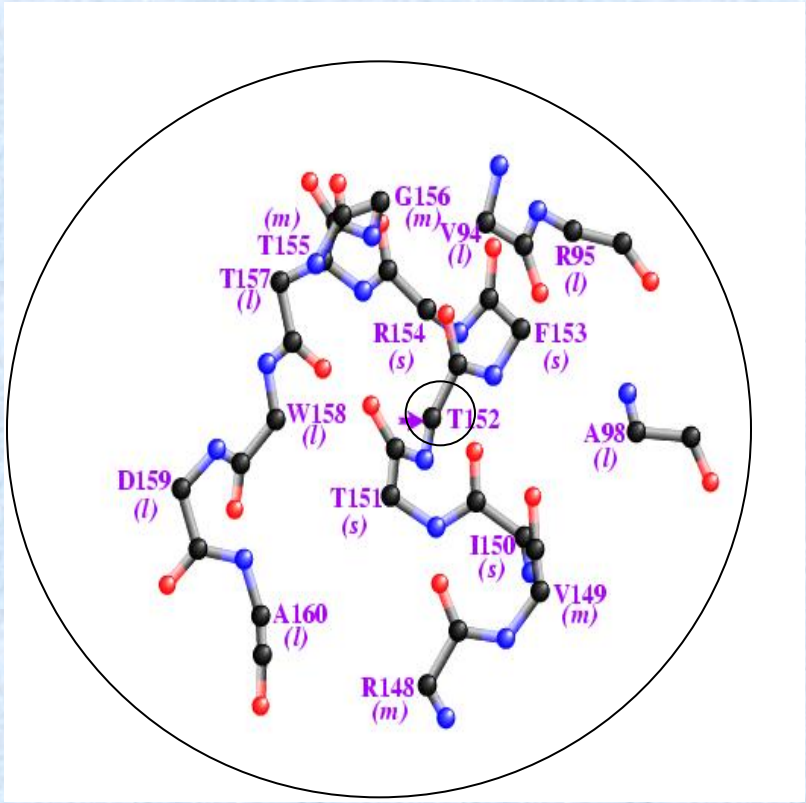
- Distance of separation between amino acid residues in a polypeptide chain

s: short (< 3 residues)

m: medium ( 3 or 4 residues)

l: long (> 4 residues)

(further divided into small bins of 10 residues)




**Gromiha, MM and Selvaraj, S. (2004) Prog. Biophys. Mol. Biol. 86, 235-277.**

# Representation

Inter-residue contacts can be pictorially represented by contact maps.

Short-range contacts:

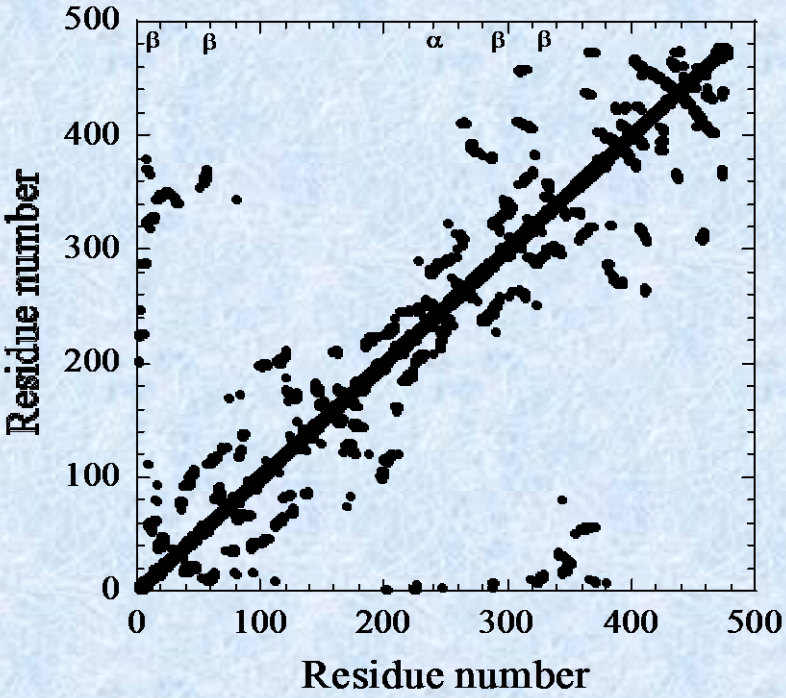
diagonal residues

Medium-range contacts:

close to diagonal residues

Long-range contacts:

far away from the diagonal.



# Secondary structure

**DSSP: Dictionary of Secondary Structure of Proteins**

**Algorithm:**

**Hydrogen bonding pattern** **ftp://ftp.cmbi.kun.nl/**

**Download the program dssp-2**

**Command:**

**./dssp-2 input output**

```
==== Secondary Structure Definition by the program DSSP, updated CMBI version by ElmK / April 1,2000 ==== DATE=2-APR-2004      .
REFERENCE W. KABSCH AND C.SANDER, BIOPOLYMERS 22 (1983) 2577-2637                                                          .
HEADER    HYDROLASE (O-GLYCOSYL)                    12-OCT-84   1LZ1                                                        .
COMPND    LYSOZYME (E.C.3.2.1.17)                                                                                           .
SOURCE    HUMAN (HOMO $SAPIENS)                                                                                             .
AUTHOR    P.J.ARTYMIUK,C.C.F.BLAKE                                                                                          .
  130  1  4  4  0 TOTAL NUMBER OF RESIDUES, NUMBER OF CHAINS, NUMBER OF SS-BRIDGES(TOTAL,INTRACHAIN,INTERCHAIN)               .
  6787.0   ACCESSIBLE SURFACE OF PROTEIN (ANGSTROM**2)                                                                      .
   89 68.5   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(J)  , SAME NUMBER PER 100 RESIDUES                            .
    2  1.5   TOTAL NUMBER OF HYDROGEN BONDS IN     PARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES                            .
   12  9.2   TOTAL NUMBER OF HYDROGEN BONDS IN ANTIPARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES                            .
    2  1.5   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-5), SAME NUMBER PER 100 RESIDUES                            .
    1  0.8   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-4), SAME NUMBER PER 100 RESIDUES                            .
    1  0.8   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-3), SAME NUMBER PER 100 RESIDUES                            .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-2), SAME NUMBER PER 100 RESIDUES                            .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-1), SAME NUMBER PER 100 RESIDUES                            .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+0), SAME NUMBER PER 100 RESIDUES                            .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+1), SAME NUMBER PER 100 RESIDUES                            .
    9  6.9   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+2), SAME NUMBER PER 100 RESIDUES                            .
   25 19.2   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+3), SAME NUMBER PER 100 RESIDUES                            .
   34 26.2   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+4), SAME NUMBER PER 100 RESIDUES                            .
    5  3.8   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+5), SAME NUMBER PER 100 RESIDUES                            .
  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30     *** HISTOGRAMS OF ***       .
  0  0  0  0  0  1  0  0  0  1  1  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    RESIDUES PER ALPHA HELIX     .
  1  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  0    PARALLEL BRIDGES PER LADDER  .
  2  1  0  1  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    ANTIPARALLEL BRIDGES PER LADDER .
  0  1  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    LADDERS PER SHEET            .
  #  RESIDUE AA STRUCTURE BP1 BP2  ACC     N-H-->O    O-->H-N    N-H-->O    O-->H-N    TCO  KAPPA ALPHA  PHI   PSI    X-CA   Y-CA   Z-CA
    1    1  K              0   0   63     0, 0.0    39,-2.6     0, 0.0     2,-0.7  0.000 360.0 360.0 360.0 146.7    1.9   20.6   21.1
    2    2  V  B       -A  39   0A  99    37,-0.3    37,-0.2    38,-0.1     2,-0.1 -0.949 360.0-145.6-104.2 112.9    4.6   19.8   18.5
    3    3  F          -   0   0   10    35,-2.4     2,-0.2    -2,-0.7     3, 0.0 -0.453  10.1-122.2 -70.2 150.5    7.4   22.4   18.9
    4    4  E        >  -   0   0  146    -2,-0.1     4,-2.6     1,-0.1     5,-0.3 -0.657  34.8-107.1 -82.3 155.8    9.4   23.7   16.0
    5    5  R  H     > S+   0   0   91    -2,-0.2     4,-2.1     1,-0.2     5,-0.1  0.886 115.0  37.2 -56.3 -50.8   13.1   23.1   16.5
    6    6  a  H     > S+   0   0   27     1,-0.2     4,-2.5     2,-0.2    -1,-0.2  0.826 112.9  58.3 -78.4 -24.6   14.3   26.7   17.2
    7    7  E  H     > S+   0   0   58     1,-0.2     4,-2.3     2,-0.2    -1,-0.2  0.907 109.6  45.4 -60.1 -45.4   11.1   27.5   19.2
    8    8  L  H     X S+   0   0    1    -4,-2.6     4,-3.2     2,-0.2     5,-0.3  0.903 110.4  54.1 -62.5 -42.1   12.0   24.6   21.5
    9    9  A  H     X S+   0   0    0    -4,-2.1     4,-2.1    -5,-0.3    -2,-0.2  0.937 111.4  43.9 -61.1 -44.3   15.7   25.6   21.7
   10   10  R  H     X S+   0   0  114    -4,-2.5     4,-2.5     2,-0.2    -1,-0.2  0.906 113.6  51.9 -67.4 -34.7   14.8   29.2   22.8
   11   11  T  H     X S+   0   0   29    -4,-2.3     4,-1.3     1,-0.2    -2,-0.2  0.910 111.0  46.7 -66.8 -47.6   12.2   27.8   25.3
   12   12  L  H    <>S+   0   0    0    -4,-3.2     5,-2.5     2,-0.2     6,-0.3  0.845 111.4  52.9 -61.5 -37.6   14.7   25.4   26.8
   13   13  K  H    ><5S+   0   0   89    -4,-2.1     3,-1.8    -5,-0.3     5,-0.3  0.970 109.3  47.8 -61.0 -44.3   17.3   28.3   27.0
   14   14  R  H    3<5S+   0   0  [illegible]  -4,-2.5    -1,-0.2     1,-0.3    -2,-0.2  0.809 106.9  58.1 -67.8 -23.4   14.9   30.5   28.9
   15   15  L  T    3<5S-   0   0   49    -4,-1.3    -1,-0.3    -5,-0.2    -2,-0.2  0.399 121.3-107.0 -91.0    6.3   14.0   27.6   31.3
   16   16  G  T    < 5S+   0   0   40    -3,-1.8    -3,-0.2    -4,-0.3    -2,-0.1  0.719  79.9 128.7  86.8   24.2   17.7   27.4   32.3
   17   17  M        > < +   0   0    0    -5,-2.5     3,-1.7     2,-0.1     2,-0.6  0.625  35.2 105.3 -87.8 -19.2   18.7   24.2   30.5
   18   18  D  T    3  S-   0   0   78    -6,-0.3     6,-0.2    -5,-0.3    4,-0.1 -0.578 104.8  -7.7 -68.4 107.8   21.8   25.4   28.6
   19   19  G  T    <  S+   0   0   44     4,-2.4     3,-2.1    -2,-0.6     2,-0.4  0.604  88.0 164.0  88.8    6.3   24.8   23.9   30.6
   20   20  Y  B  X  S-B  23  0B  54    -3,-1.7     3,-1.6     3,-0.8    -1,-0.3 -0.584  80.4 -10.2 -63.8 126.7   22.6   22.6   33.4
   21   21  R  T  3  S-   0   0 155    -2,-0.4    -1,-0.3     1,-0.2     3,-0.1  0.784 134.6 -56.6  46.1  39.0   24.8   20.0   35.2
   22   22  G  T  <  S+   0   0   75    -3,-2.1     2,-0.7     1,-0.2    -1,-0.2  0.507 104.9 132.1  83.1   6.6   27.2   20.4   32.3
   23   23  I  B  <    -B  20  0B  15    -3,-1.6    -4,-2.4    -6,-0.1    -3,-0.8 -0.753  50.3-136.2-103.2 121.5   24.8   19.5   29.6
```

