BI0505 BIOINFORMATICS- TECHNIQUES & APPLICATIONS

LAB MANUAL

Offered to

I YEAR M.TECH BIOINFORMATICS

DEPARTMENT OF BIOINFORMATICS

SCHOOL OF BIOENGINEERING
SRM UNIVERSITY

KATTANGULATHUR

Experiment 1: Biological Databases with Reference to Expasy and NCBI

Aim:

To view and use the various biological databases available on the World Wide Web.

Description:

Biological data is highly complex and interrelated. Vast amount of biological information needs to be stored organized and indexed so that the information can be retrieved and used. There are five major types of databases namely nucleotide databases, protein databases, protein structure databases, metabolic pathway databases and the bibliographic databases.

Procedure:

1. Open your web browser and type the web address of the required database.
2. Explore the database and analyze the various information available in the database.
3. Use the tools provided by the databases.
4. Save the output into a separate folder.

A. Expasy (Expert Protein analysis system):

Expasy is a proteomic server maintained by Swiss Institute of Bioinformatics for providing information on protein structures. The Database works in collaboration with European bioinformatics institute. Expasy is updated frequently with sequence information and tools for analyzing protein sequences.

Introduction:

Expasy can be reached by typing the URL www.expasy.org which is maintained by the Swiss institute of Bioinformatics. The website has a navigation column on the left side of the window, where the whole web server is categorized into various fields like proteomics, genomics, phylogeny , systems biology etc to create a better user experience. Each category is divided into two sections, Databases and tools. Since Expasy serves as the warehouse of many other databases and tools, all the available databases and tools available are characterized under these categories.

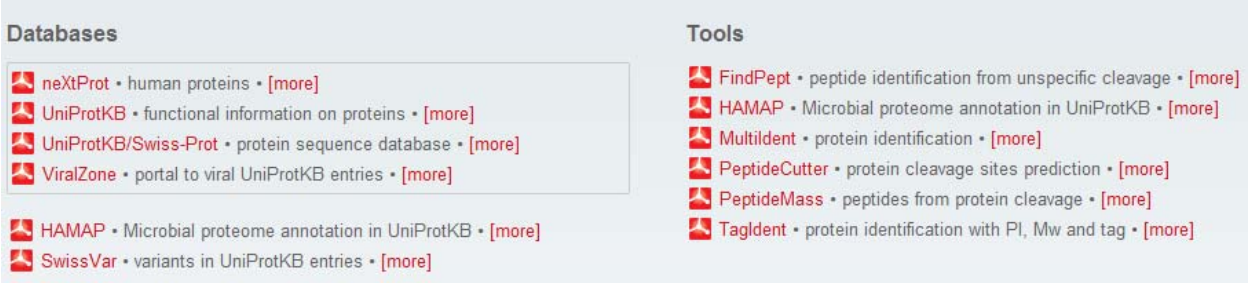
Search Tool and categories:

The home page is loaded with a query search tool where the user can search for biological information inside Expasy. A drop down menu is also provided in order to narrow down the search. The results will feature no of hits for the query with respect to each and every database based on which the user can direct himself to the location of information.

