**Semester - III**

**17BTP312 BIOINFORMATICS -PRACTICAL VI 4H-2C**

Total hours/week:L:0 T:0 **P:4** **Marks:** Internal:**40** External:**60** Total: **100**

**Bioinformatics**

1. Using RasMol through command line.
2. Quaternary structural analysis.
3. Investigation of molecular interactions using the program KineMage.
4. Similarity search using the Blast and interpretation of the results.
5. Pair-wise and multiple sequence sequence alignment by using ClustalW.
6. Introduction of BioEdit.
7. Phylogenetic analysis using web tools.
8. Protein Structure Prediction (Homology Modeling) using SPDBV.
9. Molecular modeling using SPARTAN.
10. ModelBuilding and Energy minimization.
11. Quantum chemical and molecular mechanics practicals.
12. Basic UNIX commands, pine, telnet, ftp.
13. Molecular dynamics simulation using GROMACS etc.
14. Molecular Docking and Drug designing by using Chimera.

**References**


1. Bunin Barry, A., Siesel Brian,Morales Guillermo,&Bajorath Jurgen.(2006). *Chemoinformatics.*New York:Theory, Practice, & Products Publisher&Springer.ISBN: 1402050003.
2. Gasteiger Johann, &Engel Thomas.(2003).*Chemoinformatics: A Textbook.*WileyVCH. ISBN: 3527306811.
3. Leach Andrew R., Valerie J. Gillet.(2003).*An introduction to chemoinformatics*. Kluwer academic. ISBN: 1402013477.
4. Gasteiger Johann, (2003). *Handbook of Chemoinformatics: From Data to Knowledge(Vols 4).*Wiley-VCH. ISBN:3527306803.

# Lab Manual on

# Bioinformatics

**Dr. Prabu, G.R**
Associate Professor
Department of Biotechnology
Faculty of Arts, Science and Humanities

Enable | Enlighten | Enrich
KARPAGAM
ACADEMY OF HIGHER EDUCATION
(Deemed to be University)
(Established Under Section 3 of UGC Act, 1956 )

KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University)
(Established Under Section 3 of UGC Act, 1956)
Pollachi Main Road, Eachanari Post, Coimbatore - 641 021, Tamilnadu, India.
Phone : 0422 - 2980011 - 14, 6471113, 14 | Fax : 0422 - 2980022-23 | Email : info@karpagam.com

# Contents

**Ex. No.** **Basic UNIX commands**

**Aim:** To verify the basic unix Commands and Filters

1. DATE command

Display the server date and time

Syntax :$date

Output: Mon June 12, 2010

2. CALENDER command

Display the particular month calendar or year calendar

Syntax :$cal <*month name*> or <*year*>

Eg:$ cal 2010 – Prints the calendar for the entire year 2010.

$cal 11 2009-Display Nov-2009 calendar

3. ECHO command

It is used to print the message on the screen,whatever you happen to type on the line.

Syntax :$echo<*text to be displayed on the screen*>

Eg:$echo Welcome to Bioinformatics Laboaratory 2010

4. BANNER command

It is used to print the message in large letters to gibe the impression of a banner.

Syntax :$banner<*text*>

Eg:$banner BIOTECH

5.WHO command

Display the information about all users who have logged into the system currently.

Syntax :$who

6. WHO AM I command

It gives login details of a particular system i.e it gives the user name,terminal name,date and time of login

Syntax :$who am i

7. EXIT command

It is used to logout from the user sessions.

Syntax :$exit

8. CLEAR command

It is used to clear the screen.

Syntax :$clear

9. LOGNAME commnd

It is used to display the current user name.

Syntax :$logname

10. ID command

The id command is used to display the numerical value that corresponds to your login name i.e.,every valid UNIX user is assigned a login name,a user id and a group-id.

Syntax :$id

11. TTY command

The tty(teletype) command is used to know the terminal name that we are using.

Syntax :$tty

12. UNAME command

Display the name of the operating system.

Syntax :$uname

**Ex. No.** **File and Directory Commands**

**Aim** : To create the file(s) and verify the file handling commands.

1. TOUCH command

Create file(s) with zero byte size.

Syntax :$touch <*filename*>

Eg:$touch biotech

$touch bioinfo gene dna.

2. CAT command

Create file with data and display data in the file.

(a) File creation

Syntax :$cat ><*filename*>

Eg:$cat>test

……
……
……
Ctrl+d[to close file]

Eg:$cat>>test--->append data to the file 'test'

……
……
Ctrl+d

Eg:$cat file1 file2>file3 ---->data in file1,file2 copied to file3.

(b) Display data from the file(s)

Syntax :$cat <*filename*>

Eg:$cat file1---->display data from file1

Eg:$cat file1 file2 file3

3. CP command

It is used to copy the contents of one file to another file and copies the file from one place to another.

Syntax :$cp <*source filename*> <*destination filename*>

Eg:$cp file1 file2

4. MV command

It is used to rename the file(s).

Syntax :$mv <*old filename*> <*new filename*>

Eg:$mv file1 file4--->file1 renamed to file4

5. RM command

It is used to remove the file(s)

Syntax :$rm <*filename(s)*>

Eg:$rm file1--->file1 is removed

Eg:$rm file2 file3 file4--->remove three files

Eg:$rm*--->remove all files in current directory

6. FILE command

It is used to determine the type of the file.

Syntax :$file <*file name(s)*>

Eg:$file file1

Eg:$file*

7. WC command

This command is used to display the number of lines, number of words and number of characters in a file

Syntax :$wc <*file name(s)*>

Eg:$wc file1

Eg:$wc –l file1--->display only the number of lines in file1.

Eg:$wc –w file1--->display only the number of words in file1.

Eg:$wc –c file1--->display only the number of characters in file1.

**Aim : To create the directorie(s) and verify the directory commands**

1. PWD command

It is used to know the current working directory

Syntax :$pwd

2. MKDIR command

It is used to create an empty directory

Syntax :$mkdir <*directory name*>

Eg:$mkdir program

3. CD command

It is used to move from one directory to another directory

Syntax :$cd <*directory name*>

Eg:$cd biotech

4. RMDIR command

It is used to remove the directory only if it is empty

Syntax :$rmdir <*directory name*>

Eg:$rmdir program

Syntax :$rm –r <*directory name*>

Eg:$rm –r biotech

5. MV command

It is used to rename the directory

Syntax:$mv <*old directoryname*> <*new directoryname*>

Eg:$mv program biotech--->directory'program' renamed to 'biotech'

6. LS command

It is used to view the contents of the directory

Syntax :$ls

**Ex. No.** **Queries based on Biological databases**

**Introduction:**

Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analyses.

They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics.

Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures.

Biological databases are an important tool in assisting scientists to understand and explain a host of biological phenomena from the structure of biomolecules and their interaction, to the whole metabolism of organisms and to understanding the evolution of species.

This knowledge helps facilitate the fight against diseases, assists in the development of medications and in discovering basic relationships amongst species in the history of life.

Biological knowledge is distributed amongst many different general and specialized databases. This sometimes makes it difficult to ensure the consistency of information.

Biological databases cross-reference other databases with accession numbers as one way of linking their related knowledge together. An important resource for finding biological databases is a special yearly issue of the journal Nucleic Acids Research (NAR). The Database Issue of NAR is freely available, and categorizes many of the publicly available online databases related to biology and bioinformatics.

**1. Retrieve the gene sequence in FASTA format corresponding to P00519.**

**Aim:** To retrieve the gene sequence in FASTA format corresponding to P00519

**Introduction:** A gene is a molecular unit of heredity of a living organism. It is a name given to some stretches of DNA and RNA that code for a type of protein or for an RNA chain that has a function in the organism. Knowledge of gene sequences has become indispensable for basic biological research, other research branches utilizing sequencing, and in numerous applied fields such as diagnostic, biotechnology, forensic biology and biological systematics.

In bioinformatics, **FASTA format** is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes. The format also allows for sequence names and comments to precede the

sequences. The format originates from the FASTA software package, but has now become a standard in the field of bioinformatics.

The simplicity of FASTA format makes it easy to manipulate and parse sequences using textprocessing tools and scripting languages

**Method:**

1. Open Uniprot Database www.uniprot.org

2. Enter the protein Id P00519 in search tab and click on Find

3. Click on the protein name displayed on the result page. ABL1_HUMAN

4. Obtain relevant information about protein and retrieve FASTA format of its sequence by clicking on the FASTA tab at the right corner.

**Result and inference :**

**Ex. No.** **BioEdit**

**Introduction:**

*BioEdit* is a biological sequence alignment editor written for Windows 95/98/NT/2000/XP. An intuitive multiple document interface with convenient features makes alignment and manipulation of sequences relatively easy on your desktop computer.

Several sequence manipulation and analysis options and links to external analysis programs facilitate a working environment which allows you to view and manipulate sequences with simple point-and-click operations.

**Version 5.0.6 offers the following features**:

• An easy, graphical interface for sequence manipulation and editing.

• Variable editing options, including ‘select and drag’ sliding and 'grab and drag' sliding of residues, variable selection options, mouse-click insert and delete of gaps, full column selecting, on-screen editing with cut, copy and paste, and auto-scrolling of edit window.

• Anchor alignment columns to protect fixed regions in an alignment.

• Automatically and manually annotate sequences with features such as introns, exons, promoters, CDS, and all standard GenBank feature types. Automatically annotate other sequences in an alignment using one sequence as a template.

• Group sequences into color-coded families and lock group members for synchronized hand alignment.

• User-defined character-relevance (any characters can be set to be considered as relevant bases in nucleic acid or amino acid sequences for the purposes of similarity shading, sequence identity matrices, and conservation plot views.

• User-defined motif searching using standard Prosite nomenclature and utilizing IUPAC characters to allow searching in nucleic acid or amino acid sequences, as well as exact text searches including or ignoring gaps.

• Lines may be defined as DNA, RNA, nucleic acid, protein, undefined, or Comments. Comments may be used to hold general notes or things such as secondary structure mask definitions, but do not contribute to conservation calculations.

• Rudimentary phylogenetic tree viewer which supports node flipping and printing.

• Link phylogenetic trees to alignments and save trees with BioEdit format alignment files.

• Append one alignment to the end of another.

• Configure accessory application interfaces to run external analysis programs through a graphical interface created by BioEdit. Automatically feed information to and retrieve files from external apps. External apps run in a separate thread to allow simultaneous use of BioEdit while running time-consuming processes. Output from an external program may be automatically opened by another program.

• Display, print and edit ABI trace files from ABI autosequencer model 377, 373, and 3700, as well as SCF files of version 2 and 3, such as the files output by Licor sequencers.

• RNA comparative analysis tools, including covariation, potential pairings, and mutual information analyses.

• 2-D matrix plotter for mutual information output with dynamic data viewing with the mouse pointer. (Also allows image copy/paste and bitmap save).

• Interactive 1-D plots of mutual information matrix rows and columns

• Save sequence annotation information in BioEdit or GenBank format

• Align protein-encoding nucleic acid sequences through amino acid translation.

• Search for conserved regions in an alignment (find good PCR targets or help define motifs)

• Search for user-defined motifs in nucleic acid or protein sequences or search exact text with wildcards and choice of including or ignoring gaps.

• Dynamic memory allocation with support for up to 20,000 sequences per document..

• Read and write GenBank, Fasta, Phylip and NBRF/PIR files internally and import/export several other formats with Don Gilbert's ReadSeq.

• Read and write large alignment files quickly with the BioEdit Project file format.

• ClustalW multiple sequence alignment (interface internal, external program by Des Higgins et. al.) with auto-update of aligned protein full titles and GenBank field information, as well as nucleotide coding sequence when aligned from a protein view of nucleotide sequences.

• Block copying of residues to clipboard allowing for pasting of full alignments or parts of alignments into a word processor.

• Basic sequence manipulations (copy/paste of sequences between documents, translation and degenerate encoding, RNA->DNA->RNA, reverse/complement, upper/lowercase).

• Multiple document interface (Maximum of 20 open alignment documents at a time, but no set limit on other open windows).

• Six-Frame translation of nucleic acid sequences into Fasta-format ORF lists. Tested by translating the E. coli genome (4.6 Mbases) into 10,125 sorted raw codon stretches of 100 or more amino acids and 39,880 unsorted raw codon stretches of 50 or more amino acids.