Secondary structure

Solvent accessibility

Hydrogen bonding partners/ Electrostatic energy

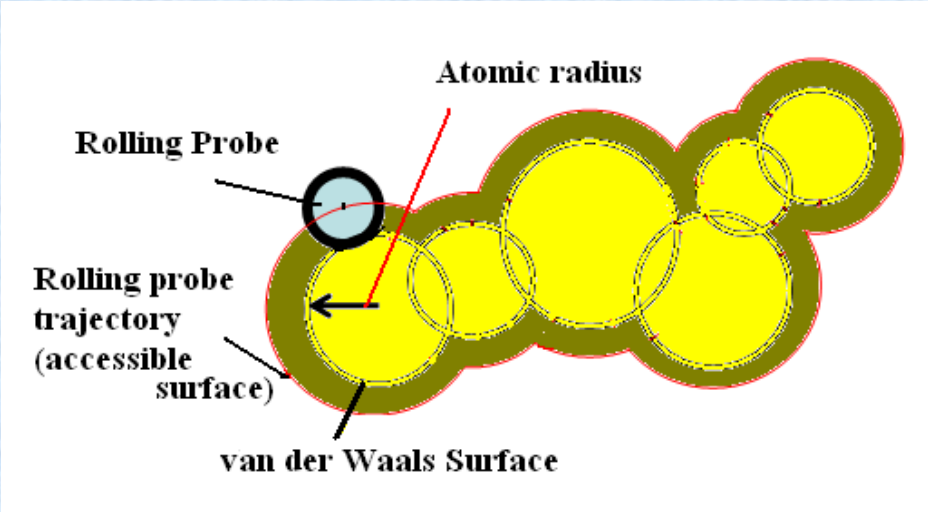
# Solvent accessibility

The solvent accessible surface area is defined as the locus of the centre of the solvent molecule as it rolls over the van der Waals surface of the protein.

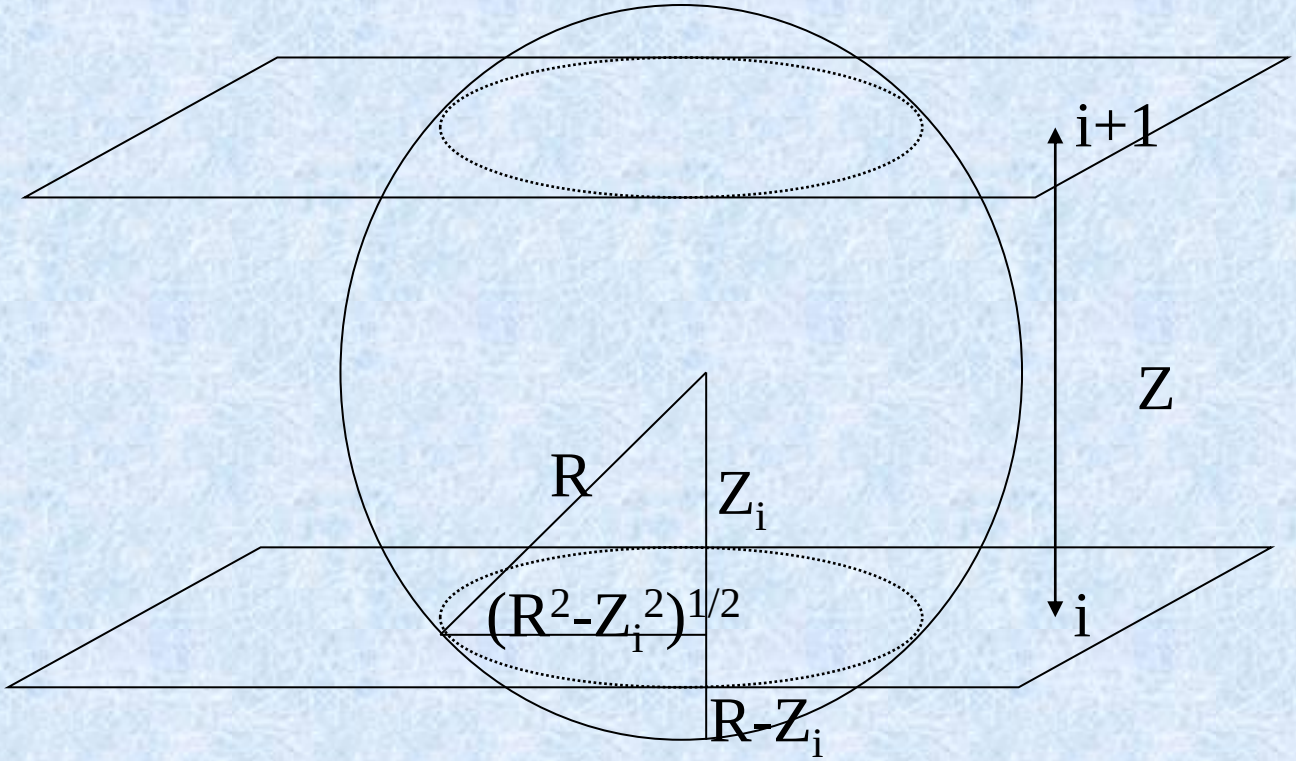
Generally, a sphere of water is assumed to be the solvent molecule with a radius 1.4 Å.

The solute molecule is represented by a set of interlocking spheres of appropriate van der Waals radii assigned to each atom

The solvent molecule is rolled along the envelope of the van der Waals surface at planes conveniently sectioned.




The accessible surface area of an atom of radius *r* is then the area on the surface of sphere of radius $R = r + r_{solv}$ on each point of which the center of the solvent molecule can be placed in contact with this atom without penetrating any other atoms of the solute molecule.




- **The solvent accessible surface area (ASA) is calculated using the formula (Lee and Richards, 1971):**

- **ASA = $[R/(R^2 - Z_i^2)^{1/2}]\ L_i\ .\ D$; D = Z/2 + 'Z**

- **where, $L_i$ is the length of the arc computed on a given section *i*, $Z_i$ is the perpendicular distance from the center of the sphere to the section *i*, *Z* is the spacing between the sections and *'Z* is *Z*/2 or $R-Z_i$, whichever is smaller. Summation is over all of the arcs drawn for the given atom.**

# Methods

ACCESS: Lee and Richards (1978)

NACCESS: Hubbard and Thornton (1993)

ASC: Eisenhaber and Argos (1993)

DSSP: Kabsch and Sandor (1983)

GETAREA: Fraczkiewicz and Braun (1998)

| Comparison index | ACCESS | DSSP | NACCESS | ASC | GETAREA |
|---|---|---|---|---|---|
| Standalone executable availability | Yes | Yes | Licensed | Yes | No |
| Online calculations/ database | No | Yes | No | Yes | Yes |
| Polar and nonpolar area | No | No | Yes | No | Yes |
| Atom-wise surface area | Yes | No | Yes | Yes | Yes |
| Source code availability | No | Yes | No | Yes | No |
| Choice of probe radius | Yes | No | Yes | Yes | No |
| Choice of van der Waals and other parameters | Yes | No | Yes | Yes | By Manual editing |
| Secondary structure | No | Yes | No | No | No |
| Reference | and Richards, 1978 | Kabsch and Sander, 1983 | Hubbard and Thornton, 1993 | Eisenhaber and , 1993 | Fraczkiewicz and Braun, 1998 |

```
==== Secondary Structure Definition by the program DSSP, updated CMBI version by ElmK / April 1,2000 ==== DATE=2-APR-2004      .
REFERENCE W. KABSCH AND C.SANDER, BIOPOLYMERS 22 (1983) 2577-2637                                                          .
HEADER    HYDROLASE (O-GLYCOSYL)                    12-OCT-84   1LZ1                                                        .
COMPND    LYSOZYME (E.C.3.2.1.17)                                                                                           .
SOURCE    HUMAN (HOMO $SAPIENS)                                                                                             .
AUTHOR    P.J.ARTYMIUK,C.C.F.BLAKE                                                                                          .
  130  1  4  4  0 TOTAL NUMBER OF RESIDUES, NUMBER OF CHAINS, NUMBER OF SS-BRIDGES(TOTAL,INTRACHAIN,INTERCHAIN)               .
  6787.0   ACCESSIBLE SURFACE OF PROTEIN (ANGSTROM**2)                                                                      .
   89 68.5   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(J)  , SAME NUMBER PER 100 RESIDUES                            .
    2  1.5   TOTAL NUMBER OF HYDROGEN BONDS IN     PARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES                            .
   12  9.2   TOTAL NUMBER OF HYDROGEN BONDS IN ANTIPARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES                            .
    2  1.5   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-5), SAME NUMBER PER 100 RESIDUES                            .
    1  0.8   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-4), SAME NUMBER PER 100 RESIDUES                            .
    1  0.8   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-3), SAME NUMBER PER 100 RESIDUES                            .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-2), SAME NUMBER PER 100 RESIDUES                            .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-1), SAME NUMBER PER 100 RESIDUES                            .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+0), SAME NUMBER PER 100 RESIDUES                            .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+1), SAME NUMBER PER 100 RESIDUES                            .
    9  6.9   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+2), SAME NUMBER PER 100 RESIDUES                            .
   25 19.2   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+3), SAME NUMBER PER 100 RESIDUES                            .
   34 26.2   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+4), SAME NUMBER PER 100 RESIDUES                            .
    5  3.8   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+5), SAME NUMBER PER 100 RESIDUES                            .
  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30     *** HISTOGRAMS OF ***       .
  0  0  0  0  0  1  0  0  0  1  1  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    RESIDUES PER ALPHA HELIX     .
  1  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  0    PARALLEL BRIDGES PER LADDER  .
  2  1  0  1  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    ANTIPARALLEL BRIDGES PER LADDER .
  0  1  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    LADDERS PER SHEET            .
  #  RESIDUE AA STRUCTURE BP1 BP2  ACC     N-H-->O    O-->H-N    N-H-->O    O-->H-N    TCO  KAPPA ALPHA  PHI   PSI    X-CA   Y-CA   Z-CA
    1    1  K              0   0   63     0, 0.0    39,-2.6     0, 0.0     2,-0.7  0.000 360.0 360.0 360.0 146.7    1.9   20.6   21.1
    2    2  V  B       -A  39   0A  99    37,-0.3    37,-0.2    38,-0.1     2,-0.1 -0.949 360.0-145.6-104.2 112.9    4.6   19.8   18.5
    3    3  F          -   0   0   10    35,-2.4     2,-0.2    -2,-0.7     3, 0.0 -0.453  10.1-122.2 -70.2 150.5    7.4   22.4   18.9
    4    4  E        >  -   0   0  146    -2,-0.1     4,-2.6     1,-0.1     5,-0.3 -0.657  34.8-107.1 -82.3 155.8    9.4   23.7   16.0
    5    5  R  H     > S+   0   0   91    -2,-0.2     4,-2.1     1,-0.2     5,-0.1  0.886 115.0  37.2 -56.3 -50.8   13.1   23.1   16.5
    6    6  a  H     > S+   0   0   27     1,-0.2     4,-2.5     2,-0.2    -1,-0.2  0.826 112.9  58.3 -78.4 -24.6   14.3   26.7   17.2
    7    7  E  H     > S+   0   0   58     1,-0.2     4,-2.3     2,-0.2    -1,-0.2  0.907 109.6  45.4 -60.1 -45.4   11.1   27.5   19.2
    8    8  L  H     X S+   0   0    1    -4,-2.6     4,-3.2     2,-0.2     5,-0.3  0.903 110.4  54.1 -62.5 -42.1   12.0   24.6   21.5
    9    9  A  H     X S+   0   0    0    -4,-2.1     4,-2.1    -5,-0.3    -2,-0.2  0.937 111.4  43.9 -61.1 -44.3   15.7   25.6   21.7
   10   10  R  H     X S+   0   0  114    -4,-2.5     4,-2.5     2,-0.2    -1,-0.2  0.906 113.6  51.9 -67.4 -34.7   14.8   29.2   22.8
   11   11  T  H     X S+   0   0   29    -4,-2.3     4,-1.3     1,-0.2    -2,-0.2  0.910 111.0  46.7 -66.8 -47.6   12.2   27.8   25.3
   12   12  L  H    <>S+   0   0    0    -4,-3.2     5,-2.5     2,-0.2     6,-0.3  0.845 111.4  52.9 -61.5 -37.6   14.7   25.4   26.8
   13   13  K  H    ><5S+   0   0   89    -4,-2.1     3,-1.8    -5,-0.3     5,-0.3  0.970 109.3  47.8 -61.0 -44.3   17.3   28.3   27.0
   14   14  R  H    3<5S+   0   0  [illegible]  -4,-2.5    -1,-0.2     1,-0.3    -2,-0.2  0.809 106.9  58.1 -67.8 -23.4   14.9   30.5   28.9
   15   15  L  T    3<5S-   0   0   49    -4,-1.3    -1,-0.3    -5,-0.2    -2,-0.2  0.399 121.3-107.0 -91.0    6.3   14.0   27.6   31.3
   16   16  G  T    < 5S+   0   0   40    -3,-1.8    -3,-0.2    -4,-0.3    -2,-0.1  0.719  79.9 128.7  86.8   24.2   17.7   27.4   32.3
   17   17  M        > < +   0   0    0    -5,-2.5     3,-1.7     2,-0.1     2,-0.6  0.625  35.2 105.3 -87.8 -19.2   18.7   24.2   30.5
   18   18  D  T    3  S-   0   0   78    -6,-0.3     6,-0.2    -5,-0.3    4,-0.1 -0.578 104.8  -7.7 -68.4 107.8   21.8   25.4   28.6
   19   19  G  T    <  S+   0   0   44     4,-2.4     3,-2.1    -2,-0.6     2,-0.4  0.604  88.0 164.0  88.8    6.3   24.8   23.9   30.6
   20   20  Y  B  X  S-B  23  0B  54    -3,-1.7     3,-1.6     3,-0.8    -1,-0.3 -0.584  80.4 -10.2 -63.8 126.7   22.6   22.6   33.4
   21   21  R  T  3  S-   0   0 155    -2,-0.4    -1,-0.3     1,-0.2     3,-0.1  0.784 134.6 -56.6  46.1  39.0   24.8   20.0   35.2
   22   22  G  T  <  S+   0   0   75    -3,-2.1     2,-0.7     1,-0.2    -1,-0.2  0.507 104.9 132.1  83.1   6.6   27.2   20.4   32.3
   23   23  I  B  <    -B  20  0B  15    -3,-1.6    -4,-2.4    -6,-0.1    -3,-0.8 -0.753  50.3-136.2-103.2 121.5   24.8   19.5   29.6
```