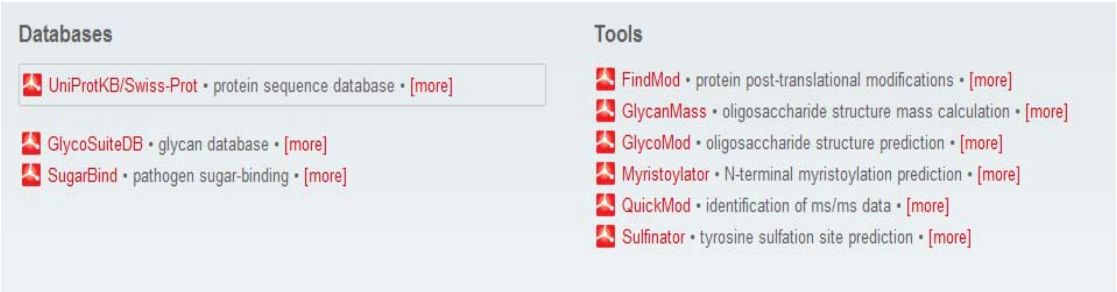
The Categories include

Proteomics

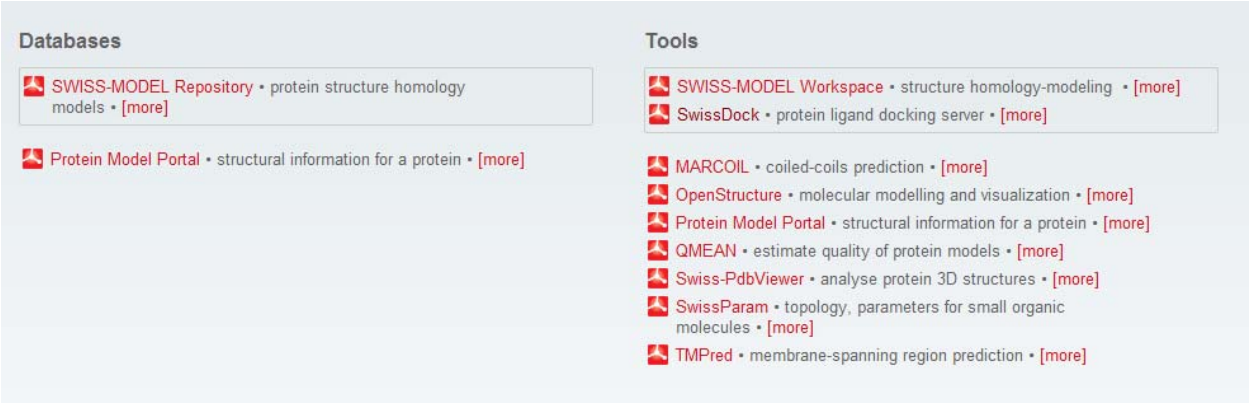
- Protein sequences and identification(Databases and tools involved in it)