• Semi-automated plasmid/vector drawing and annotation with vectored graphics, automatic restriction site and positional marking, automated polylinker view, and user-controlled drawing objects

• Save plasmid files as editable vectored graphic files or as bitmaps, copy to other graphics applications, and print plasmids at printer's full resolution.

• Amino acid and nucleotide composition summaries and plots

• 'Revert to Saved' and 'undo' functions.

• Edit both amino acid and nucleic acid sequences.

• Easy point-and-click color table editing, with different tables for protein and nucleic acid sequences.

• Alignment-responsive shading based on information content of alignment positions.

• BioEdit currently reads and writes GenBank, Fasta, NBRF/PIR, Phylip 3.2 and Phylip 4 formats and reads ClustalW and GCG formats.

• Import/Export filter for 10 additional formats (Using Don Gilbert's ReadSeq).

• Import/Append one file on to the end of another (regardless of file format).

• Basic rich-text editor.

• Internal restriction mapping utility with any or all-frames translation, multiple enzyme and output options, including enzyme suppliers, and circular DNA option.

• Browse restriction enzymes by manufacturer

• Auto-linking to your favorite Web Browser (e.g., Netscape or Internet Explorer).

• World Wide Web Bookmarks. • NCBI BLAST tools, including BLAST 3.0 Internet client and local BLAST with the ability to compile local databases from Fasta files

• Configurable formatted text print with dynamic print preview,

• Configurable formatted shaded graphical output with dynamic preview, identity and similarity shading, and ability to cut and paste directly to graphics/presentation program for generation of figures.

• Entropy (lack of information) plotting of alignments

• Hydrophobicity profiles of multiple proteins using several hydrophobicity scales, with variable window width and option to analyze degapped sequences or alignments.

**Ex. No.** **Sequence similarity searching using BLAST**

**Aim:** To determine the similarity between the sequences using web tools (BLAST).

**Introduction :**

Basic local alignment search tool (BLAST) is a sequence similarity search program.

The National Center for Biotechnology Information (NCBI) maintains a BLAST server with a home page at http://www.ncbi.nlm.nih.gov/BLAST/.

Basic local alignment search tool (BLAST) is a sequence similarity search program that can be used via a web interface or as a stand-alone tool to compare a user's query to a database of sequences.

BLAST is a heuristic that finds short matches between two sequences and attempts to start alignments from these 'hot spots'. In addition to performing alignments, BLAST provides statistical information about an alignment; this is the 'expect' value, or false-positive rate.

The National Center for Biotechnology Information (NCBI) maintains a BLAST server with a homepage at http://www.ncbi.nlm.nih.gov/BLAST/. On the homepage the different BLAST searches are listed by type: nucleotide, protein, translated and genomes.

**1. Comment on the conserved domain present in Q8NFM4.**

**Aim: To determine the conserved domain present in Q8NFM4**

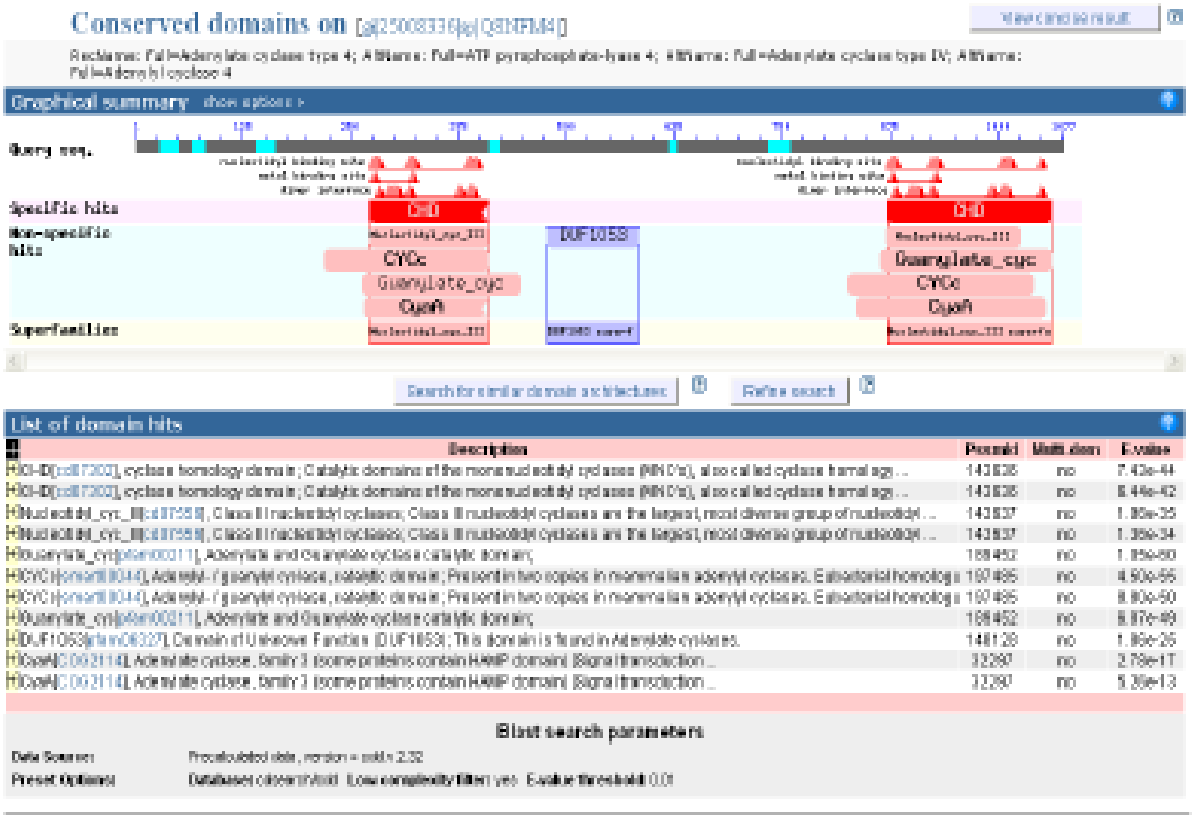
**Introduction:**

Conserved domains (CD) in proteins play a crucial role in protein interactions, DNA binding, enzyme activity, and other important cellular processes. With recently released gene number predictions in the human genome being less than many previous predictions, interactions among these domains may prove to be central to proteome complexity. Protein domains are often conserved across many species, and as such, they offer an interesting dataset in how genomes maintain them with relationship to other conserved domains, as well as to proteome size.

**Method:**

1. Retrieve the sequence from NCBI.
2. Paste the sequence in the Query box in blastp.
3. Run against a non-redundant database (nr).

**Result and inference:**




**Query ID** - gi|25008336|sp|Q8NFM4.1|ADCY4_HUMAN

**Description -** adenylate cyclase type 4 [Homo sapiens]

**Specific hit** - cd07302, cyclase homology domain;

Catalytic domains of the mononucleotidyl cyclases (MNC's), also called cyclase homology domains (CHDs), are part of the class III nucleotidyl cyclases. This class includes eukaryotic and prokaryotic adenylate cyclases (AC's) and guanylate cyclases (GC's). They seem to share a common catalytic mechanism in their requirement for two magnesium ions to bind the polyphosphate moiety of the nucleotide.

**Blast Results:**

Max score = 2214
Total score = 2214
Query coverage = 100%
E value = 0.0

**2. Find the gene sequences of Mouse origin similar to U80226.1.**

**Aim:** To find the gene sequences of Mouse origin similar to U80226.1.

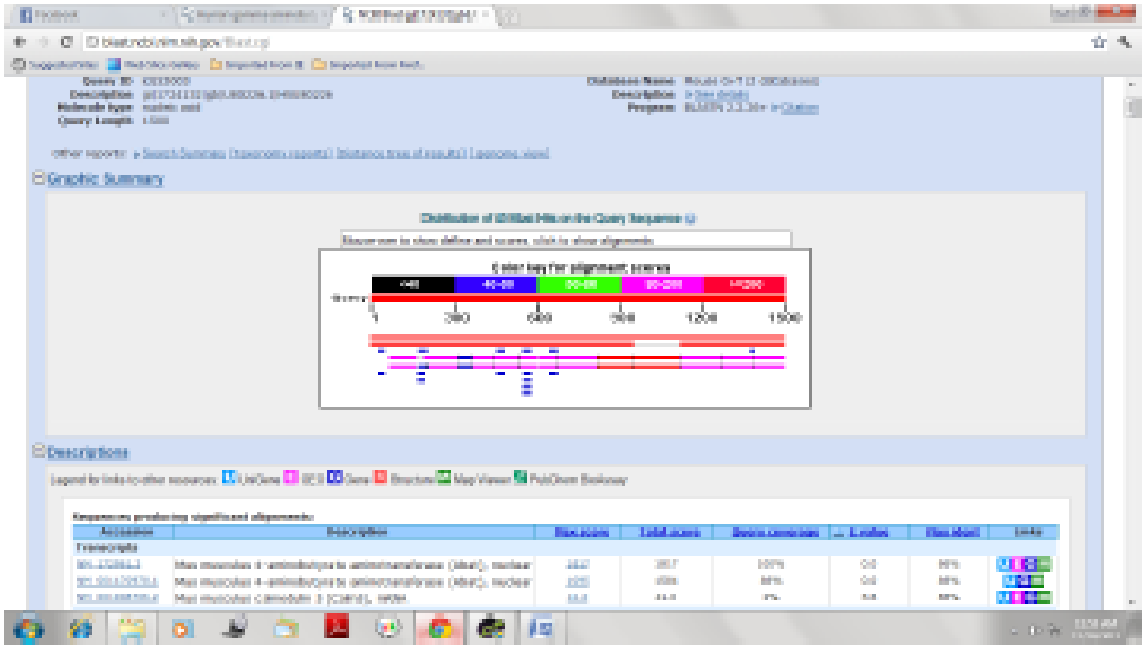
**Introduction:**

**Sequence Similarity Searching** is a method of searching sequence databases by using alignment to a query sequence. By statistically assessing how well database and query sequences match one can infer homology and transfer information to the query sequence.

**Method:**

1. Retrieve the Sequence of **U80226.1**
2. Enter the sequence in FASTA format in blastn
3. Choose Mouse genome+transcript as the Database.
4. Run blastn

**Results and inference:**




**Similar Sequence:**

NM_172961.3

Mus musculus 4-aminobutyrate aminotransferase (Abat), nuclear gene encoding mitochondrial protein, transcript variant 1,

mRNA Length=4653

GENE ID: 268860 Abat | 4-aminobutyrate aminotransferase [Mus musculus]

Score = 1817 bits (2014),
Expect = 0.0
Identities = 1305/1501 (87%), Gaps = 2/1501 (0%)
Strand=Plus/Plus

**3. Write the function of C7AE31. Find its orthologous proteins.**

**Aim**: To determine the function of **C7AE31** and to find its orthologous proteins.

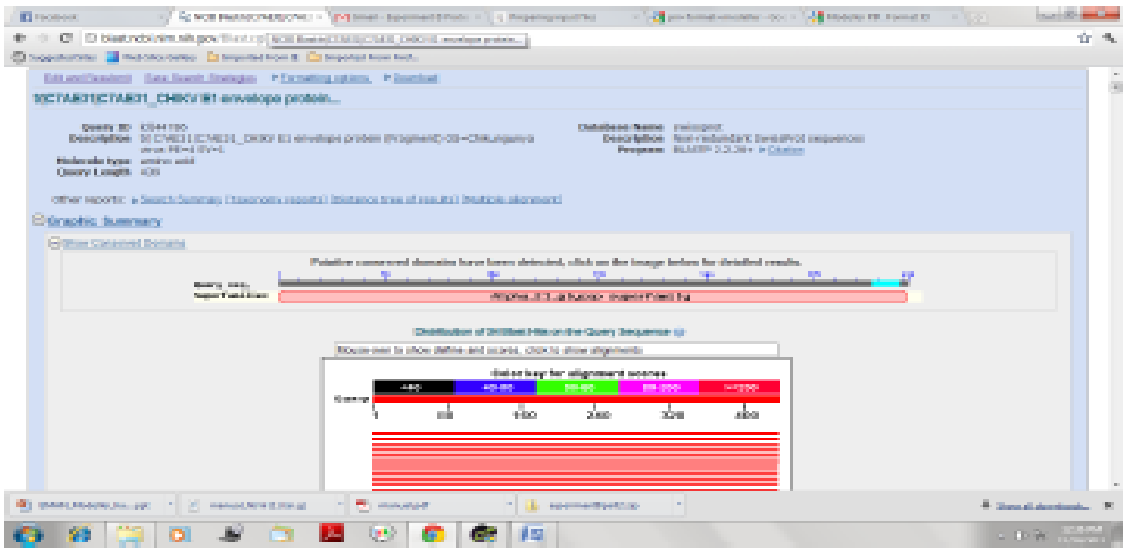
**Introduction**:

Orthologous proteins with the same function in different species, Orthologous proteins with modified function in different species, Orthologous proteins with major modification of function, Orthologous proteins that have lost their function, Orthologous proteins that have gained additional functions, The three-dimensional structure of orthologous proteins, Prediction of secondary structure of proteins, Prediction of the three-dimensional structure of proteins, Detecting sequence homology of protein-coding genes.