Secondary structure

Solvent accessibility

Hydrogen bonding partners/ Electrostatic energy

# GETAREA

## Calculation of Solvent Accessible Surface Areas, Atomic Solvation Energies and Their Gradients for Macromolecules


**Robert Fraczkiewicz and Werner Braun**

**Sealy Center for Structural Biology**
**University of Texas Medical Branch**
**Galveston, TX 77555, USA**


Quite often a biomolecular researcher wants to quickly calculate solvent accessible surface area or solvation energy of, for example, a protein molecule but does not have time/resources/skills/will to find and install an appropriate software. GETAREA, our efficient method of calculating the solvent accesible surface area [1,2] implemented in program FANTOM, can be directly accessed through this form. An on-line manual is available. Atomic coordinates should be supplied in PDB format. Please cite reference [1] in publications that use our service. Comments are welcome, please mail them to webraun@utmb.edu. Sample Getarea output: getarea.pdf

**References:**


1. Fraczkiewicz, R. and Braun, W. (1998) "Exact and Efficient Analytical Calculation of the Accessible Surface Areas and Their Gradients for Macromolecules" *J. Comp. Chem.*, **19**, 319-333.
2. Fraczkiewicz, R. and Braun, W. "A New Efficient Algorithm for Calculating Solvent Accessible Surface Areas of Macromolecules" presented at the *Third Electronic Computational Chemistry Conference*; Northern Illinois University; November 1996; World Wide Web.


**Please select your PDB File:** Choose File 2PTL.pdb

**Please Enter radius of the water probe(A):** 1.4

**Do you want gradient in calculations:** n y or n, Default is n

**Please enter your Email Address:** gromiha@iitm.ac.in For information purpose only

**Select desired level of output:** 2. Area/energy per residue

Submit for Analysis | Reset to Default values

| | | | | | |
|---|---|---|---|---|---|
| GLN | 32 | 98.49 | 34.65 | 6.46 | 9 |
| THR | 33 | 76.88 | 62.60 | 12.26 | 6 |
| ALA | 34 | 6.64 | 6.64 | 6.24 | |
| GLU | 35 | 119.15 | 44.05 | 17.76 | 10 |
| PHE | 36 | 20.53 | 20.53 | 3.05 | 1 |
| LYS | 37 | 109.82 | 52.86 | 13.38 | 9 |

| | | | | |
|---|---|---|---|---|
| 33 | CA | LYS | 3 | 9. |
| 34 | C | LYS | 3 | 0. |
| 35 | O | LYS | 3 | 3. |
| 36 | CB | LYS | 3 | 25. |
| 37 | CG | LYS | 3 | 13. |
| 38 | CD | LYS | 3 | 35. |

```
APOLAR area/energy          =        5001.36
UNKNOW area/energy          =           0.00
-------------------------------------------
Total  area/energy          =        8496.85
-------------------------------------------
Number of surface atoms     =  723
Number of buried atoms      =  586
```

# Pictorial representation: ASAView

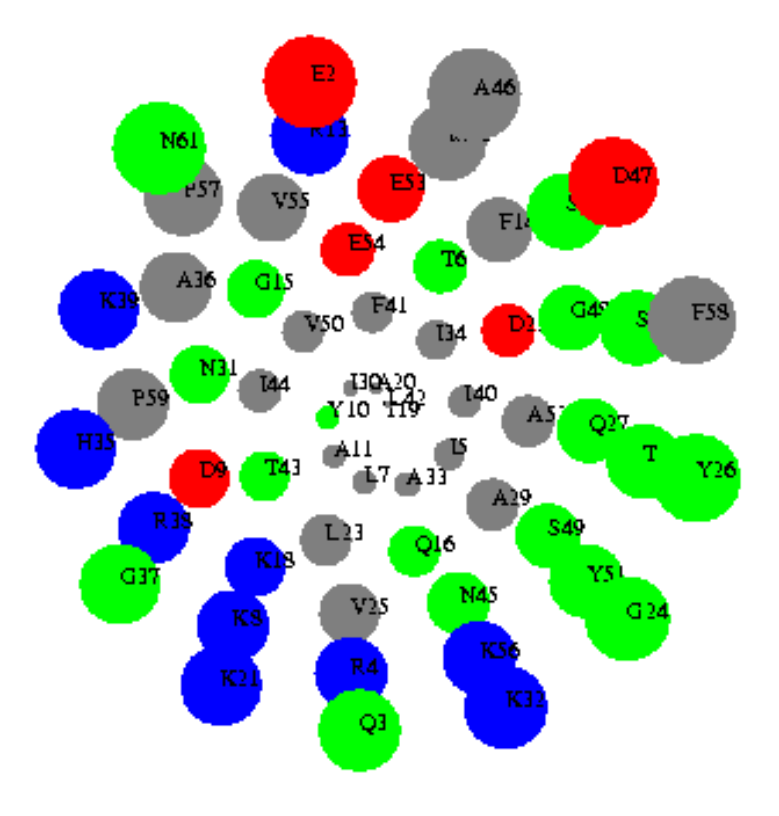
ASAview is an algorithm and a database of schematic representations of solvent accessibility of residues in a protein.

In this program, a characteristic two-dimensional spiral plot of solvent accessibility has been implemented for providing a convenient graphical view of residues in terms of their exposed surface areas.

Further, the sequential plots are also provided in the form of bar charts.

Online plots of the proteins included in the entire Protein Data Bank (PDB), are provided for the whole protein as well as their chains separately.

# ASAview

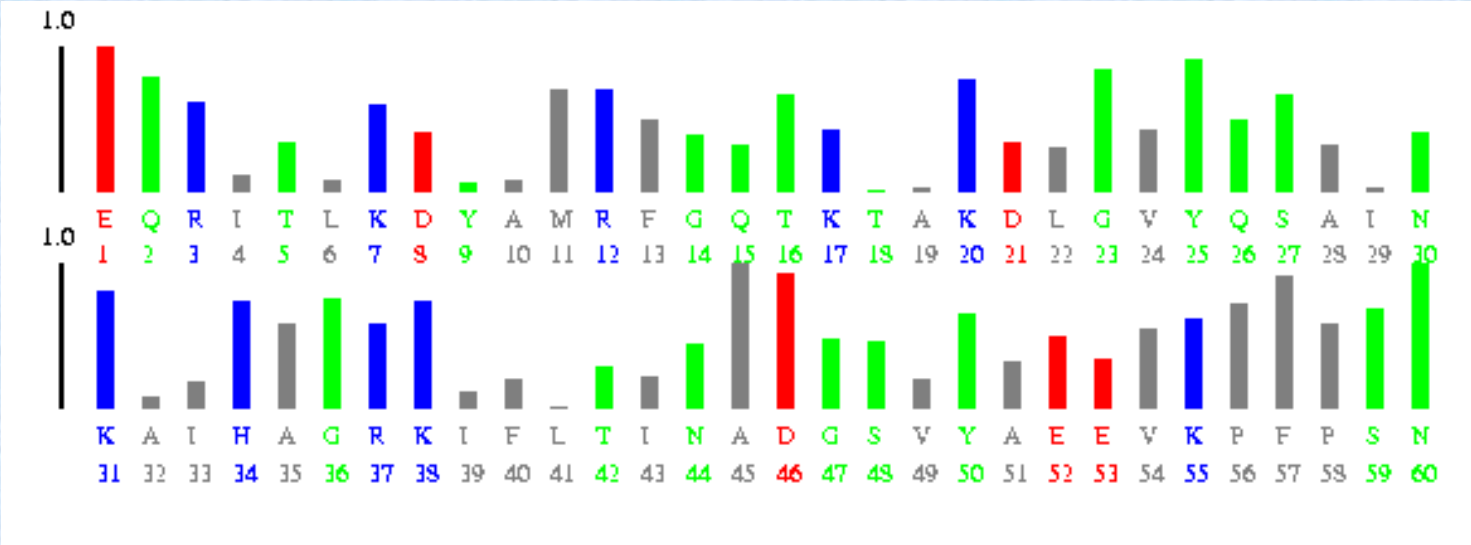



## Spiral view

**blue, red, green, gray and yellow colors indicates the positively charged, negatively charged, polar, non-polar and Cys residues, respectively. The size of the sphere shows the relative ASA**