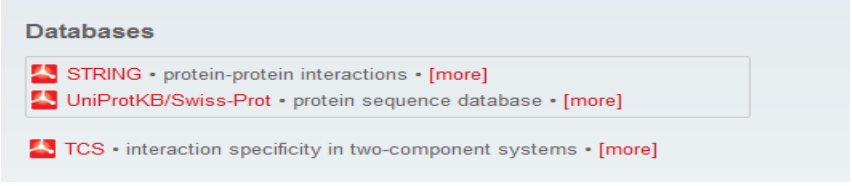
- Post translational modification



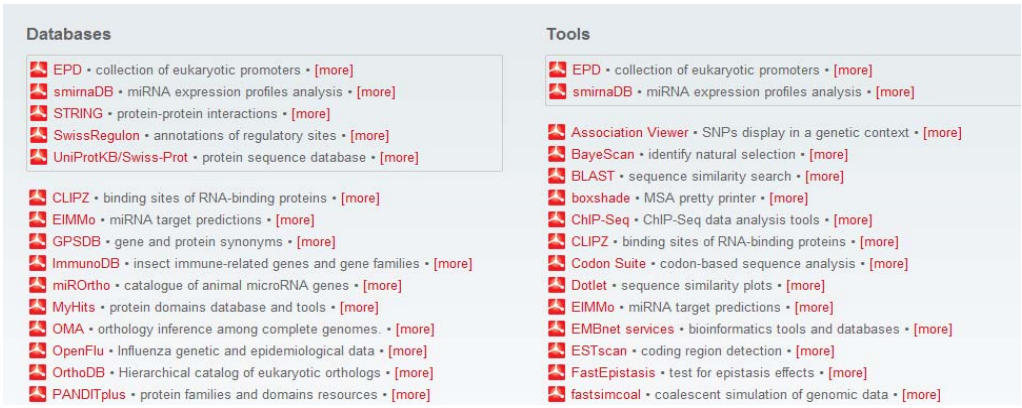
- Protein Structure



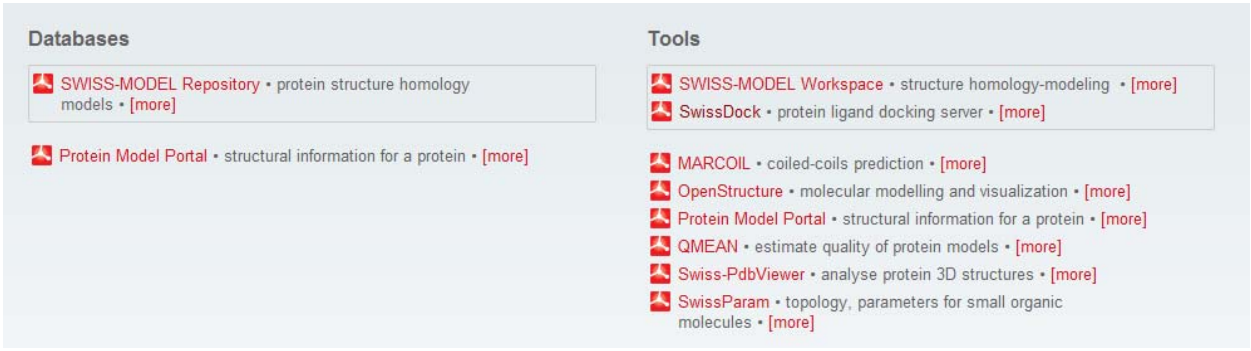
- Protein – Protein interaction



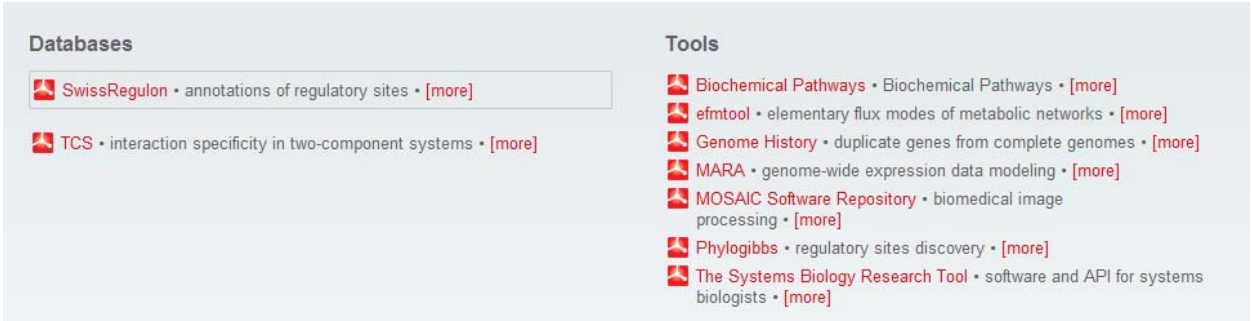
- Genomics



- Structural Bioinformatics



- System Biology



- Phylogeny/evolution

Databases

- Bgee • gene expression patterns comparison • [more]
- ImmunoDB • insect immune-related genes and gene families • [more]
- miROrtho • catalogue of animal microRNA genes • [more]
- OMA • orthology inference among complete genomes. • [more]
- OrthoDB • Hierarchical catalog of eukaryotic orthologs • [more]
- PANDITplus • protein families and domains resources • [more]
- Selectome • positive selection • [more]

Tools

- AllAll • protein sequences comparisons • [more]
- BayeScan • identify natural selection • [more]
- Codon Suite • codon-based sequence analysis • [more]
- CT-CBN • estimate conjunctive Bayesian networks • [more]
- fastsimcoal • coalescent simulation of genomic data • [more]
- Linear Classification • simple linear classification • [more]
- MLtree • maximum likelihood optimization • [more]
- MLTreeMap • phylogenetics and functionalities of metagenomes • [more]
- Newick Utilities • high-throughput phylogenetic tree processing • [more]
- OMA • orthology inference among complete genomes. • [more]
- Phylogenetic Tree • phylogenetic tree construction and printing • [more]
- RAxML • ML inference of large phylogenetic trees • [more]
- SuperTree • assemble phylogenetic trees • [more]
- TreeGen • phylogenetic tree from distance matrix • [more]
- Vertex Cover • resolves a vertex cover problem • [more]

Services :

Services etc.

- Downloads (specific databases, software tools etc.)
- Protein Spotlight • Protéines à la «Une»
- Swiss-Shop automatically obtain (by email) new sequence entries relevant to your field(s) of interest

SIB activities

- Educational activities at the SIB
- Service activities at the SIB

Old Proteomics pages

- Life Science Directory (databases, companies, journals etc.)
- Proteomics software tools

B. NCBI (National centre for Biotechnological information) :

NCBI is one of the leading online resources known for providing Biological sequence information. NCBI is maintained by two organizations in US ,National Library of Medicine (NLM) and National Institute of science (NIH). As a national resource for molecular biology information, NCBI's mission is to develop new information technologies to aid in the understanding of fundamental molecular and genetic processes that control health and disease. More specifically, the NCBI has been charged with creating automated systems for storing and analyzing knowledge about molecular biology, biochemistry, and genetics.