**Method:**

1. Retrieve the Sequence of **C7AE31** from Uniprot.
2. Enter the sequence in FASTA format in blastp
3. Run the query against SWISS-PROT database.

**Result and inference:**



**C7AE31**

O90371 POLS_ONNVI

Structural polyprotein (O'nyong-nyong virus (strain Igbo Ora))

O90369 POLS_ONNVS

Structural polyprotein (O'nyong-nyong virus (strain SG650))

P22056 POLS_ONNVG

Structural polyprotein (O'nyong-nyong virus (strain Gulu))

**4. Write the function of P80404. Find its paralogous proteins.**

**Aim**: To determine the function of **P80404** and its paralogous proteins

**Introduction:**

Paralogous means genes that have arisen from a common ancestor and are present in the same genome. Paralgous may or may not have the same function. Paralogous proteins are proteins that have arisen by gene duplication. The group of paralogous proteins that are descended from a common ancestor by gene duplication is called a protein family.

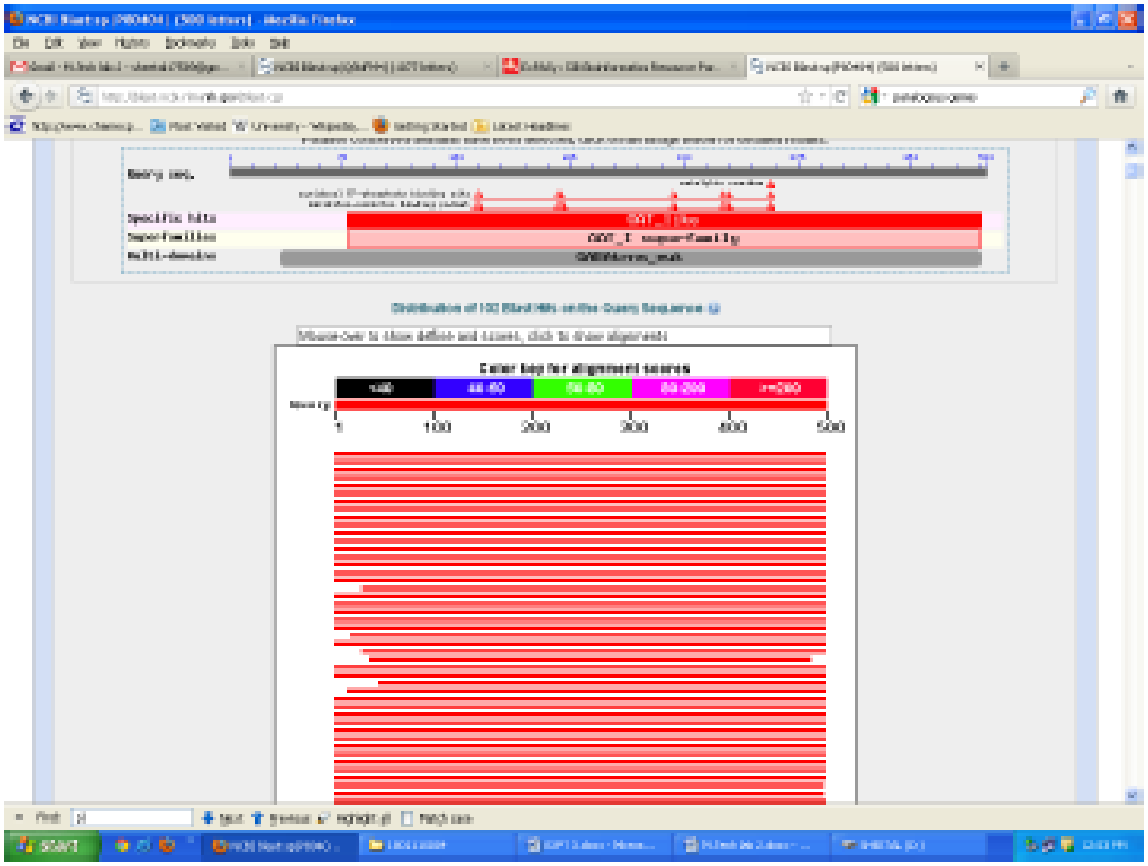
**Method:**

1. Retrieve the sequence from NCBI.
2. Paste the sequence in the Query box in blastp.
3. Run against a non-redundant database (nr).

**Result and inference:**

**Query ID** - gi|48429239|sp|P80404.3|GABT_HUMAN

**Description** - 4-aminobutyrate aminotransferase, mitochondrial precursor [Homo sapiens]



**Paralogous protein:**

AAB38510.1 gamma-aminobutyric acid transaminase [Homo sapiens]

Score = 1000 bits (2585), Expect = 0.0, Method: Compositional matrix adjust.

Identities = 482/500 (96%), Positives = 483/500 (97%), Gaps = 0/500 (0%)

**Ex. No.** **Pairwise sequence alignment**

**Introduction:**

A sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns. Pairwise sequence alignment methods are used to find the best-matching piecewise (local) or global alignments of two query sequences. Pairwise alignments can only be used between two sequences at a time, but they are efficient to calculate and are often used for methods that do not require extreme precision (such as searching a database for sequences with high similarity to a query).

The three primary methods of producing pairwise alignments are dot-matrix methods, dynamic programming, and word methods; however, multiple sequence alignment techniques can also align pairs of sequences.

1. Perform the local alignment between following sequences using any two variants of BLOSUM. Comment on the result.

**Aim: To** perform the local alignment between the given sequences using any two variants of BLOSUM

**Introduction:**

The **BLOSUM** (**BLO**cks of Amino Acid **SU**bstitution **M**atrix) matrix is a substitution matrix used for sequence alignment ofproteins. BLOSUM matrices are used to score alignments between evolutionarily divergent protein sequences. They are based on local alignments. BLOSUM matrices were first introduced in a paper by Henikoff and Henikoff. They scanned the BLOCKS database for very conserved regions of protein families (that do not have gaps in the sequence alignment) and then counted the relative frequencies of amino acids and their substitution probabilities. Then, they calculated a log-odds score for each of the 210 possible substitutions of the 20 standard amino acids. All BLOSUM matrices are based on observed alignments; they are not extrapolated from comparisons of closely related proteins like the PAM Matrices. Several sets of BLOSUM matrices exist using different alignment databases, named with numbers. BLOSUM matrices with high numbers are designed for comparing closely related sequences, while those with low numbers are designed for comparing distant related sequences. For example, BLOSUM80 is used for less divergent alignments, and BLOSUM45 is used for more divergent alignments. The matrices were created by merging (clustering) all sequences that were more similar than a given percentage into one single sequence and then comparing those sequences (that were all more divergent than the given percentage value) only; thus reducing the

contribution of closely related sequences. The percentage used was appended to the name, giving BLOSUM80 for example where sequences that were more than 80% identical were clustered.

**Method:**

1. Enter the Given Sequences in Blastp.

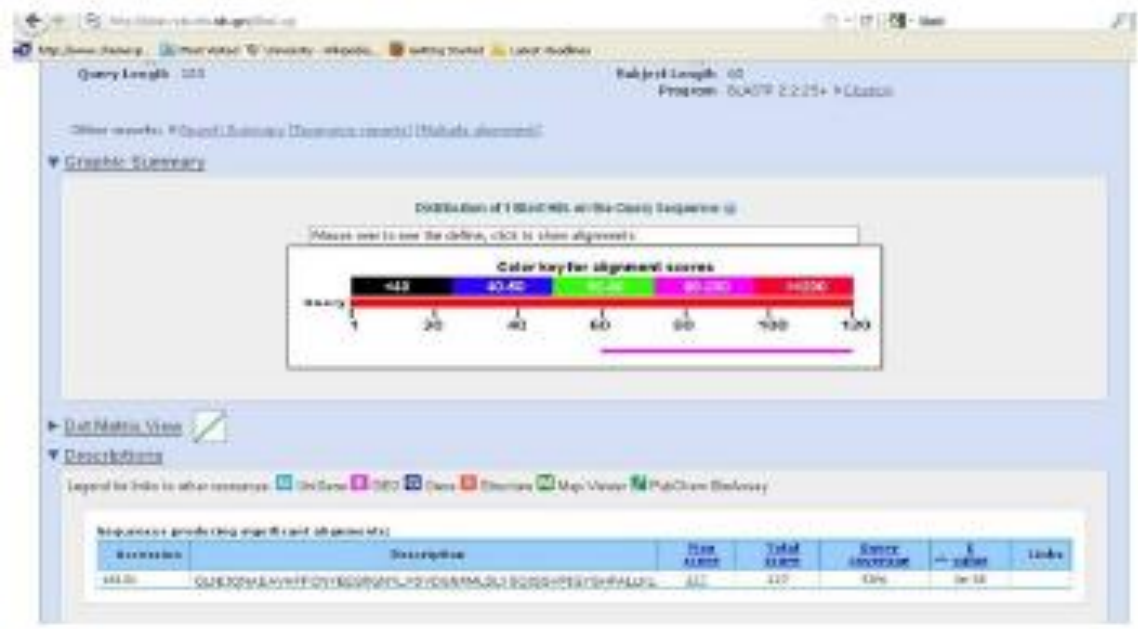
MASMLLAQRLACSFQHSYRLLVPGSRHISQAAAKVDVEFDYDGPLMKTEVPGPRSQEL
MKQLNIIQNAEAVHFFCNYEESRGNYLVDVDGNRMLDLYSQISSVPIGYSHPALLKLIQQ
PQNASMFVNRPALGILPPENFVEKLRQSLLSVAPKGMSQLITMACGSCSNENALKTIFM
WYRQLNIIQNAEAVHFFCNYEESRGNYLVDVDGNRMLDLYSQISSVPIGYSHPALLKLIQ
QPQNASMFVNRPALGILPPENFVEKLRQSLLSVAPKGMSQLITMACGSCSNENALKTIFM
WYR