**S. Ahmad, M.M. Gromiha and A. Sarai, (2004) BMC Bioinformatics**

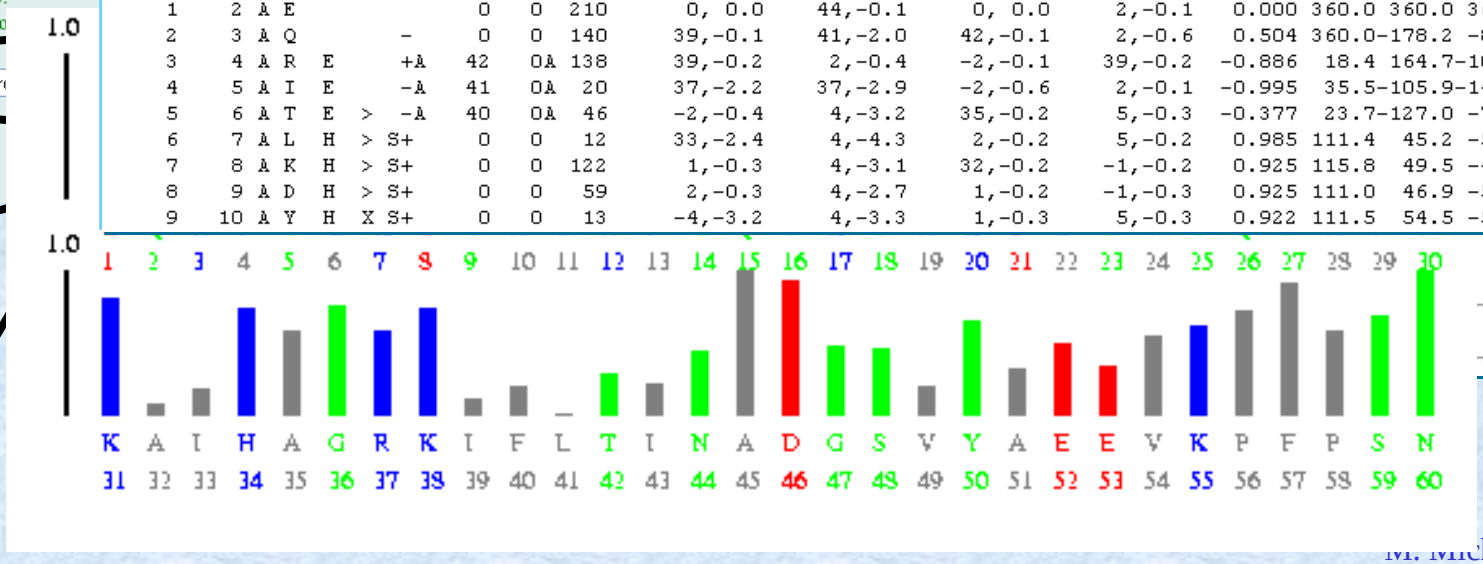
# ASAview


1.0
E Q R I T L K D Y A M R F G Q T K T A K D L G V Y Q S A I N
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
1.0
K A I H A G R K I F L T I N A D G S V Y A E E V K P F P S N
31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60

```
==== Secondary Structure Definition by the program DSSP, updated CMBI version by ElmK / April 1,2000 ==== DATE=2
REFERENCE W. KABSCH AND C.SANDER, BIOPOLYMERS 22 (1983) 2577-2637
HEADER    GENE REGULATION/DNA                     22-APR-98   6CRO
COMPND    2 MOLECULE: DNA (5'-
SOURCE    2 SYNTHETIC: YES;
AUTHOR    R.A.ALBRIGHT,B.W.MATTHEWS
   60  1  0  0  0 TOTAL NUMBER OF RESIDUES, NUMBER OF CHAINS, NUMBER OF SS-BRIDGES(TOTAL,INTRACHAIN,INTERCHAIN)
  4773.0   ACCESSIBLE SURFACE OF PROTEIN (ANGSTROM**2)
   40 66.7   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(J)  , SAME NUMBER PER 100 RESIDUES
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS IN     PARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES
   11 18.3   TOTAL NUMBER OF HYDROGEN BONDS IN ANTIPARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-5), SAME NUMBER PER 100 RESIDUES
    1  1.7   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-4), SAME NUMBER PER 100 RESIDUES
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-3), SAME NUMBER PER 100 RESIDUES
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-2), SAME NUMBER PER 100 RESIDUES
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-1), SAME NUMBER PER 100 RESIDUES
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+0), SAME NUMBER PER 100 RESIDUES
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+1), SAME NUMBER PER 100 RESIDUES
    3  5.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+2), SAME NUMBER PER 100 RESIDUES
    4  6.7   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+3), SAME NUMBER PER 100 RESIDUES
   20 33.3   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+4), SAME NUMBER PER 100 RESIDUES
    2  3.3   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+5), SAME NUMBER PER 100 RESIDUES
  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30     *** HISTOGRAMS OF
  0  0  0  0  0  0  0  2  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    RESIDUES PER ALPHA
  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  0  0    PARALLEL BRIDGES P
  0  0  1  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    ANTIPARALLEL BRIDG
  0  1  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    LADDERS PER SHEET
  #  RESIDUE AA STRUCTURE BP1 BP2  ACC     N-H-->O    O-->H-N    N-H-->O    O-->H-N    TCO  KAPPA ALPHA  PHI    P
    1    2 A E              0   0  210     0, 0.0    44,-0.1     0, 0.0     2,-0.1   0.000 360.0 360.0 360.0  8
    2    3 A Q        -     0   0  140    39,-0.1    41,-2.0    42,-0.1     2,-0.6   0.504 360.0-178.2 -87.1 14
    3    4 A R  E     +A   42  0A 138    39,-0.2     2,-0.4    -2,-0.1    39,-0.2  -0.886  18.4 164.7-100.9 12
    4    5 A I  E     -A   41  0A  20    37,-2.2    37,-2.9    -2,-0.6     2,-0.1  -0.995  35.5-105.9-146.6 15
    5    6 A T  E  >  -A   40  0A  46    -2,-0.4     4,-3.2    35,-0.2     5,-0.3  -0.377  23.7-127.0 -71.5 14
    6    7 A L  H  > S+     0   0  12    33,-2.4     4,-4.3     2,-0.2     5,-0.2   0.985 111.4  45.2 -58.0 -5
    7    8 A K  H  > S+     0   0 122     1,-0.3     4,-3.1    32,-0.2    -1,-0.2   0.925 115.8  49.5 -45.6 -5
    8    9 A D  H  > S+     0   0  59     2,-0.3     4,-2.7     1,-0.2    -1,-0.3   0.925 111.0  46.9 -50.4 -5
    9   10 A Y  H  X S+     0   0  13    -4,-3.2     4,-3.3     1,-0.3     5,-0.3   0.922 111.5  54.5 -57.6 -4
```

```
E2 100.0
Q3 78.4
R4 60.3
I5 10.8
T6 33.2
L7 6.6
K8 59.3
D9 40.9
Y10 6.1
A11 6.4
M12 70.0
R13 70.3
F14 48.3
G15 38.1
Q16 30.8
T17 65.6
K18 41.3
T19 0.0
A20 1.8
K21 75.8
D22 32.5
L23 30.6
G24 83.9
V25 41.6
Y26 90.8
Q27 48.2
S28 65.7
A29 31.8
I30 2.7
N31 39.6
K32 80.7
A33 7.3
I34 18.4
H35 74.2
A36 58.1
G37 75.0
R38 58.5
K39 73.9
```

Z-CA
89.8
90.3
87.0
87.0
85.2
86.5
85.9
87.4
90.4



# Percentage Accessibility

**Percentage accessibility**

**Ratio between**

**Accessible surface area computed with 3D structure and**

**Accessible surface area in extended state**

**Gly-X-Gly or Ala-X-Ala**

The values are Ala-110.2; Asp-144.1; Cys-140.4; Glu-174.7; Phe-200.7; Gly-78.7; His-181.9; Ile-185.0; Lys-205.7; Leu-183.1; Met-200.1; Asn-146.4; Pro-141.9; Gln-178.6; Arg-229.0; Ser-117.2; Thr-138.7; Val-153.7; Trp-240.5; Tyr-213.7 (the units are in A**2).

**ASA less than 5%: Buried**

**ASA between 5% and 20%: Partially buried**

**ASA between 20% and 50%: Partially exposed**

**ASA more than 50%: Exposed.**

# Advanced Bioinformatics

## Development of algorithms
## Genomic analysis

## Machine learning techniques
## Features, applications and validation procedures

# Research themes

## Classification problems

Helix, strand and loop (secondary structures)

DNA/RNA binding proteins or NOT

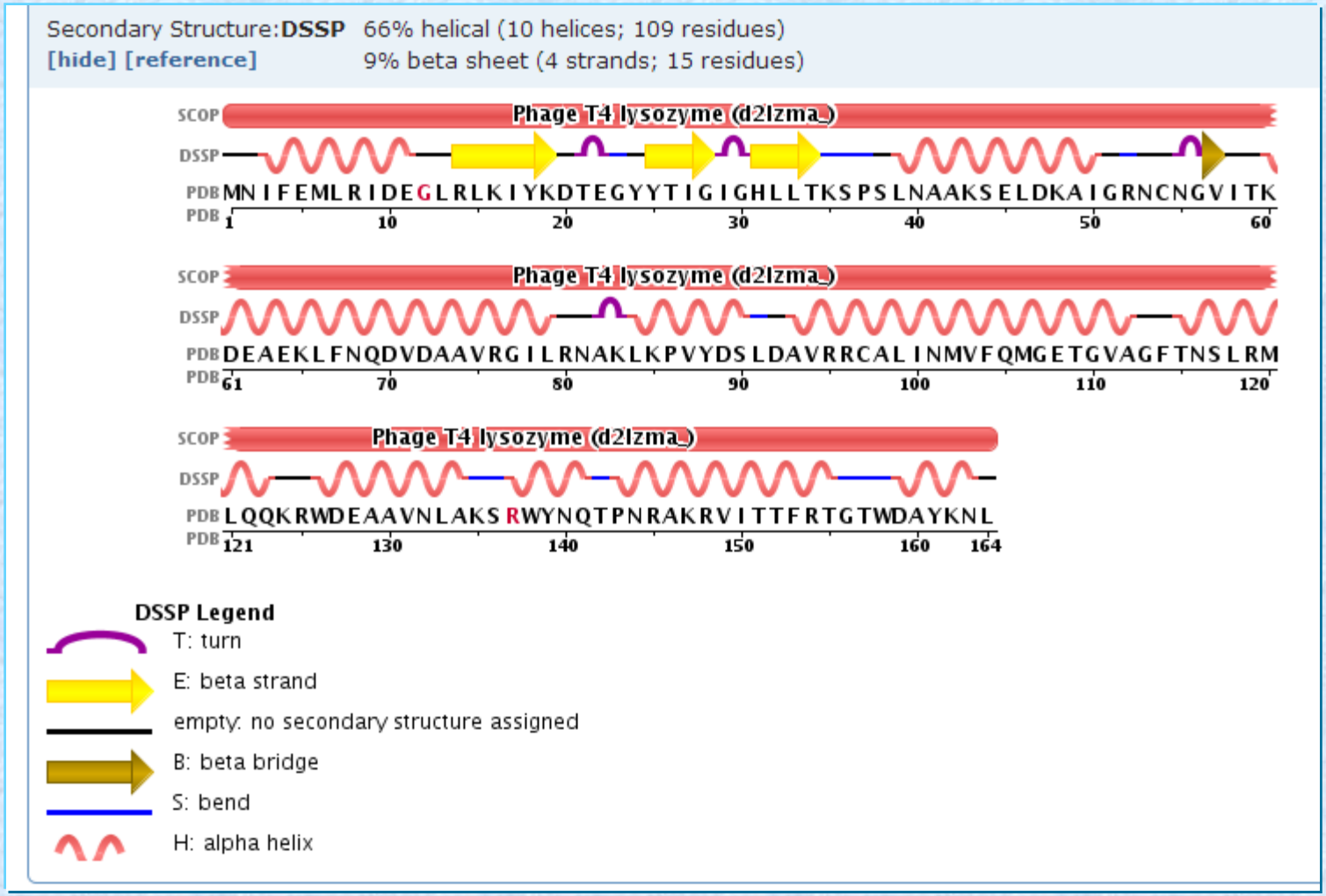
Binding sites or NOT

Residues in membrane or NOT

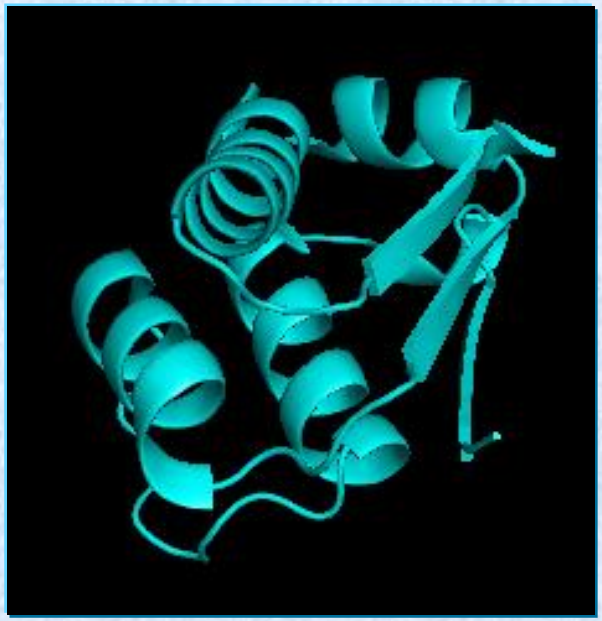
Perform specific function (transporters) or NOT