NCBI is connected to various other sequence databases in order to be more efficient in answering sequence queries. The user queries and sequence information are delivered through NCBI's search tool called the "entrez".

Home Page:

NCBI has a simplified homepage from where the user can navigate to different resources. The left side pane of the Homepage has a site map followed by different categories which narrows down the possibility of finding the right sequence. On the right side , you can see the list of popular resources which is very useful for first time users.

GenBank

The GenBank sequence database is an open access, annotated collection of all publicly available nucleotide sequences and their protein translations. This database is produced and maintained by the National Center for Biotechnology Information (NCBI) as part of the International Nucleotide Sequence Database Collaboration (INSDC). The National Center for Biotechnology Information is a part of the National Institutes of Health in the United States. GenBank and its collaborators receive sequences produced in laboratories throughout the world from more than 100,000 distinct organisms. In more than 20 years since its establishment, GenBank has become the most important and most influential database for research in almost all biological fields, whose data were accessed and cited by millions of researchers around the world. GenBank continues to grow at an exponential rate, doubling every 18 months.

Entrez:

The NCBI database accepts queries and delivers data via a custom made search engine called Entrez. The Home page of NCBI has a search box which directs the user to entrez. Entrez is internally connected to various biological databases which increases the probability of getting the correct information

BLAST:

BLAST stands for Basic Local Alignment Search Tool.BLAST is a tools that is used to find the seqyuences homologous to a particular sequence.BLAST compares all the sequences in the database with the one that is searched for and provides many hits which are usually arranged in the increasing order of the scored obtained

BLAST is available at the URLhttp://blast.ncbi.nlm.nih.gov/

BLAST uses PAM and BLOSUM matrices for scoring the alignment.

PubMed :

This is an online Bibliographic database which has a collection of the research papers, journals and other bibliographic data. The Database is internally connected with other Bibliographic databases like Medline, Biomedcentral etc.

Pubchem :

This contains data about the chemical compounds that are used for insillico analysis

Database of SNP's:

This database contains data about SNP's (Single Nucleotide polymorphism)

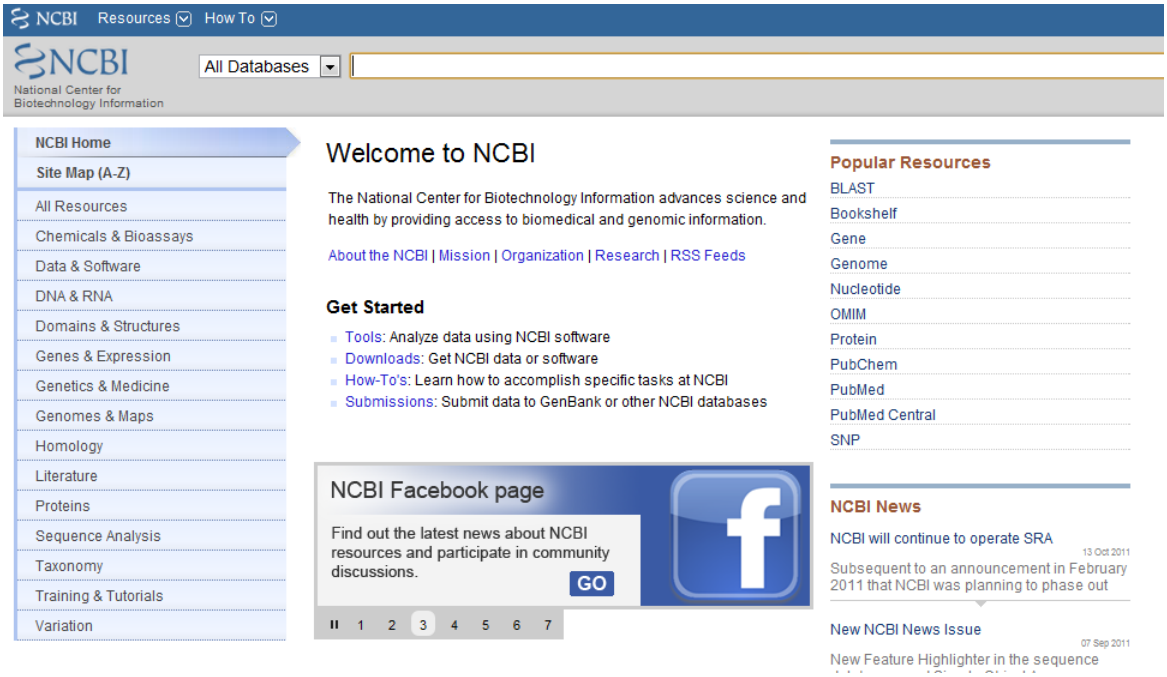
OMIM:

OMIM stand for Online Mendelian Inheritance in Man. This database contains information about the genetical disorders. OMIM gives complete data on the diseases the genetical background behind it and also the corresponding journal resources.