1. Run Blast with following algorithmic parameters-

Matrix: BLOSUM 62

Gap costs: Existence 11 extension 1

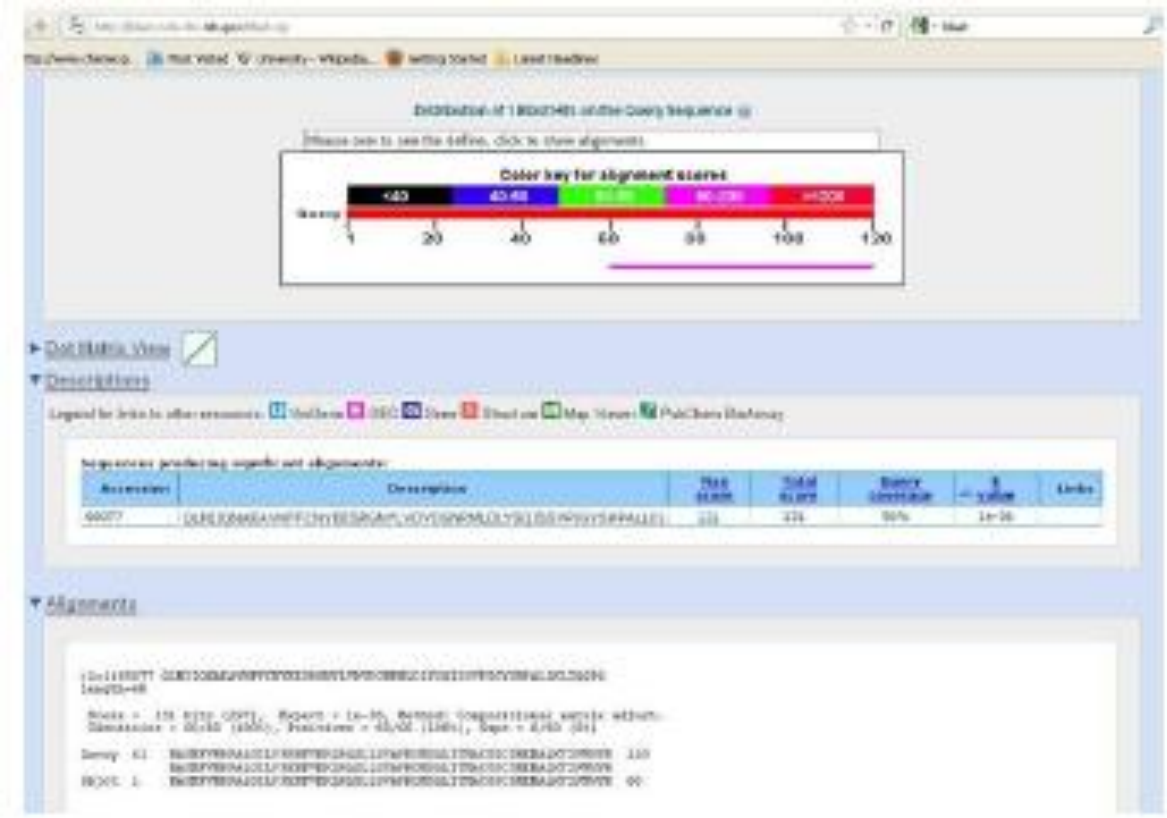
**Result and inference:**




>lcl|16131
QLNIIQNAEAVHFFCNYEESRGNYLVDVDGNRMLDLYSQISSVPIGYSHPALLKLIQQPQ

Length=60

Score = 127 bits (318), Expect = 3e-35, Method: Compositional matrix adjust.
Identities = 60/60 (100%), Positives = 60/60 (100%), Gaps = 0/60 (0%)
Matrix: BLOSUM 80
Gap costs: Existence 10 extension 1




>lcl|60077
QLNIIQNAEAVHFFCNYEESRGNYLVDVDGNRMLDLYSQISSVPIGYSHPALLKLIQQPQ
Length=60
Score = 131 bits (297), Expect = 1e-36, Method: Compositional matrix adjust.
Identities = 60/60 (100%), Positives = 60/60 (100%), Gaps = 0/60 (0%)

## 1. Obtain the global alignment between the following sequences.

QLNIIQNAEAVHFFCNYEESRGNYLVDVDGNRMLDLYSQISSVPIGYSHPALLKLIQQPQ
NASMFVNRPALGILPPENFVEKLRQSLLSVAPKGMSQLITMACGSCSNENALKTIFMWY
RQLNIIQNAEAVHFFCNYEESRGNYLYSQISSVPASMFVNRPALGILPPENFVSCSNENAL
KTIFMWY

**Aim:** To obtain the global alignment between the following sequences

**Introduction:**

Global Alignment is an alignment that assumes that the two proteins are basically similar over the entire length of one another. The alignment attempts to match them to each other from end to end, even though parts of the alignment are not very convincing

**Method:**

1. Choose the Needleman-Wunsch **Global Sequence Alignment Tool.**
2. Enter the two query sequences to b matched.

**Result and inference:**

>lcl|1345 unnamed protein product
Length=67
NW Score = 260
Identities = 67/120 (56%), Positives = 67/120 (56%), Gaps = 53/120 (44%)

**Ex. No.** **Multiple Sequence and Phylogenetic Analysis**

**Aim:** To identify the 10- homologues sequences of P68871 of various origins. Find the conserved region existing between them comment on the same. Comment on the evolutionary relationship between the sequences.

**Introduction:**

**Multiple sequence alignment (MSA)** is a sequence alignment of three or more biological sequences, generally protein, DNA, or RNA. In many cases, 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. From the resulting MSA, sequence homology can be inferred and phylogenetic analysis can be conducted to assess the sequences' shared evolutionary origins. Multiple sequence alignment is often used to assess sequence conservation of protein domains, tertiary and secondary structures, and even individual amino acids or nucleotides. **Conserved domains (CD)** in proteins play a crucial role in protein interactions, DNA binding, enzyme activity, and other important cellular processes**.** Protein domains are often conserved across many species, and as such, they offer an interesting dataset in how genomes maintain them with relationship to other conserved domains, as well as to proteome size. A **phylogenetic tree** or **evolutionary tree** is a branching diagram or "tree" showing the inferred evolutionary relationships among various biological species or other entities based upon similarities and differences in their physical and/or genetic characteristics. The taxa joined together in the tree are implied to have descended from a common ancestor . A **cladogram** is a diagram used in cladistics which shows ancestral relations between organisms, to represent the evolutionary tree of life.

**Phylogeny.fr** has been designed to provide a high performance platform that transparently chains programs relevant to phylogenetic analysis in a comprehensive, and flexible pipeline. Although phylogenetic aficionados will be able to find most of their favorite tools and run sophisticated analysis, the primary philosophy of Phylogeny.fr is to assist biologists with no experience in phylogeny in analyzing their data in a robust way. The Phylogeny.fr platform offers a phylogeny pipeline which can be executed through three main modes:

The "One Click mode" targets users that do not wish to deal with program and parameter selection. By default, the pipeline is already set up to run and connect programs recognized for their accuracy and speed (MUSCLE for multiple alignment and PhyML for phylogeny) to reconstruct a robust phylogenetic tree from a set of sequences.

In the "Advanced mode", the Phylogeny.fr server proposes the succession of the same programs but users can choose the steps to perform (multiple sequence alignment, phylogenetic reconstruction, tree drawing) and the options of each program.

The "A la carte mode" offers the possibility of running and testing more alignment and phylogeny programs, such as MUSCLE, ClustalW, T-Coffee, PhyML,BioNJ, TNT.

**Method:**

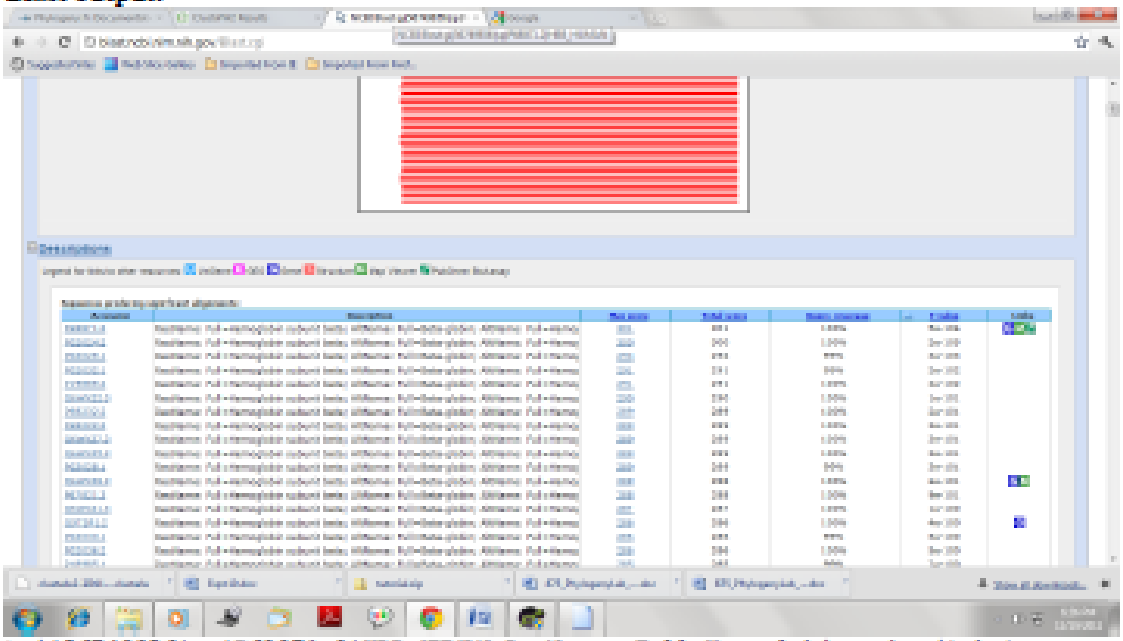
1. Run a blastp for the protein Id: P68871
2. Choose 10 homologous proteins and save in .txt format
3. Input these sequences in Clustalw. Determine conserved reagions.
4. Access the "One Click" mode

This is a "default" mode which proposes a pipeline already set up to run and connect programs recognized for their accuracy and speed (MUSCLE for multiple alignment, optionally Gblocks for alignment curation, PhyML for phylogeny and finally TreeDyn for tree drawing) to reconstruct a robust phylogenetic tree from a set of sequences.

1. Copy and paste the set of sequences in the FASTA, all the parameters are those of programs by default.

**Results and inference:**

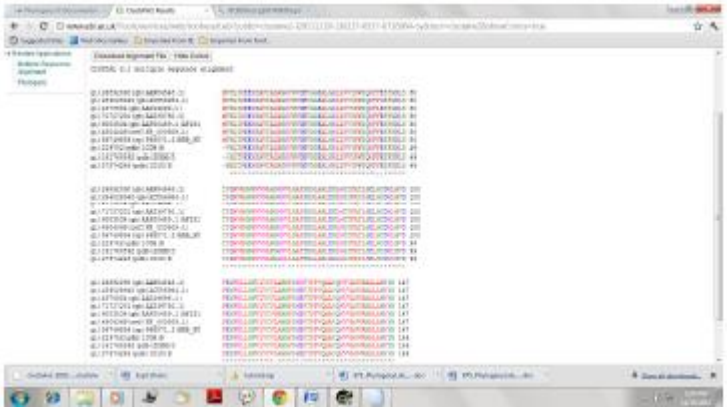
**Blast output:**



>gi|56749856|sp|P68871.2|HBB_HUMAN RecName: Full=Hemoglobin subunit beta; AltName: Full=Beta-globin; AltName: Full=Hemoglobin beta chain; Contains: RecName: Full=LVV-hemorphin-7
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH

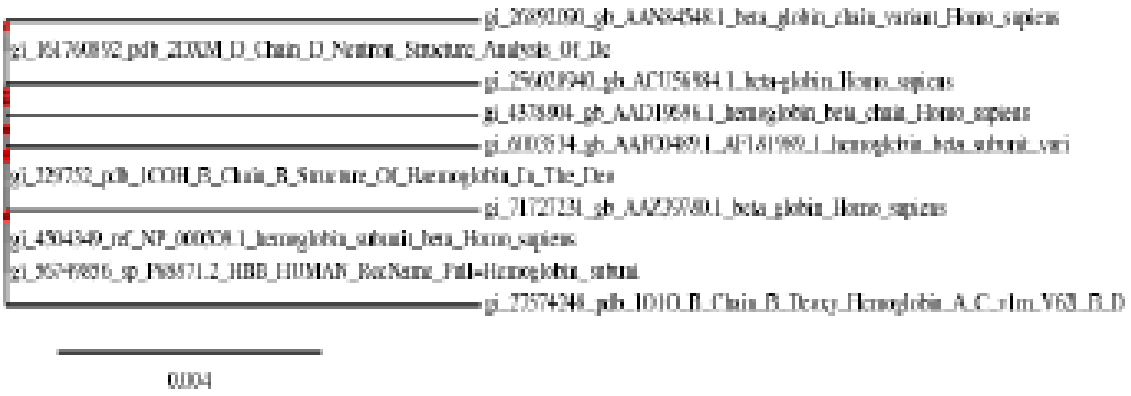
```
>gi|229752|pdb|1COH|B Chain B, Structure Of Haemoglobin In The Deoxy
Quaternary State With Ligand Bound At The Alpha Haems
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGA
FSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANA
LAHKYH
>gi|6003534|gb|AAF00489.1|AF181989_1 hemoglobin beta subunit variant [Homo
sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVPAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
>gi|71727231|gb|AA239780.1| beta globin [Homo sapiens]
MVHLTPKEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
>gi|4378804|gb|AAD19696.1| hemoglobin beta chain [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFLESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
>gi|26892090|gb|AAN84548.1| beta globin chain variant [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKKFTPPVQAAYQKVVAGVAN
ALAHKYH
>gi|256028940|gb|ACU56984.1| beta-globin [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFKSFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
>gi|161760892|pdb|2DXM|D Chain D, Neutron Structure Analysis Of Deoxy Human
Hemoglobin
XHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGA
FSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANA
LAHKYH
>gi|27574248|pdb|1010|B Chain B, Deoxy Hemoglobin (A,C:vlm,V621;
B,D:vlm,V671)
MHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKLLGA
FSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANA
LAHKYH
```