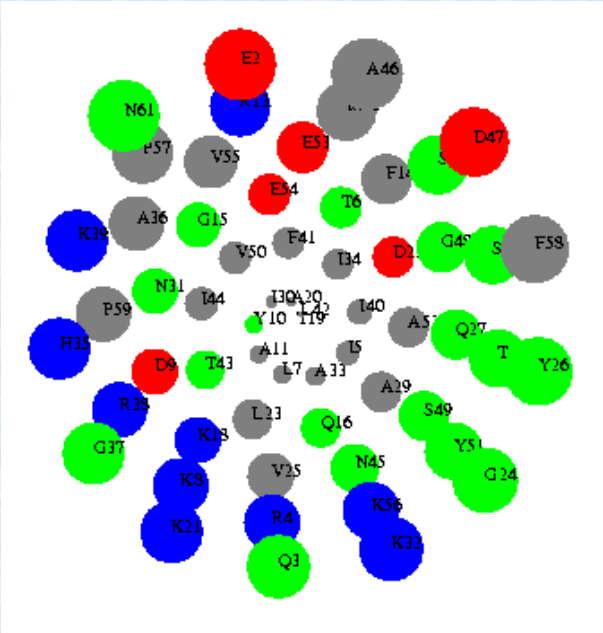
Mutants stabilize or destabilize a protein

# Protein secondary structure



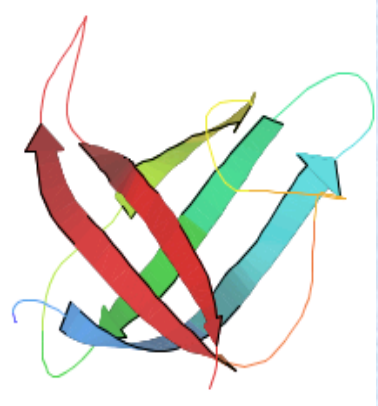
# Solvent accessibility



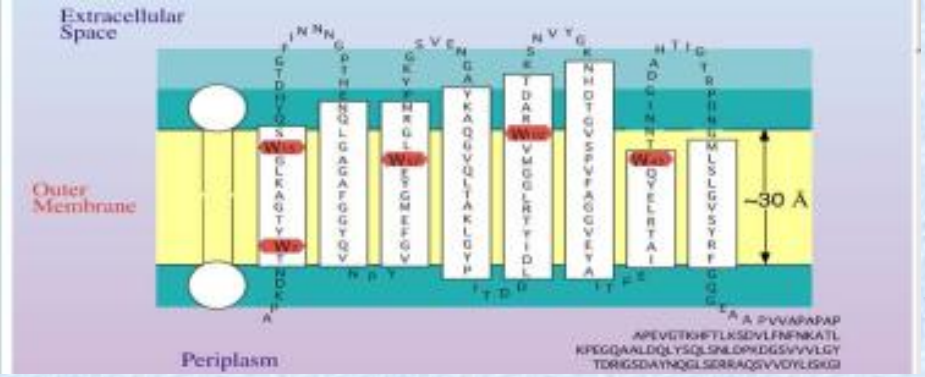


# Membrane proteins

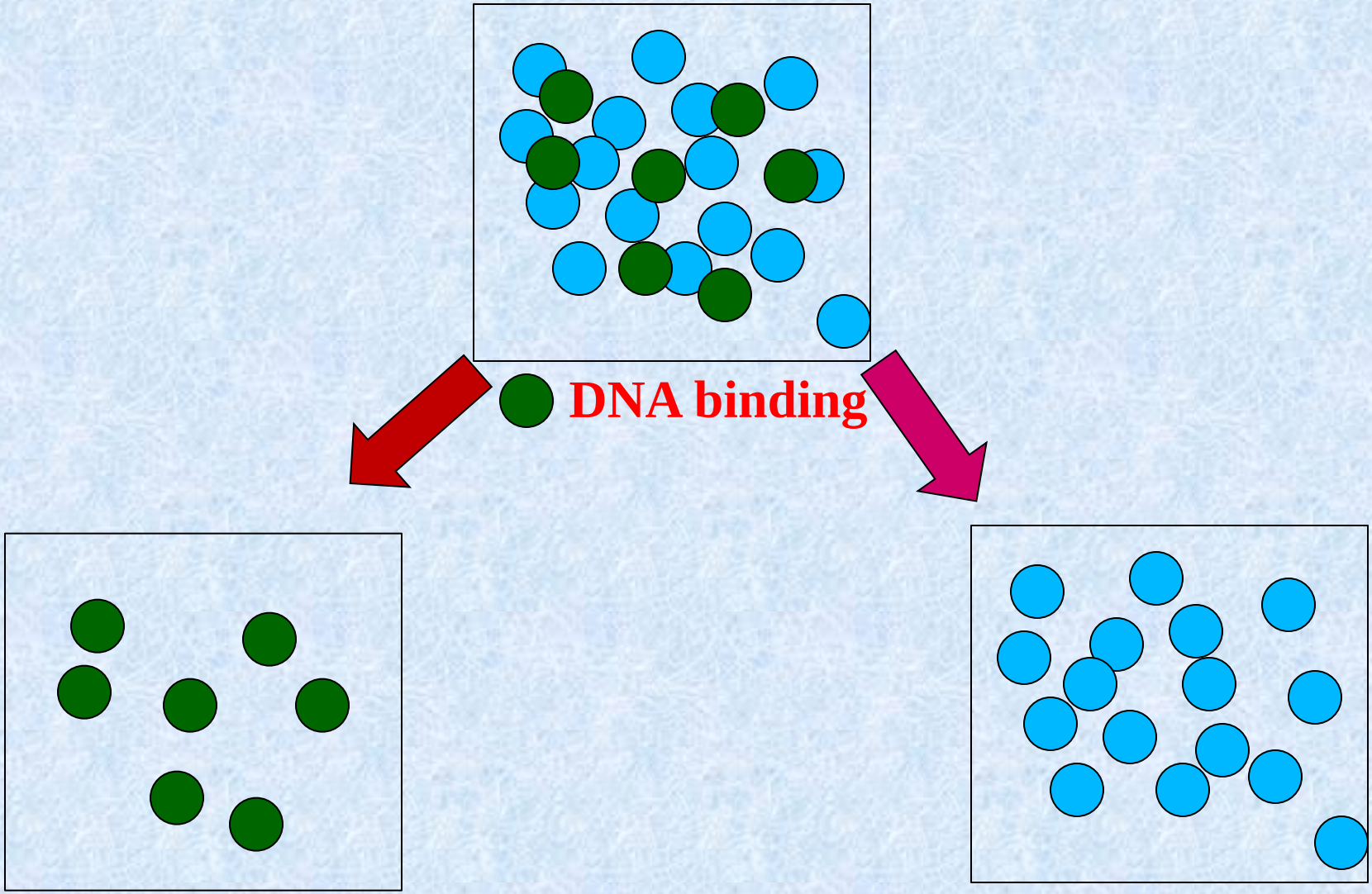
## Discrimination



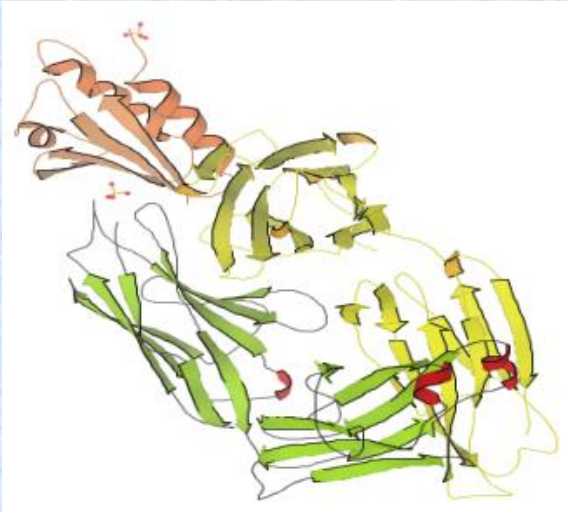
## Prediction

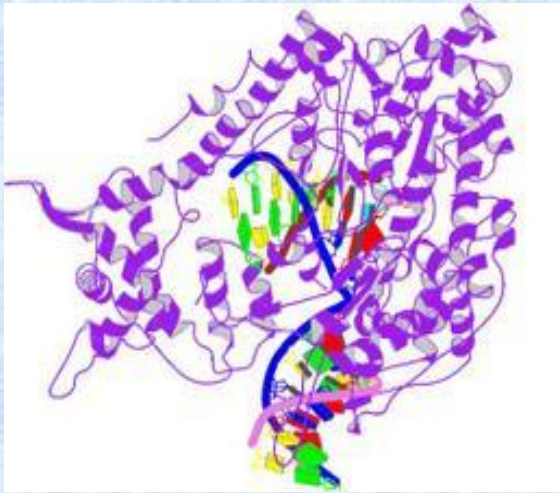


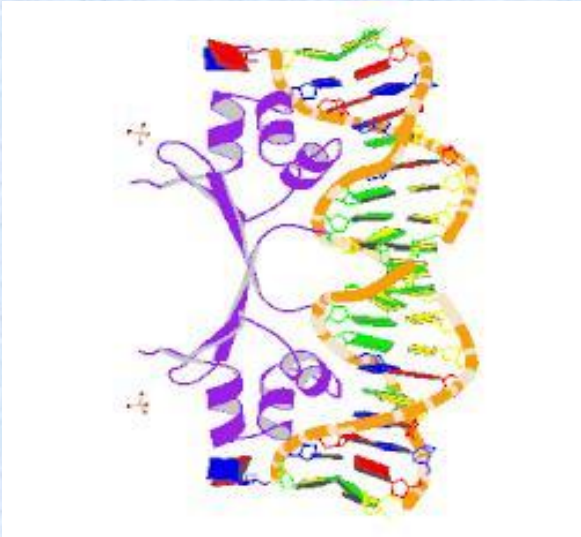
# Discrimination of DNA binding proteins