OMIA:

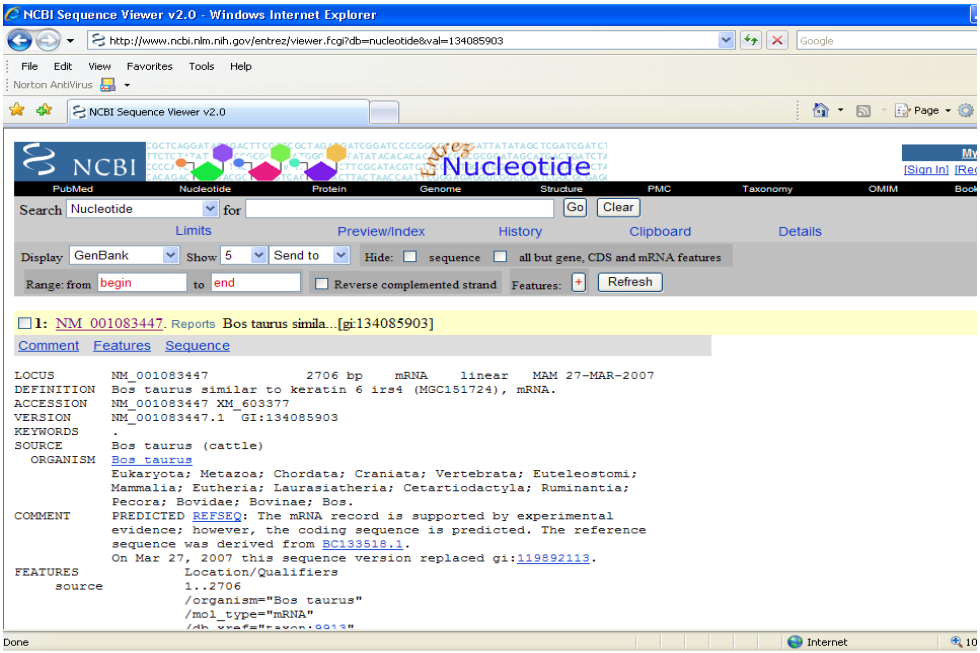
This database is similar to OMIM, but contains data about the diseases of all the other animals at the genetic level except human.

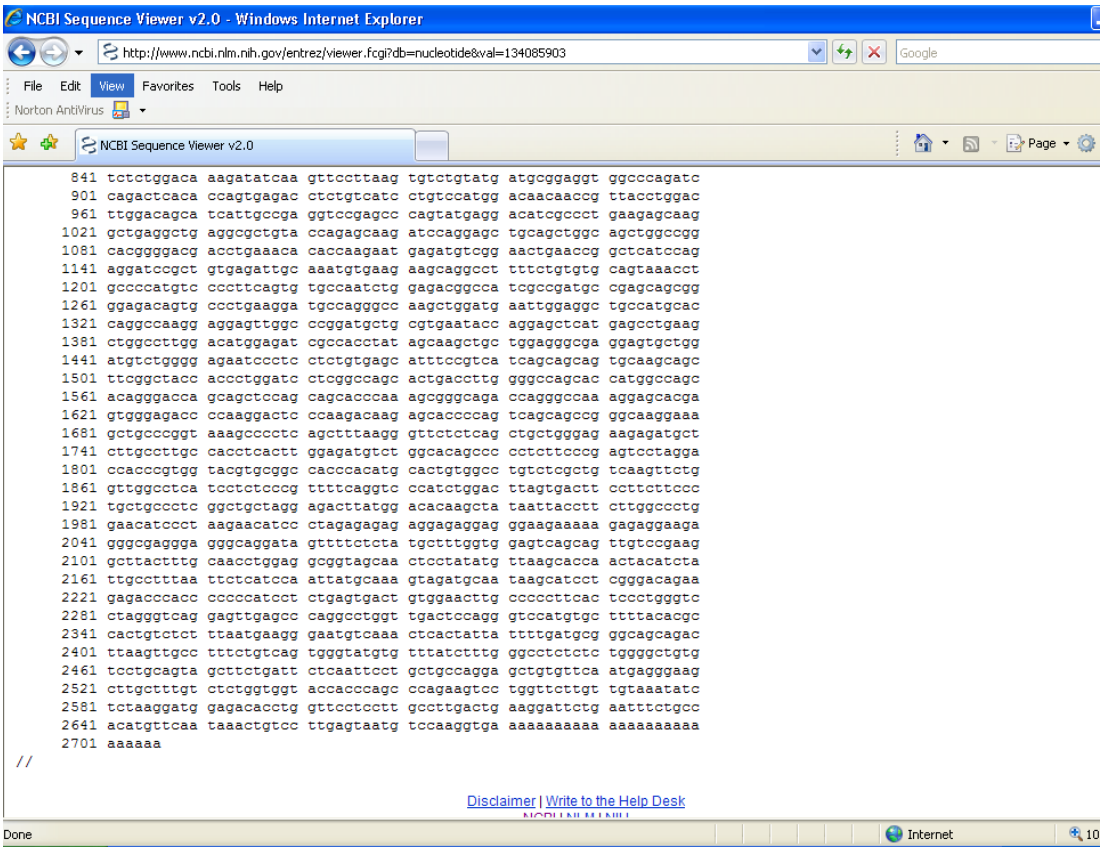
The home page of NCBI can be seen as follows:



Output:

The file format of the particular protein keratin can be shown follows:




NCBI Sequence Viewer v2.0 - Windows Internet Explorer
http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=134085903
File Edit View Favorites Tools Help
Norton AntiVirus
NCBI Sequence Viewer v2.0
Page
841 tctctggaca aagatatcaa gttccttaag tgtctgtatg atgcggaggt ggcccagatc
901 cagactcaca ccagtgagac ctctgtcatc ctgtccatgg acaacaaccg ttacctggac
961 ttggacagca tcattgccga ggtccgagcc cagtatgagg acatcgccct gaagagcaag
1021 gctgagggctg aggcgctgta ccagagcaag atccaggagc tgcagctggc agctggccgg
1081 cacggggacg acctgaaaca caccaagaat gagatgtcgg aactgaaccg gctcatccag
1141 aggatccgct gtgagattgc aaatgtgaag aagcaggcct tttctgtgtg cagtaaacct
1201 gccccatgtc cccttcagtg tgccaatctg gagacggcca tcgccgatgc cgagcagcgg
1261 ggagacagtg ccctgaagga tgccagggcc aagctggatg aattggaggc tgccatgcac
1321 caggccaagg aggagttggc ccggatgctg cgtgaatacc aggagctcat gagcctgaag
1381 ctggccttgg acatggagat cgccacctat agcaagctgc tggagggcga ggagtgctgg
1441 atgtctgggg agaatccctc ctctgtgagc atttccgtca tcagcagcag tgcaagcagc
1501 ttcggctacc accctggatc ctcggccagc actgaccttg gggccagcac catggccagc
1561 acagggacca gcagctccag cagcacccaa agcgggcaga ccaggggccaa aggagcacga
1621 gtgggagacc ccaaggactc ccaagacaag agcaccccag tcagcagccg ggcaaggaaa
1681 gctgcccggt aaagcccctc agctttaagg gttctctcag ctgctgggag aagagatgct
1741 cttgccttgc cacctcactt ggagatgtct ggcacagccc cctcttcccg agtcctagga
1801 ccacccgtgg tacgtgcggc cacccacatg cactgtggcc tgtctcgctg tcaagttctg
1861 gttggcctca tcctctcccg ttttcaggtc ccatctggac ttagtgactt ccttcttccc
1921 tgctgccctc ggctgctagg agacttatgg acacaagcta taattacctt cttggccctg
1981 gaacatccct aagaacatcc ctagagagag aggagaggag ggaagaaaaa gagaggaaga
2041 gggcgaggga gggcaggata gttttctcta tgctttggtg gagtcagcag ttgtccgaag
2101 gcttactttg caacctggag gcggtagcaa ctcctatatg ttaagcacca actacatcta
2161 ttgcctttaa ttctcatcca attatgcaaa gtagatgcaa taagcatcct cgggacagaa
2221 gagacccacc cccccatcct ctgagtgact gtggaacttg ccccccttcac tccctgggtc
2281 ctagggtcag gagttgagcc caggcctggt tgactccagg gtccatgtgc ttttacacgc
2341 cactgtctct ttaatgaagg gaatgtcaaa ctcactatta ttttgatgcg ggcagcagac
2401 ttaagttgcc tttctgtcag tgggtatgtg tttatctttg ggcctctctc tgggggctgtg
2461 tcctgcagta gcttctgatt ctcaattcct gctgccagga gctgtgttca atgagggaag
2521 cttgctttgt ctctggtggt accacccagc ccagaagtcc tggttcttgt tgtaaatatc
2581 tctaaggatg gagacacctg gttcctcctt gccttgactg aaggattctg aatttctgcc
2641 acatgttcaa taaactgtcc ttgagtaatg tccaaggtga aaaaaaaaaa aaaaaaaaaa
2701 aaaaaa
//
Disclaimer | Write to the Help Desk
Done
Internet
10