**Clustaw Alignment:**

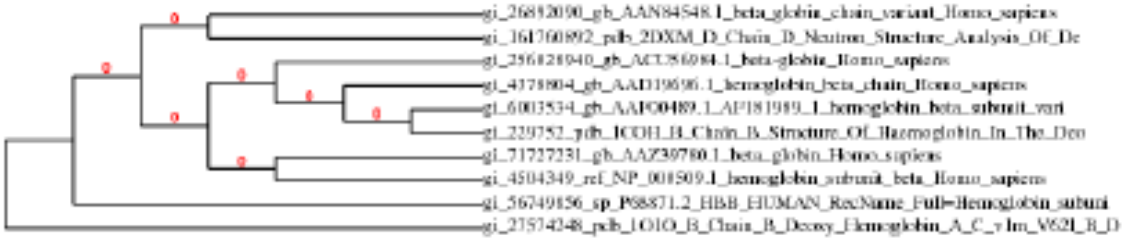


Red: small, hydrophobic, aromatic, not Y.
Blue: acidic. Magenta: basic.
Green: hydroxyl, amine, amide, basic.
Gray: others.
"*": identical.
":": conserved substitutions (same colour group).
".": semi-conserved substitution (similar shapes)
For the first 50 aa of the query,
Conserved regions:
HLTP
Semi-conserved:
E
Conserved:
EKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRF
Semi-conserved:
FE
Conserved:
SFGDLS

**One-Click Mode:**




**PHYLOGRAM**



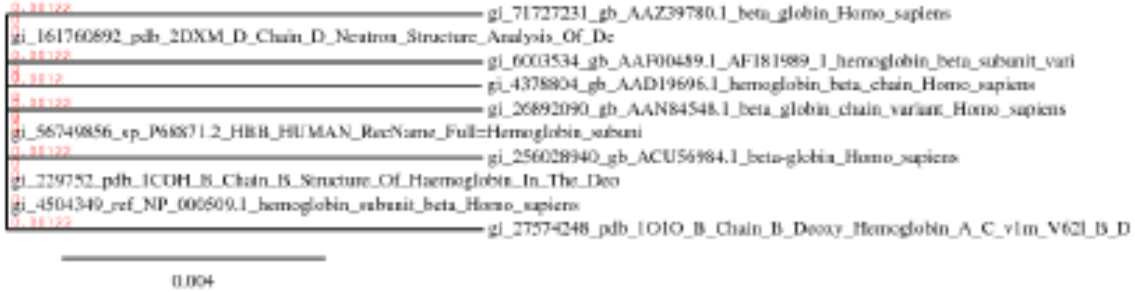
## CLADOGRAM

>gi|26892090|gb|AAN84548.1| beta globin chain variant [Homo sapiens]
>gi|161760892|pdb|2DXM|D Chain D, Neutron Structure Analysis Of Deoxy Human Hemoglobin
>gi|256028940|gb|ACU56984.1| beta-globin [Homo sapiens]
>gi|4378804|gb|AAD19696.1| hemoglobin beta chain [Homo sapiens]
>gi|6003534|gb|AAF00489.1|AF181989_1 hemoglobin beta subunit variant [Homo sapiens]
>gi|229752|pdb|1COH|B Chain B, Structure Of Haemoglobin In The Deoxy Quaternary State With Ligand Bound At The Alpha Haems
>gi|71727231|gb|AAZ39780.1| beta globin [Homo sapiens]
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
>gi|56749856|sp|P68871.2|HBB_HUMAN RecName: Full=Hemoglobin subunit beta
>gi|27574248|pdb|1O1O|B Chain B, Deoxy Hemoglobin (A,C:v1m,V62l; B,D:v1m,V67l)

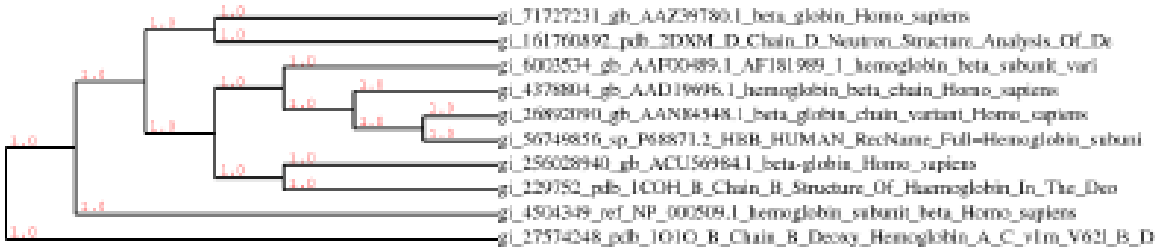
**A la Carte Result:**
MA: ClustalW
Align curation: GBlocks
Contruction of tree: PhyML
(http://www.phylogeny.fr/version2_cgi/one_task.cgi?task_type=phyml)

View: TreeDyn

## PHYLOGRAM




## CLADOGRAM



Sequences gi_71727231_gb_AAZ39780.1_beta_globin_Homo_sapiens and gi_161760892_pdb_2DXM_D_Chain_D_Neutron_Structure_Analysis_Of_De are apart by a distance of 1.0
Sequences gi_56749856_sp_P68871.2_HBB_HUMAN_RecName_Full=Hemoglobin_subunit and gi_26892090_gb_AAN84548.1_beta_globin_chain_variant_Homo_sapiens are also apart by a distance of 1.0
gi_256028940_gb_ACU56984.1_beta-globin_Homo_sapiens and gi_229752_pdb_1COH_B_Chain_B_Structure_Of_Haemoglobin_In_The_Deo are apart by a distance of 1.0
These are the closely related sequences.
Sequence gi_27574248_pdb_1O1O_B_Chain_B_Deoxy_Hemoglobin_A_C_v1m_V62l_B_D is the most distantly related sequence to the closely related sequences.
The distance between gi_4378804_gb_AAD19696.1_hemoglobin_beta_chain_Homo_sapiens and the cluster formed by
gi_56749856_sp_P68871.2_HBB_HUMAN_RecName_Full=Hemoglobin_subunit and
gi_26892090_gb_AAN84548.1_beta_globin_chain_variant_Homo_sapiens is 3.0
Sequences gi_56749856_sp_P68871.2_HBB_HUMAN_RecName_Full=Hemoglobin_subunit, gi_26892090_gb_AAN84548.1_beta_globin_chain_variant_Homo_sapiens and

gi_4378804_gb_AAD19696.1_hemoglobin_beta_chain_Homo_sapiens forms a new cluster which is apart from gi_6003534_gb_AAF00489.1_AF181989_1_hemoglobin_beta_subunit_vari by a distance of 1.0

The distance between the above cluster and the cluster formed by from gi_6003534_gb_AAF00489.1_AF181989_1_hemoglobin_beta_subunit_vari,
gi_56749856_sp_P68871.2_HBB_HUMAN_RecName_Full=Hemoglobin_subunit,
gi_26892090_gb_AAN84548.1_beta_globin_chain_variant_Homo_sapiens and
gi_4378804_gb_AAD19696.1_hemoglobin_beta_chain_Homo_sapiens is 1.0
gi_4504349_ref_NP_000509.1_hemoglobin_subunit_beta_Homo_sapiens is distantly related to above clusters.

**Ex.No.** **Find the presence of super secondary structure**

**Aim:** To determine the presence of super secondary structure in Q9H6F5

**Introduction:**

A supersecondary structure is a compact three-dimensional protein structure of several adjacent elements of secondary structure that is smaller than a protein domain or a subunit. Super secondary structures can act as nucleations in the process of protein folding. Examples include β-hairpins, α-helix hairpins, and β-α-β motifs. Supersecondary structures involve the association of secondary structures in a particular geometric arrangement. If we think of each secondary structure as a 'unit' then a supersecondary structure would be comprised of at least two 'units' of secondary structure. Some of these supersecondary structures are known to have a specific biological, or structural, role but for others their role is unknown.

This presentation outlines some supersecondary structures, or structural motifs, seen in proteins.

**Method:**

1. Open Uniprot http://www.uniprot.org/

2. Enter the protein ID.

**Result and inference:**

Q9H6F5 (CCD86_HUMAN)

Source: Human

Description: Coiled-coil domain-containing protein

Supersecondary Structure: It contains one coiled coil domain, a type of secondary structure composed of two or more alpha helices which entwine to form a cable structure. 1-360 residues

**Ex. No.** **Secondary Structure prediction**

**Aim : To predict secondary structure of the give protein sequences**

**Introduction:**

Protein secondary structure includes the regular polypeptide folding patterns such as helices, sheets, and turns. The backbone or main chain of a protein refers to the atoms that participate in peptide bonds, ignoring the side chains of the amino acid.

The conformation of the backbone can therefore be described by the torsion angles (also called dihedral angles or rotation angles) around the Phi and the Psi of each residue. The helix structure looks like a spring. The most common shape is a right handed a-helix defined by the repeat length of 3.6 amino acid residues and a rise of 5.4 Å per turn.

Secondary structure in proteins consists of local inter-residue interactions mediated by hydrogen bonds, or not. The most common secondary structures are alpha helices and beta sheets. Other helices, such as the $3_{10}$ helix and $\pi$ helix, are calculated to have energetically favorable hydrogen-bonding patterns but are rarely if ever observed in natural proteins except at the ends of α helices due to unfavorable backbone packing in the center of the helix. Other extended structures such as the polyproline helix and alpha sheet are rare in native state proteins but are often hypothesized as important protein folding intermediates. Tight turns and loose, flexible loops link the more "regular" secondary structure elements. The random coil is not a true secondary structure, but is the class of conformations that indicate an absence of regular secondary structure.

Amino acids vary in their ability to form the various secondary structure elements. Proline and glycine are sometimes known as "helix breakers" because they disrupt the regularity of the α helical backbone conformation; however, both have unusual conformational abilities and are commonly found in turns. Amino acids that prefer to adopt helical conformations in proteins include methionine, alanine, leucine, glutamate and lysine ("MALEK" in amino-acid 1-letter codes); by contrast, the large aromatic residues (tryptophan, tyrosine and phenylalanine) and Cβ-branched amino acids (isoleucine, valine, and threonine) prefer to adopt β-strand conformations. However, these preferences are not strong enough to produce a reliable method of predicting secondary structure from sequence alone.

There are several methods for defining protein secondary structure (e.g. DEFINE, DSSP, STRIDE (protein)).

**Structural features of the three major forms of protein helices**

| Geometry attribute | α-helix | $3_{10}$ helix | π-helix |
|---|---|---|---|
| Residues per turn | 3.6 | 3.0 | 4.4 |
| Translation per residue | 1.5 Å (0.15 nm) | 2.0 Å (0.20 nm) | 1.1 Å (0.11 nm) |
| Radius of helix | 2.3 Å (0.23 nm) | 1.9 Å (0.19 nm) | 2.8 Å (0.28 nm) |
| Pitch | 5.4 Å (0.54 nm) | 6.0 Å (0.60 nm) | 4.8 Å (0.48 nm) |

1. **To Compare the secondary structures of the following sequences and comment on the result.**

>1
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE
LKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDF
GADAQGAMNKALELFRKDMASNYKELGFQG
>2
MDPKQTTLLCLVLCLGQRIQAQEGDFPMPFISAKSSPVIPLDGSVKIQCQAIREAYLTQLMIIK
STYREIGRRLKFWNETDPEFVIDHMDANKAGRYQCQYRIGHYRFRYSDTLELVVTGLYGKP
FLSADRGLVLMPGENISLTCSSAHIPFDRFSLAKEGELSLPQHQSGEHPANFSLGPVDLNVSGI
RCYGWYNRSPYLWSFPSNALELVVTDSIHQDYTTQNLIRMAVAGLVLVALLAILVENWHSH
ALNKEASADVAEPSWSQQMCQPGLTFARTPSVCK
**Methods:**

1. Take the sequence from uniprot or copy the sequence if already given
2. Go to http://www.compbio.dundee.ac.uk/www-jpred/
3. Paste the sequence and click on make prediction
4. Wait for the software to predict the structure
5. Once Job is done . Save the output.