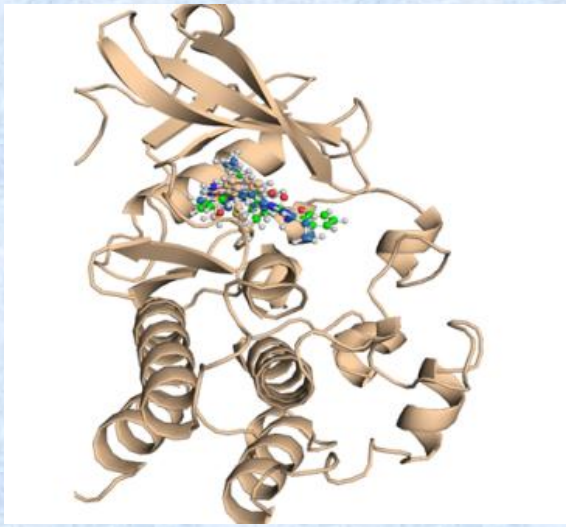


# Binding sites in protein complexes





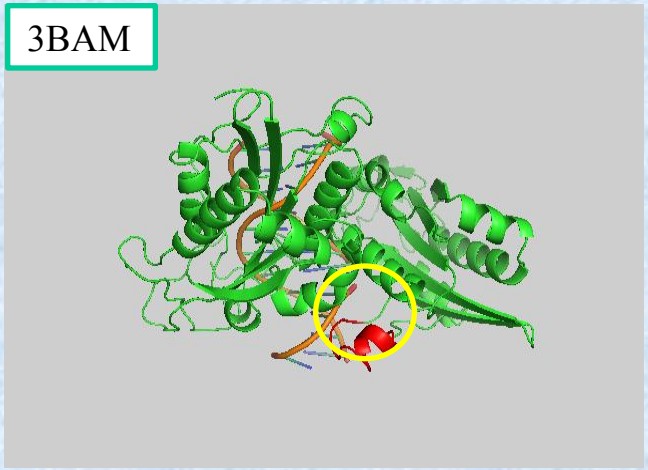




# Disordered regions

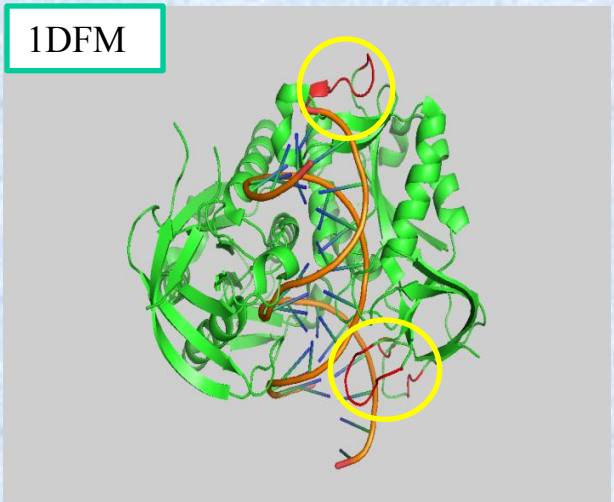
Ex : 1 Free protein : 1BAM:A Complex : 3BAM:A

```
MEVEKEFITDEAKELLSKDKLIQQAYNEVKTSICSPIWPATSKTFTINNTEKNCN
GVPIKELCYTLLEDTYNWYREKPLDILKLEKKKGGPIDVYKEFIENSELKRVGME
FETGNISSAHRSMNKLLLGLKHGEIDLAIILMPIKQLAYYLTDRVTNFEELEPYF
ELTEGQPFIFIGFNAEAYNSNVPLIPKGSDGMSKRSIKKWKDKVENK
```



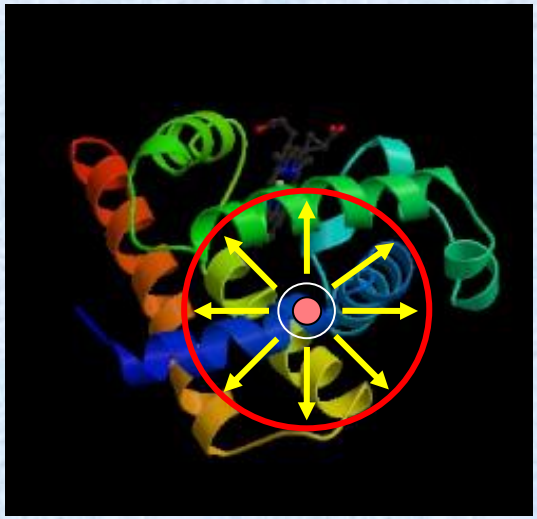

Ex : 2 Free protein : 1ES8:A Complex : 1DFM:B

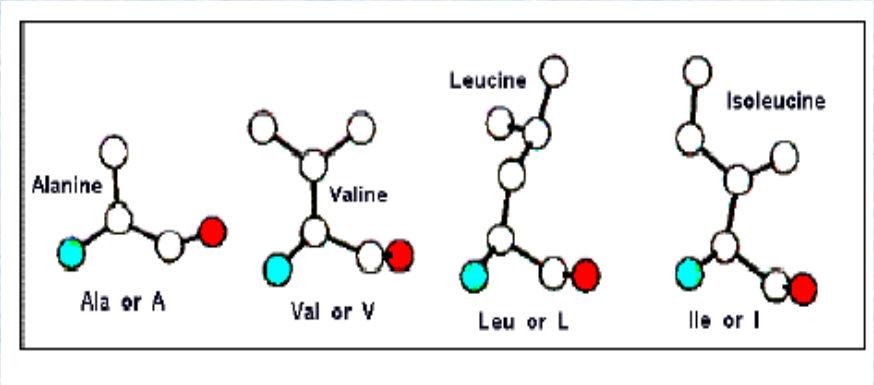
```
MKIDITDYNHADEILNPQLWKEIEETLLKMPLHVKASDQASKVGSLIFDPVGTNQ
YIKDELVPKHWKNNIPIPKRFDFLGTDIDFGKRDTLVEVQFSNYPFLLNNTVRSE
LFHKSNMDIDEEGMKVAIIITKGHMFPASNSSLYYEQAQNQLNSLAEYNVFDVPI
RLVGLIEDFETDIDIVSTTYADKRYSRTITKRDTVKGKVIDTNTPNTRRRRKRGT
IVTY
```

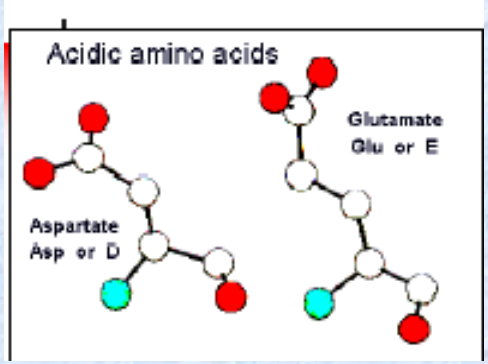


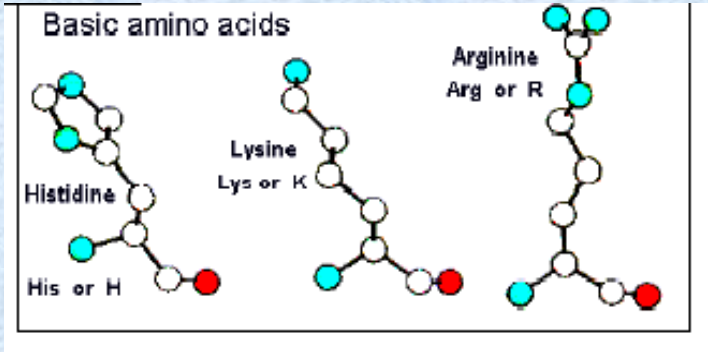

The regions which are transformed from disordered to ordered state during complex formation are indicated in red color which are encircled in the figures

# Protein Mutant stability





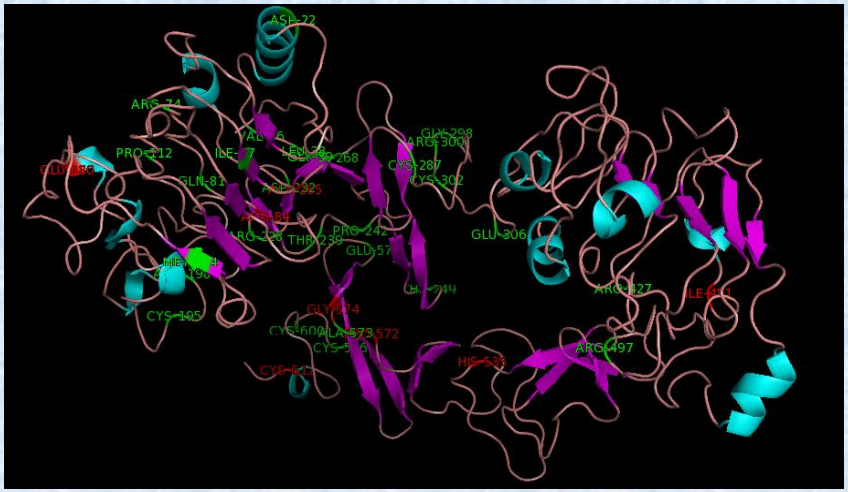



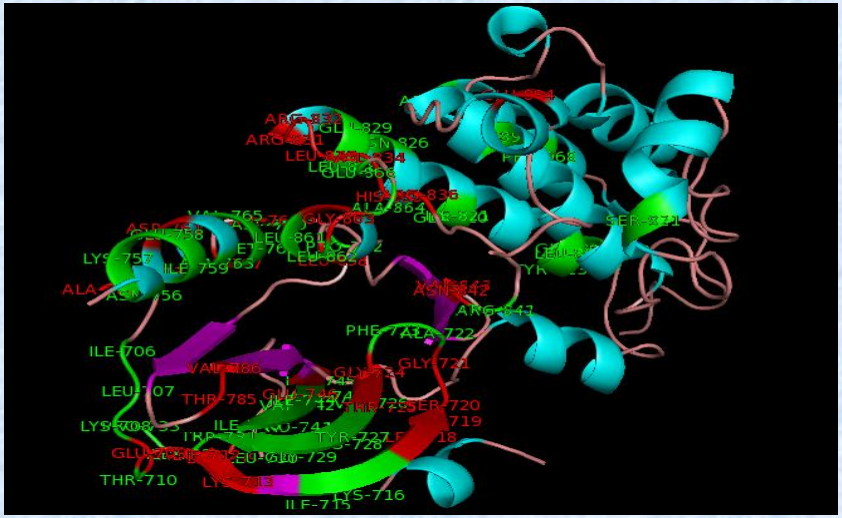

| | |
|---|---|
| Ala | Val: collide OR well packed |
| Val | Ala: cavity OR freely move |
| Ala | Asp: electrostatic |
| Lys | Leu: break ion pairs |

- In a protein, the **replacement** of one amino acid into others may disturb the structure, stability and function.

Hemoglobin Glu 6 Val (sickle cell anemia); p53 mutants: cancer

# Driver and passenger cancer mutations in EGFR



# Research themes

## Real value prediction

- Stability change upon mutation
- Binding affinity of protein complexes
- Protein folding rates
- Solvent accessibility

# Construction of datasets

**Non-redundant dataset**

**Main dataset**

**Test set**

**Blind data**

**Dataset is very sensitive**

# Feature selection

## Evaluate different features

- Physical/chemical Properties
- Specific interactions
- Amino acid composition
- Residue pair preference
- Motifs

**Selecting important features plays a key role**

# Statistical methods

**Regression equations**

**(understanding the influence of specific features)**

**Simple equation:**

**Relationship between two variables**

**E.g. Physiological conditions and growth of a plant**

**Maximum temperature**

**Minimum temperature**

**Maximum relative humidity**

**Minimum relative humidity**

**Rainfall**

**Rainy days per week**

# Statistical methods

Multiple regression technique

$$Y = a + b_1 X_1 + b_2 X_2 + \ldots\ldots\ldots + b_n X_n$$

Y: experimental data (growth of a plant)

$X_n$ = independent variable (physiological conditions).