**Results:** <u>Output for seq 1 and 2:</u>
Colour code for alignment:
Blue - Complete identity at a position
Shades of red - The more red a position is, the higher the level of
conservation of chemical properties of the amino acids
Jnet - Final secondary structure prediction for query
jalign - Jnet alignment prediction
jhmm - Jnet hmm profile prediction
jpssm - Jnet PSIBLAST pssm profile prediction
Lupas - Lupas Coil prediction (window size of 14, 21 and 28)
Note on coiled coil predictions - = less than 50% probability
c = between 50% and 90% probability C = greater than 90% probability
Jnet_25 - Jnet prediction of burial, less than 25% solvent accesibility
Jnet_5 - Jnet prediction of burial, less than 5% exposure
Jnet_0 - Jnet prediction of burial, 0% exposure
Jnet Rel - Jnet reliability of prediction accuracy, ranges from 0 to 9, bigger is better.

**Ex. No.** **Tertiary Structure Predicition**

**Aim:** Determine the 3d structure of human gaba transaminase using homology modeling

**Introduction:**

The tertiary structure of a protein or any other macromolecule is its three-dimensional structure, as defined by the atomic coordinates.

Tertiary structure is considered to be largely determined by the protein's primary structure - the sequence of amino acids of which it is composed. Efforts to predict tertiary structure from the primary structure are known generally as protein structure prediction. However, the environment in which a protein is synthesized and allowed to fold are significant determinants of its final shape and are usually not directly taken into account by current prediction methods.

Most such methods do rely on comparisons between the sequence to be predicted and sequences of known structure in the Protein Data Bank and thus account for environment indirectly, assuming the target and template sequences share similar cellular contexts.

In globular proteins, tertiary interactions are frequently stabilized by the sequestration of hydrophobic amino acid residues in the protein core, from which water is excluded, and by the consequent enrichment of charged or hydrophilic residues on the protein's water-exposed surface.

In secreted proteins that do not spend time in the cytoplasm, disulfide bonds between cysteine residues help to maintain the protein's tertiary structure. A variety of common and stable tertiary structures appear in a large number of proteins that are unrelated in both function and evolution - for example, many proteins are shaped like a TIM barrel, named for the enzyme triosephosphateisomerase. Another common structure is a highly stable dimeric coiled coil structure composed of 2-7 alpha helices. Proteins are classified by the folds they represent in databases like SCOP and CATH.

Homology modeling, also known as comparative modeling of protein refers to constructing an atomic-resolution model of the "target" protein from its amino acid sequence and an experimental three-dimensional structure of a related homologous protein (the "template").

Homology modeling relies on the identification of one or more known protein structures likely to resemble the structure of the query sequence, and on the production of an alignment that maps residues in the query sequence to residues in the template sequence.

It has been shown that protein structures are more conserved than protein sequences amongst homologues, but sequences falling below a 20% sequence identity can have very

different structure. Homology modeling aims to build three-dimensional protein structure models using experimentally determined structures of related family members as templates.

SWISS-MODEL workspace is an integrated Web-based modeling expert system.

For a given target protein, a library of experimental protein structures is searched to identify suitable templates. On the basis of a sequence alignment between the target protein and the template structure, a three-dimensional model for the target protein is generated.

Model quality assessment tools are used to estimate the reliability of the resulting models. Homology modeling is currently the most accurate computational method to generate reliable structural models and is routinely used in many biological applications.

Typically, the computational effort for a modeling project is less than 2 h. However, this does not include the time required for visualization and interpretation of the model, which may vary depending on personal experience working with protein structures.

**Swiss PDB viewer and swiss modeler are used as homology modeling software and workspace.**

Swiss-Pdb Viewer provides a user friendly interface allowing to analyze several proteins at the same time.

1. Superimposition - structural alignments and compare their active sites or any other relevant parts

2. . Make amino acid mutations

3. Generate Hydrogen bonds

4. Calculate angles and distances between atoms

5. Tightly linked to Swiss-Model, an automated homology modeling server

6. Thread a protein primary sequence onto a 3D template

7. Build missing loops and refine sidechain packing

8. Read electron density maps and build into the density

9. Perform energy minimization

10. POV-Ray scenes can be generated for stunning ray-traced quality images

**Swiss Modeller**

The SWISS-MODEL Workspace is a web-based integrated service dedicated to protein structure homology modelling. It assists and guides the user in building protein homology models at different levels of complexity.

Successful model building requires at least one experimentally determined 3D structure (template) that shows significant amino acid sequence similarity with the target sequence.

**Building a homology model comprises four main steps:**

identification of structural template(s),

alignment of target sequence and template structure(s),

model building, and

model quality evaluation.

These steps can be repeated until a satisfying modelling result is achieved. Each of the four steps requires specialized software and access to up-to-date protein sequence and structure databases.

Protein sequence and structure databases necessary for modelling are accessible from the workspace and are updated in regular intervals.

Software tools for template selection, model building, and structure quality evaluation can be invoked from within the workspace.

A personal working environment (workspace), where several modelling projects can be carried out in parallel, is provided for each user.

**Methods:**

1.load the 1OHV protein

2.select the chain A, in control panel and in the menu bar click the bulid option and select the inverse selection and then click on the remove selected residues.

3. save it separately as 1OHVA.pdb

4. open the empty window again, and click the swissmodel to load the raw sequence.

5.open the pdb file through the import structures in the "File" menubar.

6. Click the magic fit, iterative magic fit from Fit option in the menubar.

7. Open the alignment window from the wind and select the residues which are not aligned.

8.Delete the residues which are not aligned using the Build option in the menubar and click the remove residues and save it.

9. Now submit this to the swiss modelling request for the raw

10. Download the modelled protein and open in the swiss viewer.

11. In Bulid option, click the energy minimization.

12. open the seq-structure aligned protein (step 8) and energy minimized protein in the viewer and click the improve fit

13. Calculate the RMS value from Fit option

14. Render the model in 3D view.

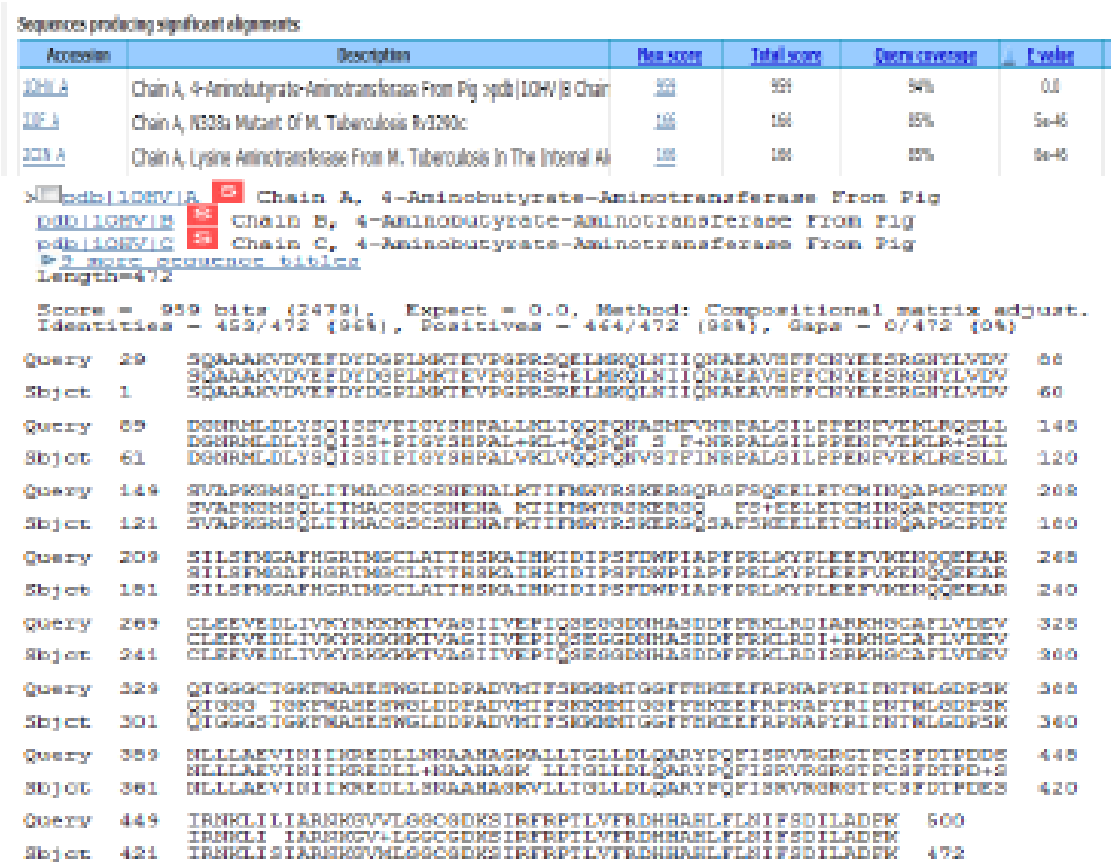
15. Use Protein Structure & Model Assessment Tools for analyzing the protein.

**Result and Inference:**

Query sequence (gabat.txt)

>sp|P80404|GABT_HUMAN 4-aminobutyrate aminotransferase, mitochondrial OS=Homo sapiens GN=ABAT PE=1 SV=3

MASMLLAQRLACSFQHSYRLLVPGSRHISQAAAKVDVEFDYDGPLMKTEVPGPRSQELMKQLNI
IQNAEAVHFFCNYEESRGNYLVDVDGNRMLDLYSQISSVPIGYSHPALLKLIQQPQNASMFVNRP
ALGILPPENFVEKLRQSLLSVAPKGMSQLITMACGSCSNENALKTIFMWYRSKERGQRGFSQEEL
ETCMINQAPGCPDYSILSFMGAFHGRTMGCLATTHSKAIHKIDIPSFDWPIAPFPRLKYPLEEFVKE
NQQEEARCLEEVEDLIVKYRKKKKTVAGIIVEPIQSEGGDNHASDDFFRKLRDIARKHGCAFLVD
EVQTGGGCTGKFWAHEHWGLDDPADVMTFSKKMMTGGFFHKEEFRPNAPYRIFNTWLGDPSK
NLLLAEVINIIKREDLLNNAAHAGKALLTGLLDLQARYPQFISRVRGRGTFCSFDTPDDSIRNKLIL
IARNKGVVLGGCGDKSIRFRPTLVFRDHHAHLFLNIFSDILADFK



Sequences producing significant alignments:

| Accession | Description | Max score | Total score | Query coverage | E value |
|---|---|---|---|---|---|
| 1OHV A | Chain A, 4-Aminobutyrate-Aminotransferase From Pig >pdb\|1OHV\|B Chain | 959 | 959 | 94% | 0.0 |
| 1SF A | Chain A, W338a Mutant Of M. Tuberculosis Rv3290c | 186 | 186 | 85% | 5e-46 |
| 2CSJ A | Chain A, Lysine Aminotransferase From M. Tuberculosis In The Internal Al | 186 | 186 | 85% | 6e-46 |

>pdb|1OHV|A Chain A, 4-Aminobutyrate-Aminotransferase From Pig
pdb|1OHV|B Chain B, 4-Aminobutyrate-Aminotransferase From Pig
pdb|1OHV|C Chain C, 4-Aminobutyrate-Aminotransferase From Pig
More sequence titles
Length=472

Score = 959 bits (2479), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 453/472 (96%), Positives = 464/472 (98%), Gaps = 0/472 (0%)

Query 29 SQAAAKVDVEFDYDGPLMKTEVPGPRSQELMKQLNIIQNAEAVHFFCNYEESRGNYLVDV 88
SQAAAKVDVEFDYDGPLMKTEVPGPRS+ELMKQLNIIQNAEAVHFFCNYEESRGNYLVDV
Sbjct 1 SQAAAKVDVEFDYDGPLMKTEVPGPRSRELMKQLNIIQNAEAVHFFCNYEESRGNYLVDV 60

Query 89 DGNRMLDLYSQISSVPIGYSHPALLKLIQQPQNASMFVNRPALGILPPENFVEKLRQSLL 148
DGNRMLDLYSQISS+PIGYSHPAL+KL+QQPQN S F+NRPALGILPPENFVEKLR+SLL
Sbjct 61 DGNRMLDLYSQISSIPIGYSHPALVKLVQQPQNVSIFINRPALGILPPENFVEKLRESLL 120

Query 149 SVLPKGHNSQLITMACGSCSNENALKTIFMWYRSKERSQRGFSQEELETCMINQAPGCPDY 208
SVLPKGHNSQLITMACGSCSNENA KTIFMWYRSKERSQ FS+EELETCMINQAPGCPDY
Sbjct 121 SVLPKGHNSQLITMACGSCSNENAFKTIFMWYRSKERSQSAFSNEELETCMINQAPGCPDY 180

Query 209 SILSFMGAFHGRTMGCLATTRSKAIHKIDIPSFDWPIAPFPRLKYPLEEFVKENQQEEAR 268
SILSFMGAFHGRTMGCLATTRSKAIHKIDIPSFDWPIAPFPRLKYPLEEFVKENQQEEAR
Sbjct 181 SILSFMGAFHGRTMGCLATTRSKAIHKIDIPSFDWPIAPFPRLKYPLEEFVKENQQEEAR 240

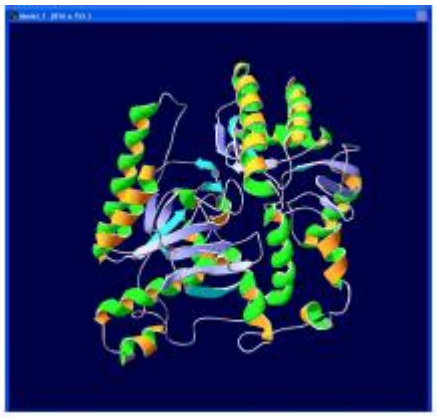
Query 269 CLEEVEDLIVKYRKKKKTVAGIIVEPIQSEGGDNHASDDFFRKLRDIARKHGCAFLVDEV 328
CLEEVEDLIVKYRKKKKTVAGIIVEPIQSEGGDNHASDDFFRKLRDI+RKHGCAFLVDEV
Sbjct 241 CLEEVEDLIVKYRKKKKTVAGIIVEPIQSEGGDNHASDDFFRKLRDISRKHGCAFLVDEV 300

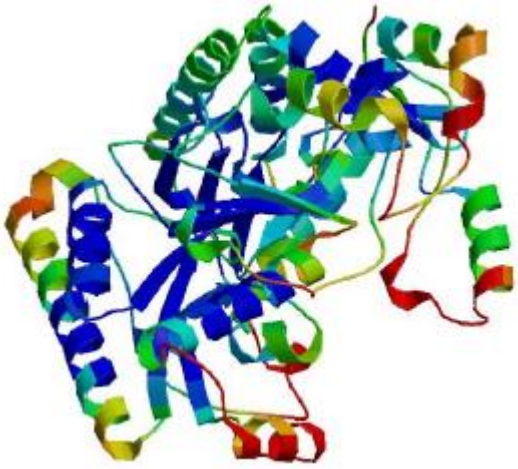
Query 329 QIGGGCTGKFWAHEHWGLDDPADVMTFSKKMMIGGFFHKEEFRPNAPYRIFNTWLGDPSK 388
QIGGG TGKFWAHEHWGLDDPADVMTFSKKMMIGGFFHKEEFRPNAPYRIFNTWLGDPSK
Sbjct 301 QIGGGSTGKFWAHEHWGLDDPADVMTFSKKMMIGGFFHKEEFRPNAPYRIFNTWLGDPSK 360

Query 389 NLLLAEVINIIKREDLLNNAANAGKALLTGLLDLQARYPQFISRVRGRGTFCSFDTPDDS 448
NLLLAEVINIIKREDLL+NAANAGK LLTGLLDLQARYPQFISRVRGRGTFCSFDTPD+S
Sbjct 361 NLLLAEVINIIKREDLLSNAANAGKVLLTGLLDLQARYPQFISRVRGNGTFCSFDIPDES 420

Query 449 IRNKLILIARSKGVVLGGCGDKSIRFRPTLVFRDHHAHLFLNIFSDILADFK 500
IRNKLI IARSKGV+LGGCGDKSIRFRPTLVFRDHHAHLFLNIFSDILADFK
Sbjct 421 IRNKLISIARSKGVVLGGCGDKSIRFRPTLVFRDHHAHLFLNIFSDILADFK 472

**Gabat.txt and 1OHVA.pdb Modeled Structure at swisspdb viewer and swiss modeler**





Energy minimization score: -26789.707

RMSD: 0.07A

**Quality information:**
QMEAN Z-Score: -1.129

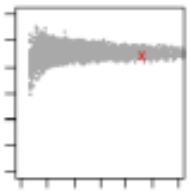
**Ligand information:**

**Global Model Quality Estimation:**

**QMEAN4 global scores:**

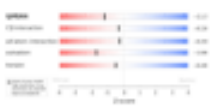
QMEANscore4: 0.695

Estimated absolute model quality:



Z-Score: -1.129

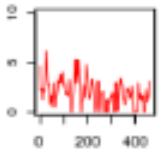
Score components:



Local scores:

Coloring by residue error:

Residue error plot:




**QMEAN4 global scores:**

The QMEAN4 score is a composite score consisting of a linear combination of 4 statistical potential terms (estimated model reliability between 0-1). The pseudo-energies of the contributing terms are given below together with their Z-scores with respect to scores obtained for high-resolution experimental structures of similar size solved by X-ray crystallography:

| Scoring function term | Raw score | Z-score |
|---|---|---|
| C_beta interaction energy | -150.16 | -0.28 |
| All-atom pairwise energy | -12126.34 | -0.20 |
| Solvation energy | -22.16 | -1.68 |
| Torsion angle energy | -115.93 | -0.48 |
| QMEAN4 score | 0.695 | -1.13 |

**Procheck: [+/-]**

```
 +----------<<<  P  R  O  C  H  E  C  K     S  U  M  M  A  R  Y  >>>----------+
 |                                                                             |
 | input_atom_only.pdb   2.5                                    461 residues   |
 |                                                                             |
*| Ramachandran plot:   88.8% core   10.2% allow    0.5% gener    0.5% disall  |
 |                                                                             |
+| All Ramachandrans:   10 labelled residues (out of 459)                      |
+| Chi1-chi2 plots:      2 labelled residues (out of 296)                      |
 | Main-chain params:    6 better     0 inside      0 worse                    |
 | Side-chain params:    5 better     0 inside      0 worse                    |
 |                                                                             |
+| Residue properties: Max.deviation:    11.0            Bad contacts:    1    |
+|                     Bond len/angle:    4.4    Morris et al class:  1  1  2  |
+|     3 cis-peptides                                                          |
 | G-factors           Dihedrals:  -0.02  Covalent:   0.42    Overall:   0.15   |
 |                                                                             |
 | M/c bond lengths:100.0% within limits   0.0% highlighted                    |
 | M/c bond angles:  99.6% within limits   0.4% highlighted                    |
*| Planar groups:    89.5% within limits  10.5% highlighted       1 off graph  |
 |                                                                             |
 +-----------------------------------------------------------------------------+
   + May be worth investigating further.  * Worth investigating further.
```

The qmean score(-1.129) and procheck (rc plot : 99.5% in allowed region)score were within ranges proving protein structure as stable.

**Ex. No.** **Homology Modeling Using Modeller**

**AIM:** To do homology modeling for human gaba transaminase using MODELLER.

**Introduction:**

MODELLER is used for homology or comparative modeling of protein three-dimensional structures. The user provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms.

MODELLER implements comparative protein structure modeling by satisfaction of spatial restraints and can perform many additional tasks, including de novo modeling of loops in protein structures, optimization of various models of protein structure with respect to a flexibly defined objective function, multiple alignment of protein sequences and/or structures, clustering, searching of sequence databases, comparison of protein structures, etc. MODELLER is available for download for most Unix/Linux systems, Windows, and Mac.

MODELLER is used for homology or comparative modeling of protein three-dimensional structures. The user provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms.There are 5 modeling examples that the user can follow:

**Basic Modeling**. Model a sequence with high identity to a template.This exercise introduces the use of MODELLER in a simple case where the template selection and target-template alignments are not a problem.

**Advanced Modeling**. Model a sequence based on multiple templates and bound to a ligand.This exercise introduces the use of multiple templates, ligands and loop refinement in the process of model building with MODELLER.

**Iterative Modeling**. Increase the accuracy of the modeling exercise by iterating the 4 step process.This exercise introduces the concept of MOULDING to improve the accuracy of comparative models.

**Difficult Modeling**. Model a sequence based on a low identity to a template.This exercise uses resources external to MODELLER in order to select a template for a difficult case of protein structure prediction.

**Modeling with cryo-EM.** Model a sequence using both template and cryo-EM data.This exercise assesses the quality of generated models and loops by rigid fitting into cryo-EM maps, and improves them with flexible EM fitting.

**Method:**

1. Take query sequence whose structure needs to be modelled (e.g gabat) in PIR format.

2. Save the file with .ali extension in the bin folder of modeller.

3. Open build_profile.py file. Change the append filename to the query sequence(gabat.ali).

4. Open the command line by clicking the 'Modeller' link from the Start Menu in Windows.

5. Run the build_profile.py.This will search for potentially related sequences of known structure.

Two files are created build_profile_gabat.ali file and build_profile_gabat.prf file.

6. Open the build_profile.prf file and select the sequences which has an e value 0.0 .

7. Download the structures of the selected protein from the PDB and save it in bin folder of modeller.

8. Open the compare.py file.Write the the name of the selected proteins.

9. Run compare.py command in command line. A compare.log output file is created.

10. Choose the sequence with high resolution and moderate identity.

11. Align the query sequence with the template by using align2d command.

12. Two output files are created .pap file and .ali file.

13. Open model_single.py file .Use the above created .ali file .Run the model_single.py command in the command line.

14. 5 possible models are generated .Select the best model which has the lowest dope score.