In this equation, the regression coefficients (or *B* coefficients) represent the *independent* contributions of each independent variable to the prediction of the dependent variable.

**Principle of least squares**

# Assessment

## Correlation coefficient

$$r \quad [N \quad {}_{i\ 1}^{N} X_i Y_i \quad ( \quad {}_{i\ 1}^{N} X_i \quad {}_{i\ 1}^{N} Y_i)] / \sqrt{[N \quad {}_{i\ 1}^{N} X_i^2 \quad ( \quad {}_{i\ 1}^{N} X_i)^2][N \quad {}_{i\ 1}^{N} Y_i^2 \quad ( \quad {}_{i\ 1}^{N} Y_i)^2]},$$

## Mean Absolute Error

$$\mathrm{MAE} \quad \frac{1}{\mathrm{N}} \quad {}_{i\ 1}^{\mathrm{N}} \left| V_i^P \quad V_i^E \right|$$

# Machine learning techniques

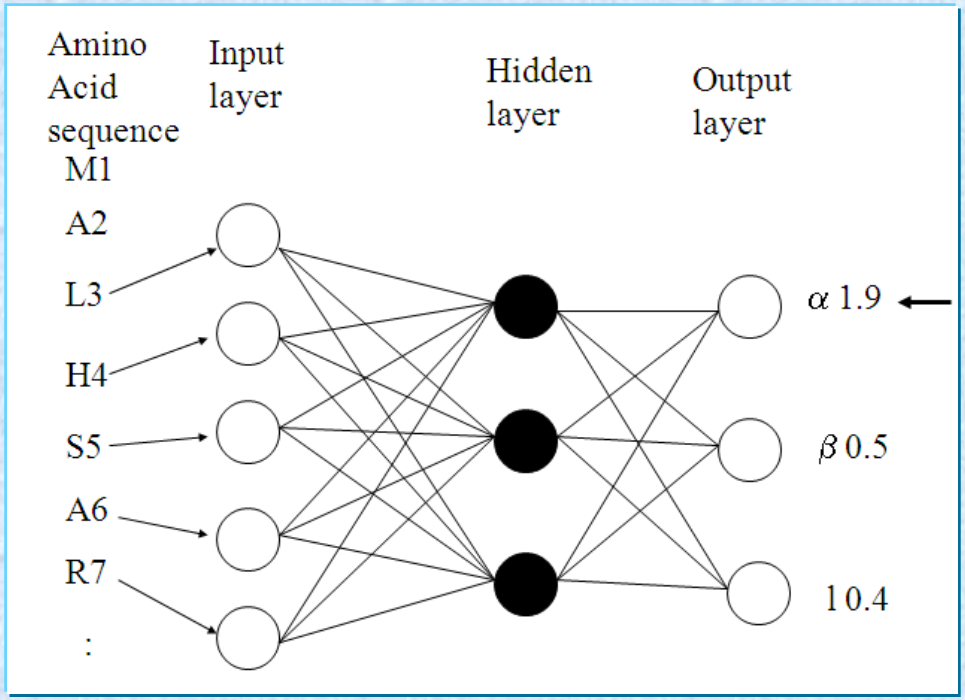
Machine learning techniques are popular in several biological applications.

This technique fits the experimental data with given input parameters and automatically selects the weights for each parameter.

Examples: Neural networks, support vector machines, classification and regression tool etc.

# Neural networks

**Input layer**
**Hidden layer**
**Output layer**




Each node in a layer is connected to all nodes in the preceding layer.

In this case the layers are fully connected.

A weight is associated with each connecting line.

For each node an activation value is calculated based on the node's input value.

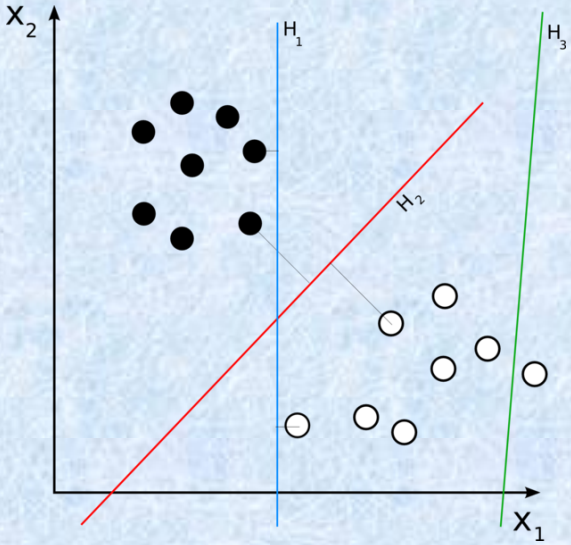
**Training:** Finding the values for the weights , which give high accuracy

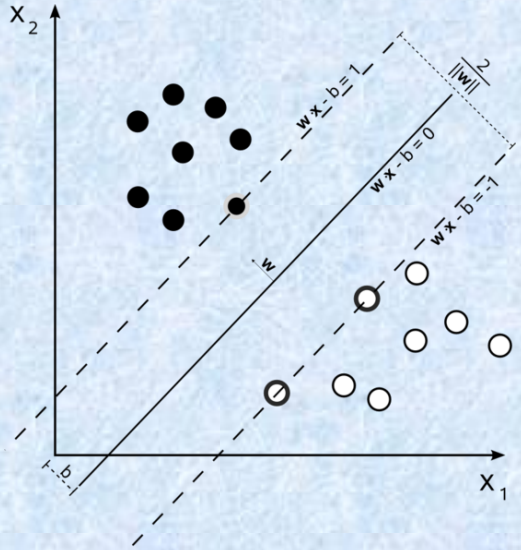
# Support vector machines

It is a learning algorithm, which from a set of positively and negatively labeled training vectors learns a classifier that can be used to classify new unlabelled test samples.

SVM learns the classifier by mapping the input training samples into a possibly high dimensional feature space, and seeking a hyperplane in this space which separates the two types of examples with the largest possible margin, i.e., distance to the nearest points.

If the training set is not linearly separable, SVM finds a hyperplane, which optimizes a trade-off between good classification and large margin.



# Assessment measures

**2-groups classification**

- **Sensitivity = TP/(TP+FN)**
- **Specificity = TN/(TN+FP)**
- **Accuracy = (TP+TN)/(TP+TN+FP+FN)**

| | | Predicted | |
|---|---|---|---|
| | | + | - |
| Experimental | + | TP | FN |
| | - | FP | TN |

**TP: True positive**
**FP: False positive**
**TN: True negative**
**FN: False negative**

```
Experimental:    NNNNNBBBBNNBNNNNBBBNNNNNBNNNBNNNBBN
Predicted:       NNNNNBBBBNNNNBBBBNNNNNNNNNNNNBBBNNN
```

**Assessment with different measures is necessary**

# Discrimination results: β-barrel membrane proteins

**Datasets**: 208 TMβs and 879 non-outer membrane (673 globular and 206 α-helical membrane; nTMβ) proteins

| Method | Parameters | TMβ | nTMβ | Accuracy (%) | |
|---|---|---|---|---|---|
| Statistical | AA | 89 | 79 | 83 | **Bioinformatics, 2005** |
| Statistical | Pairs | 95 | 79 | 86 | **CBAC, 2005** |
| Statistical | Motifs | 96 | 86 | 90 | **BPC, 2005** |
| SVM | 18 AA + 10 pairs | 91 | 95 | 94 | **Bioinformatics, 2005** |
| NN | 49 properties | 81 | 98 | 94 | **BBA, 2006** |
| RBF | PSSM profiles | 89 | 98 | 96 | **CBAC, 2008** |

18 AA: Except Ala and Glu; 10 pairs: QA, DF, DA, KK, EF, NK, DR, YN, FF and LI

Statistical: High sensitivity (correctly identifying TMBs)
Machine learning: High specificity (correctly excluding non-TMBs)

# Machine learning techniques: Cautions

**Over-fitting**

Number of data

Number of features

**Cautious about over-fitting of parameters**

# Validation procedures

1. Self assessment (back-check)
2. Training and test set
3. n-fold cross validation
4. Jack-knife test
5. Split sampling

**Systematic validation procedure shows the performance**

# Comparison with other methods

Different datasets

Measures

Validations

New data

**Systematic comparison shows the superiority**

# Constructing web servers

Beneficial to other researchers

- User friendly
- Effective algorithms
- Reliable results

# Gene and Genome

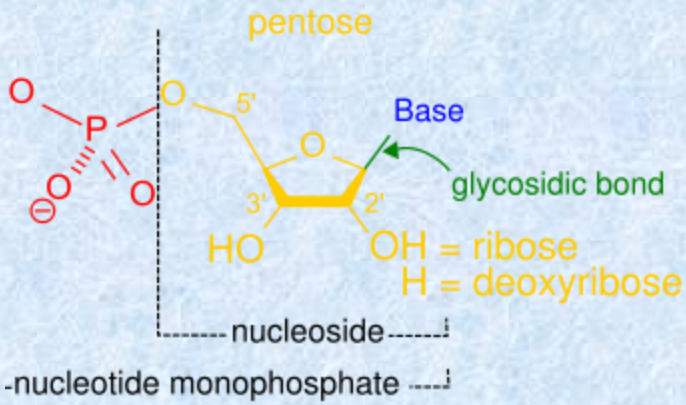
**A gene is a unit of heredity in a living organism. It is normally a stretch of DNA that codes for a type of protein or for an RNA chain that has a function in the organism.**

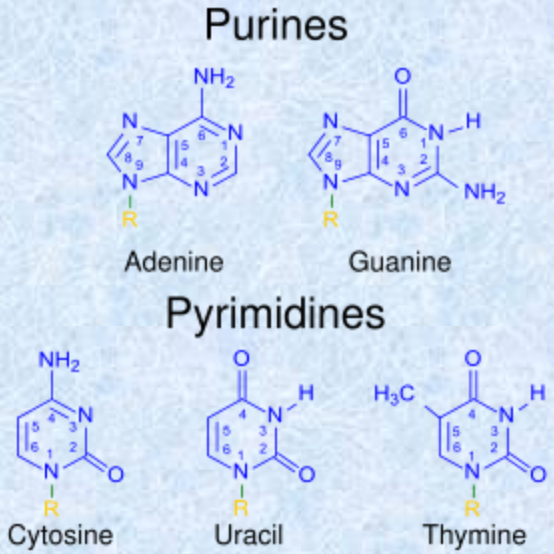
In most living organisms genetic information is stored in the molecule deoxyribonucleic acid, or DNA.

DNA is made and resides in the nucleus of living cells.