15. Run evaluate_model.py command for evaluating the selected model.Note the Dope score.

16. Run evaluate_template.py command for evaluating the template. Note the Dope score.

17. Compare the dope score of both model and template.

**Results and Inference:**

**Build_profile_gabat.ali (output for build_profile.py)**

```
>P1;gabat
sequence:gabat: 0: : 0: :::-1.00:-1.00
MASMLLAQRLACSFQHSYRLLVPGSRHISQAAAKVDVEFDYDGPLMKTEVPGPRSQELMKQLNIIQNAEA
VHFFCNYEESRGNYLVDVDGNRMLDLYSQISSVPIGYSHPALLKLIQQPQNASMFVNRPALGILPPENFV
EKLRQSLLSVAPKGMSQLITMACGSCSNENALKTIFMWYRSKERGQRGFSQEELETCMINQAPGCPDYSI
LSFMGAFHGRTMGCLATTHSKAIHKIDIPSFDWPIAPFPRLKYPLEEFVKENQQEEARCLEEVEDLIVKY
RKKKKTVAGIIVEPIQSEGGDNHASDDFFRKLRDIARKHGCAFLVDEVQTGGGCTGKFWAHEHWGLDDPA
```

```
DVMTFSKKMMTGGFFHKEEFRPNAPYRIFNTWLGDPSKNLLLAEVINIIKREDLLNNAAHAGKALLTGLL
DLQARYPQFISRVRGRGTFCSFDTPDDSIRNKLILIARNKGVVLGGCGDKSIRFRPTLVFRDHHAHLFLN
IFSDILADFK*

>P1;2oatA
structure:2oatA: 28: : 404: :::-1.00:-1.00
-----------------------------------------------------------------------
-------ERGKGIYLWDVEGRKYFDFLSSYSAVNQGHCHPKIVNALKSQVDKLTLTSRAVLG--
YEEYITKL---------FNYHKVLPMNTGVEAGETACKLARKW---------GYTVKGIQKYKA------
---KIVFAAGNFWGRTLSAISSS-------TDPTSYD-GFGPF----MPGFDIIPYND------
PALERAL-----
QDPNVAAFMVEPIQGEAGVVVPDPGYLMGVRELCTRHQVLFIADEIQTGLARTGRWLAVDYENV--
RPDIVLLG-
KALSGGLYDDDIMLTIKPGEHGSTYGGNPLGCRVAIAALEVLEEENLAENADKLGIILRNELMKLPS---
DVVTAVRGKGLLNAIVIKEDWDA
WKVCLRLRDNGLLAKPTHGDIIRFAPPLVIKEDELRESIEIINKTILSF-*

>P1;1d7uA
structure:1d7uA: 28: : 427: :::-1.00:-1.00
-----------------------------------------------------------------------
-----
--ERAKGSFVYDADGRAILDFTSGQMSAVLGHCHPEIVSVIGEYAGKSGMLSRP----------
VVDLATRLANI
TPPGLDRALLLSTGAESNEAAIR----------------------MAKLVTG--
KYEIVGFAQSWHGMTGAAA
SATYSKGVGPAAVGSFAIP-APFPR-------FERNGAYDYLAELDYAFDLI--
DRQSSGNLAAFIAEPILSSGG
IIELPDGYMAALKRKCEARGMLLILDEAQTGVGRTGTMFACQRDGV—
TPDILTLSKTLGAGTSAAIEERAHELG

YLFYTTHVSDPLPAAVGLRVLDVVQRDGLVARANVMGDRLRRGLLDLMERF-
DCIGDVRGRGLLLGVEEPADGLG
AKITRECMNLGVQLPGMGG-VFRIAPPLTVSEDEIDLGLSLLGQAI----*
>P1;1s0aA
structure:1s0aA: 32: : 261: :::-1.00:-1.00
-----------------------------------------------------------------------
-----
----AEGCELILSDGRRLVDGMSSWWAAIHGYNHPQLNAAMKSQIDAMSHVMFGGITHAP----
AIELCRKLVAM
TPQPLECVFLADSGSVAVEVAMKMALQYWQAKGEARQRF--------------------
LTFRNGYHGDTFGAM
SVCDDNSMHSL------WKFAPAPQSR--MGEWDERDMVGFAR-------LMAAHRHE---
IAAVIIEPIQGAGG
MRMYHPEWLKRIRKICDREGILLIADEIATGFGRTGKLFACEH--------------------------
-----
-----------------------------------------------------------------------
-----
-------------------------------------------------*
```

```
>P1;2gsaA
structure:2gsaA: 38: : 338: :::-1.00:-1.00
-------------------------------------------------------------------------
-----
-FDRVKDAYAWDVDGNRYIDYVGTWGPAICGHAHPEVIEALKVAMEKGTSFGAPC----
ALENLAEMVNDAVPSI
E---MVRFVNSGTEACM---AVLRLMRAYTGRDK------------------------
IIKFEGCYHGHADMFL
VKAGS-GVATLGLPSS--PGVP-----------
KKTTANTLTTPYNDLEAVKALFAENPGEIAGVILEPIVGNSG
FIVPDAGFLEGLREITLEHDALLVFDEVMTGGGVQEKFGV--------
TPDLTTLGKGLPVGAYGGKREIAPAGP
MYQAGTLSGNPLAMTAGIKTLELLRQPGTYEYLDQITKRLSDGLL------------------------
-----
-------------------------------------------------*
>P1;1ohvA
structure:1ohvA: 1: : 461: :::-1.00:-1.00
-------------------------------------
FDYDGPLMKTEVPGPRSRELMKQLNIIQNAEAVHFFC
NYEESRGNYLVDVDGNRMLDLYSQISSIPIGYSHPALVKLVQQPQNVSTFINRPALGILPPENFVEKLRE
SLLSV
APKGMSQLITMACGSCSNENAFKTIFMWYRSKERGQSAFSKEELETCMINQAPGCPDYSILSFMGAFHGR
TMGCL
ATTHSKAIHKIDIPSFDWPIAPFPRLKYPLEEFVKENQQEEARCLEEVEDLIVKYRKKKKTVAGIIVEPI
QSEGG
DNHASDDFFRKLRDISRKHGCAFLVDEVQTGGGSTGKFWAHEHWGLDDPADVMTFSKKMMTGGFFHKEEF
RPNAP
YRIFNTWLGDPSKNLLLAEVINIIKREDLLSNAAHAGKVLLTGLLDLQARYPQFISRVRGRGTFCSFDTP
DESIR
NKLISIARNKGVMLGGCGDKSIRFRPTLVFRDHHAHLFLNIFSDILADF-*

>P1;1sffA

structure:1sffA: 36: : 424: :::-1.00:-1.00

---------------------------------------------------------------------

----------DVEGREYLDFAGGIAVLNTGHLHPKVVAAVEAQLKK---LSHTCFQVLAYEPYLELCEIMNQKV

PGDFAKKTLLVTTGSEAVENAVKI------ARAATKRS----------GTIAFSGAYHGRTHYTL

ALT-----GKVNPYSAGMGLMPVYRALYPCP--LHGISEDDA--IASIH-RIFKNDAAPEDIAAIVIEPVQGEGG

FYASSPAFMQRLRALCDEHGIMLIADEVQSGAGRTGTLFAMEQMGV--APDLTTFAKS-IAGGFGRAEVMDAVAP

GGLGGTYAGNPIACVAALEVLKVFEQENLLQKANDLGQKLKDGLLAIAEKHPE-IGDVRGLGAMIAIELFEDGDH

NKIVARARDKGLILLSCGPNVLRILVPLTIEDAQIRQGLEIISQCFDEAK*
```

**Compare.log (output for compare.py)**

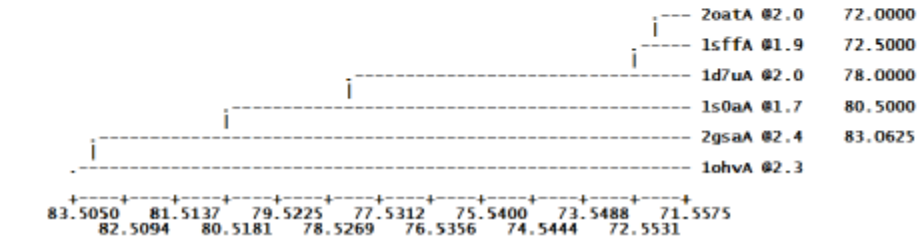
## Align2d.

```
Sequence identity comparison (ID_TABLE):

   Diagonal        ... number of residues;
   Upper triangle ... number of identical residues;
   Lower triangle ... % sequence identity, id/min(length).

          2oatA @21d7uA @21s0aA @12gsaA @21ohvA @21sffA @1
2oatA @2      404      108       93       84       76      112
1d7uA @2       27      431       86       76       86      117
1s0aA @1       23       20      427       79       72      107
2gsaA @2       21       18       19      427       63       97
1ohvA @2       19       20       17       15      461      102
1sffA @1       28       28       25       23       24      425

Weighted pair-group average clustering based on a distance matrix:
```




## Align2d.ali

```
>P1;1ohvA
structureX:1ohv.pdb: 11 :A:+461 :A:MOL_ID 1; MOLECULE 4-AMINOBUTYRATE
AMINOTRANSFERASE; CHAIN A, B, C, D; FRAGMENT RESIDUES 29-500; SYNONYM
GAMMA-AMINO-N-BUTYRATE TRANSAMINASE, GABA TRANSAMI GABA
AMINOTRANSFERASE, GABA-AT, GABA-T; EC 2.6.1.19:MOL_ID 1;
ORGANISM_SCIENTIFIC SUS SCROFA; ORGANISM_COMMON PIG; ORGANISM_TAXID
9823; ORGAN LIVER: 2.30:-1.00
-------------------------------------
FDYDGPLMKTEVPGPRSRELMKQLNIIQNAEAVHFFCNYEESRGNYLVDVDGNRMLDLYSQISSIPIGYS

HPALVKLVQQPQNVSTFINRPALGILPPENFVEKLRESLLSVAPKGMSQLITMACGSCSNENAFKTIFMW
YRSKERGQSAFSKEELETCMINQAPGCPDYSILSFMGAFHGRTMGCLATTHSKAIHKIDIPSFDWPIAPF
PRLKYPLEEFVKENQQEEARCLEEVEDLIVKYRKKKKTVAGIIVEPIQSEGGDNHASDDFFRKLRDISRK
HGCAFLVDEVQTGGGSTGKFWAHEHWGLDDPADVMTFSKKMMTGGFFHKEEFRPNAPYRIFNTWLGDPSK
NLLLAEVINIIKREDLLSNAAHAGKVLLTGLLDLQARYPQFISRVRGRGTFCSFDTPDESIRNKLISIAR
NKGVMLGGCGDKSIRFRPTLVFRDHHAHLFLNIFSDILADF-*

>P1;gabat
sequence:gabat: : : : ::: 0.00: 0.00
MASMLLAQRLACSFQHSYRLLVPGSRHISQAAAKVDVEFDYDGPLMKTEVPGPRSQELMKQLNIIQNAEA
VHFFCNYEESRGNYLVDVDGNRMLDLYSQISSVPIGYSHPALLKLIQQPQNASMFVNRPALGILPPENFV
EKLRQSLLSVAPKGMSQLITMACGSCSNENALKTIFMWYRSKERGQRGFSQEELETCMINQAPGCPDYSI
LSFMGAFHGRTMGCLATTHSKAIHKIDIPSFDWPIAPFPRLKYPLEEFVKENQQEEARCLEEVEDLIVKY
RKKKKTVAGIIVEPIQSEGGDNHASDDFFRKLRDIARKHGCAFLVDEVQTGGGCTGKFWAHEHWGLDDPA
DVMTFSKKMMTGGFFHKEEFRPNAPYRIFNTWLGDPSKNLLLAEVINIIKREDLLNNAAHAGKALLTGLL
DLQARYPQFISRVRGRGTFCSFDTPDDSIRNKLILIARNKGVVLGGCGDKSIRFRPTLVFRDHHAHLFLN
IFSDILADFK*
```

**Model-single.py (model generated gabat-1ohvA with dope score)**

```
<< end of ENERGY.
DOPE score               : -55550.527344
>> Model assessment by GA341 potential

Surface library          : C:\Program Files\Modeller9v7/modlib/surf5.de
Pair library             : C:\Program Files\Modeller9v7/modlib/pair9.de
Chain identifier         : _
% sequence identity      :   95.878998
Sequence length          :         500
Compactness              :    0.092349
Native energy (pair)     :  -563.688055
Native energy (surface)  :    -3.234556
Native energy (combined) :    -8.943275
Z score (pair)           :   -10.823216
Z score (surface)        :    -6.227564
Z score (combined)       :   -11.747523
GA341 score              :    1.000000


>> Summary of successfully produced models:
Filename                          molpdf     DOPE score    GA341 score
------------------------------------------------------------------------
gabat.B99990001.pdb           2768.29199   -55550.52734        1.00000
```

**Evaluate_template.py**

```
openf___224_> Open          1ohvA.profile
# Energy of each residue is written to: 1ohvA.profile
# The profile IS normalized by the number of restraints.
# The profiles are smoothed over a window of residues:   13
# The sum of all numbers in the file:     -17.5030

<< end of ENERGY.
DOPE score               : -56652.394531

Dynamically allocated memory at          finish [B,KiB,MiB]:     21326537   20826.695    20.339
Starting time                                               : 2011/11/19 23:00:14
Closing time                                                : 2011/11/19 23:00:29
Total CPU time [seconds]                                    :      15.56
```

**Evaluate_model.py**

```
openf___224_> Open          gabat.profile
# Energy of each residue is written to: gabat.profile
# The profile IS normalized by the number of restraints.
# The profiles are smoothed over a window of residues:   13
# The sum of all numbers in the file:     -18.5110

<< end of ENERGY.
DOPE score               : -55550.566406

Dynamically allocated memory at          finish [B,KiB,MiB]:     22408256   21883.062    21.370
Starting time                                               : 2011/11/19 23:02:54
Closing time                                                : 2011/11/19 23:03:10
Total CPU time [seconds]                                    :      16.23
```

Using **gabat** as query sequence and **1ohvA** as a template **"gabat.B99990001.pdb(gabat-1ohvA)"** structure was modeled using modeler with **dope score as -55550.52734.**