DNA gets its name from the sugar molecule contained in its backbone (deoxyribose);

Four different nucleotide bases occur in DNA: **adenine (A), cytosine (C), guanine (G), and thymine (T).**




**Genome is the entirety of an organism's hereditary information.**

# Human genome

**Human genome occupies a total of just over 3 billion DNA base pairs.**

**Human genome contains ca. 23,000 protein-coding genes.**

**A personal genome sequence is a complete sequencing of the chemical base pairs that make up the DNA of a single person.**

**Because medical treatments have different effects on different people because of genetic variations the analysis of personal genomes may lead to personalized medical treatment.**

**The genome of Kim (Korea) had 1.58 million alterations**

**Six out of 10,000 DNA bases are unique to Koreans**

**Homo sapiens chromosome 1 genomic contig, GRCh37**

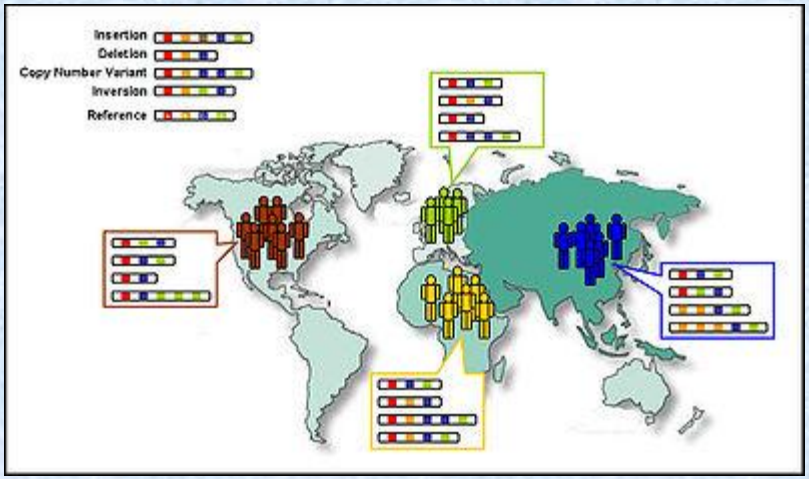
NCBI Reference Sequence: NT_077402.2

```
>gi|224514618|ref|NT_077402.2| Homo sapiens chromosome 1 genomic contig,
reference primary assembly
TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCAACCCTAACCCTAACCCTAACCCTAACCCTAA
CCCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCTAACCCTAACCCTAACCCTAACCCTAACCCT
AACCCTAACCCTAACCCTAACCCCTAACCCTAACCCTAAACCCTAAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAACCCCAACCCCAACCCTAACCCTAACCCTAACCCTAACC
CTACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCCTAACCCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTCGCGGTACCCTCAGCCGGCCCG
CCCGCCCGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAGAGTACCACCGAAATCTGTGCAGAGGAC
AACGCAGCTCCGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGCAACTCCGCCGTTGCAAAGG
CGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCG
GCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAG
AGAGGCGCGCCGCGCCGGCGCAGGCGCAGACACATGCTAGCGCGTCGGGGTGGAGGCGTGGCGCAGGCGC
AGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGACACATGCTACCGCGTCCAGGGGTGGAGGCGTGGCGC
AGGCGCAGAGAGGCGCACCGCGCCGGCGCAGGCGCAGAGACACATGCTAGCGCGTCCAGGGGTGGAGGCG
TGGCGCAGGCGCAGAGACGCAAGCCTACGGGCGGGGGTTGGGGGGGCGTGTGTTGCAGGAGCAAAGTCGC
ACGGCGCCGGGCTGGGGCGGGGGGAGGGTGGCGCCGTGCACGCGCAGAAACTCACGTCACGGTGGCGCGG
CGCAGAGACGGGTAGAACCTCAGTAATCCGAAAAGCCGGGATCGACCGCCCCTTGCTTGCAGCCGGGCAC
TACAGGACCCGCTTGCTCACGGTGCTGTGCCAGGGCGCCCCCTGCTGGCGACTAGGGCAACTGCAGGGCT
CTCTTGCTTAGAGTGGTGGCCAGCGCCCCCTGCTGGCGCCGGGGCACTGCAGGGCCCTCTTGCTTACTGT
ATAGTGGTGGCACGCCGCCTGCTGGCAGCTAGGGACATTGCAGGGTCCTCTTGCTCAAGGTGTAGTGGCA
GCACGCCCACCTGCTGGCAGCTGGGGACACTGCCGGGCCCTCTTGCTCCAACAGTACTGGCGGATTATAG
GGAAACACCCGGAGCATATGCTGTTTGGTCTCAGTAGACTCCTAAATATGGGATTCCTGGGTTTAAAAGT
AAAAAATAAATATGTTTAATTTGTGAACTGATTACCATCAGAATTGTACTGTTCTGTATCCCACCAGCAA
TGTCTAGGAATGCCTGTTTCTCCACAAAGTGTTTACTTTTGGATTTTTGCCAGTCTAACAGGTGAAGCCC
TGGAGATTCTTATTAGTGATTTGGGCTGGGGCCTGGCCATGTGTATTTTTTTAAATTTCCACTGATGATT
TTGCTGCATGGCCGGTGTTGAGAATGACTGCGCAAATTTGCCGGATTTCCTTTGCTGTTCCTGCATGTAG
TTTAAACGAGATTGCCAGCACCGGGTATCATTCACCATTTTTCTTTTCGTTAACTTGCCGTCAGCCTTTT
CTTTGACCTCTTCTTTCTGTTCATGTGTATTTGCTGTCTCTTAGCCCAGACTTCCCGTGTCCTTTCCACC
GGGCCTTTGAGAGGTCACAGGGTCTTGATGCTGTGGTCTTCATCTGCAGGTGTCTGACTTCCAGCAACTG
CTGGCCTGTGCCAGGGTGCAAGCTGAGCACTGGAGTGGAGTTTTCCTGTGGAGAGGAGCCATGCCTAGAG
TGGGATGGGCCATTGTTCATCTTCTGGCCCCTGTTGTCTGCATGTAACTTAATACCACAACCAGGCATAG
GGGAAAGATTGGAGGAAAGATGAGTGAGAGCATCAACTTCTCTCACAACCTAGGCCAGTAAGTAGTGCTT
GTGCTCATCTCCTTGGCTGTGATACGTGGCCGGCCCTCGCTCCAGCAGCTGGACCCCTACCTGCCGTCTG
CTGCCATCGGAGCCCAAAGCCGGGCTGTGACTGCTCAGACCAGCCGGCTGGAGGGAGGGGCTCAGCAGGT
CTGGCTTTGGCCCTGGGAGAGCAGGTGGAAGATCAGGCAGGCCATCGCTGCCACAGAACCCAGTGGATTG
GCCTAGGTGGGATCTCTGAGCTCAACAAGCCCTCTCTGGGTGGTAGGTGCAGAGACGGGAGGGGCAGAGC
CGCAGGCACAGCCAAGAGGGCTGAAGAAATGGTAGAACGGAGCAGCTGGTGATGTGTGGGCCCACCGGCC
CCAGGCTCCTGTCTCCCCCCAGGTGTGTGGTGATGCCAGGCATGCCCTTCCCCAGCATCAGGTCTCCAGA
GCTGCAGAAGACGACGGCCGACTTGGATCACACTCTTGTGAGTGTCCCCAGTGTTGCAGAGGTGAGAGGA
GAGTAGACAGTGAGTGGGAGTGGCGTCGCCCCTAGGGCTCTACGGGGCCGGCGTCTCCTGTCTCCTGGAG
AGGCTTCGATGCCCCTCCACACCCTCTTGATCTTCCCTGTGATGTCATCTGGAGCCCTGCTGCTTGCGGT
GGCCTATAAAGCCTCCTAGTCTGGCTCCAAGGCCTGGCAGAGTCTTTCCCAGGGAAAGCTACAAGCAGCA
AACAGTCTGCATGGGTCATCCCCTTCACTCCCAGCTCAGAGCCCAGGCCAGGGGCCCCCAAGAAAGGCTC
TGGTGGAGAACCTGTGCATGAAGGCTGTCAACCAGTCCATAGGCAAGCCTGGCTGCCTCCAGCTGGGTCG
ACAGACAGGGGCTGGAGAAGGGGAGAAGAGGAAAGTGAGGTTGCCTGCCCTGTCTCCTACCTGAGGCTGA
GGAAGGAGAAGGGGATGCACTGTTGGGGAGGCAGCTGTAACTCAAAGCCTTAGCCTCTGTTCCCACGAAG
GCAGGGCCATCAGGCACCAAAGGGATTCTGCCAGCATAGTGCTCCTGGACCAGTGATACACCCGGCACCC
TGTCCTGGACACGCTGTTGGCCTGGATCTGAGCCCTGGTGGAGGTCAAAGCCACCTTTGGTTCTGCCATT
GCTGCTGTGTGGAAGTTCACTCCTGCCTTTTCCTTTCCCTAGAGCCTCCACCACCCCGAGATCACATTTC
TCACTGCCTTTTGTCTGCCCAGTTTCACCAGAAGTAGGCCTCTTCCTGACAGGCAGCTGCACCACTGCCT
```

# 1000 genomes project

**The 1000 Genomes Project**, launched in January 2008, is an international research effort to establish the most detailed catalogue of human genetic variation.

Sequence the genomes of at least **one thousand** anonymous participants from a number of different ethnic groups within the next three years, using newly developed technologies which are faster and less expensive.




## An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aim
build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1,
individuals from 14 populations, constructed using a combination of low-coverage whole-genome and ex
sequencing. By developing methods to integrate information across several algorithms and diverse data sources,
provide a validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions
deletions, and more than 14,000 larger deletions. We show that individuals from different populations carry diffe
profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiat
which is further increased by the action of purifying selection. We show that evolutionary conservation and co
consequence are key determinants of the strength of purifying selection, that rare-variant load varies substanti
across biological pathways, and that each individual contains hundreds of rare non-coding variants at conserved si
such as motif-disrupting changes in transcription-factor-binding sites. This resource, which captures up to 98%
accessible single nucleotide polymorphisms at a frequency of 1% in related populations, enables analysis of common
low-frequency variants in individuals from diverse, including admixed, populations.

**Nature, 2010**

# Cancer genome project

**The Cancer Genome Project aims to identify sequence variants/mutations critical in the development of human cancers.**

**Cancer Genome Project represents an effort to improve cancer diagnosis, treatment, and prevention through a better understanding of the molecular basis of this disease.**

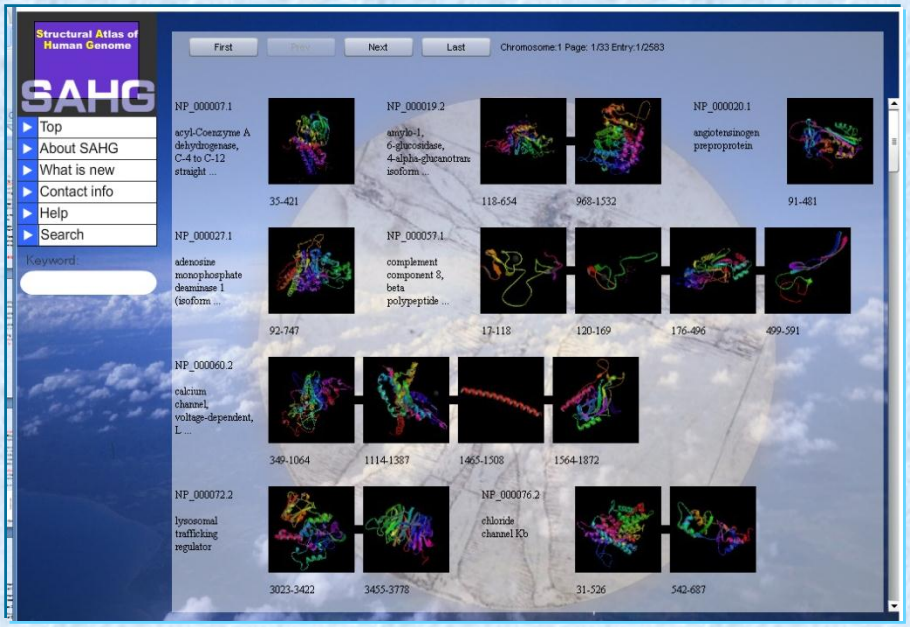
**The Cancer Genome Project combines knowledge of the human genome sequence with high throughput mutation detection techniques.**

# Human genome: Bioinformatics

## Structural Atlas of Human Genome (Japan)

**Structural and Functional Annotation**

- **Protein-protein complexes**
- **Protein-DNA complexes**
- **Protein-RNA complexes**
- **Protein-small molecule complexes**
- **Membrane proteins**
- **Subcellular localization**
- **Transporters**
- **Receptors**
- **Enzymes**
- **Coiled coil proteins**



# Open problems

**Annotate genomes based on structure and function of proteins**

**Cross-genome analysis**

**Personalized medicine**