Experiment 2: 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 ofbioinformatics.

The simplicity of FASTA format makes it easy to manipulate and parse sequences using text-processing 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 :

ABL1_HUMAN

P00519, A3KFJ3, Q13869, Q13870, Q16133, Q17R61, Q45F09

Tyrosine-protein kinase ABL1

Organism: Homo sapiens

Function of Protein:

Non-receptor tyrosine-protein kinase that plays a role in many key processes linked to cell growth and survival such as cytoskeleton remodeling in response to extracellular stimuli, cell motility and adhesion, receptor endocytosis, autophagy, DNA damage response and apoptosis. Coordinates actin remodeling through tyrosine phosphorylation of proteins controlling cytoskeleton dynamics.

Catalytic Activity: ATP + a [protein]-L-tyrosine = ADP + a [protein]-L-tyrosine phosphate

Cofactor: Magnesium or manganese

Enzyme regulation: Stabilized in the inactive form by an association between the SH3 domain and the SH2-TK linker region, interactions of the N-terminal cap, and contributions from an N-terminal myristoyl group and phospholipids.

Protein attributes

Sequence length	1130 AA.
Sequence status	Complete.

FASTA format of sequence:

>sp|P00519|ABL1_HUMAN Tyrosine-protein kinase ABL1 OS=Homo sapiens GN=ABL1 PE=1 SV=4

MLEICLKLVGCKSKKGLSSSSSCYLEEALQRPVASDFEPQGLSEAARWNSKENLLAGPSE
NDPNLFVALYDFVASGDNTLSITKGEKLRVLGYNHNGEWCEAQTKNGQGWVPSNYITP
VNSLEKHSWYHGPVSRNAAEYLLSSGINGSFLVRESESSPGQRSISLRYEGRVYHYRINT
ASDGKLYVSSESRFNTLAELVHHHSTVADGLITTLHYPAPKRNKPTVYGVSPNYDKWE
MERTDITMKHKLGGGQYGEVYEGVWKKYSLTVAVKTLKEDTMEVEEFLKEAAVMKEI
KHPNLVQLLGVCTREPPFYIITEFMTYGNLLDYLRECNRQEVNAVVLLYMATQISSAME
YLEKKNFIHRDLAARNCLVGENHLVKVADFGLSRLMTGDTYTAHAGAKFPIKWTAPES
LAYNKFSIKSDVWAFGVLLWEIATYGMSPYPGIDLSQVYELLEKDYRMERPEGCPEKVY
ELMRACWQWNPSDRPSFAEIHQAFETMFQESSISDEVEKELGKQGVRGAVSTLLQAPEL
PTKTRTSRRAAEHRDTTDVPEMPHSKGQGESDPLDHEPAVSPLLPRKERGPPEGGLNED
ERLLPKDKKTNLFSALIKKKKKTAPTPPKRSSSFREMDGQPERRGAGEEEGRDISNGALA
FTPLDTADPAKSPKPSNGAGVPNGALRESGGSGFRSPHLWKKSSTLTSSRLATGEEEGG
GSSSKRFLRSCSASCVPHGAKDTEWRSVTLPRDLQSTGRQFDSSTFGGHKSEKPALPRKR
AGENRSDQVTRGTVTPPPRLVKKNEEAADEVFKDIMESSPGSSPPNLTPKPLRRQVTVAP
ASGLPHKEEAGKGSALGTPAAAEPVTPTSKAGSGAPGGTSKGPAEESRVRRHKHSSESP
GRDKGKLSRLKPAPPPPPAASAGKAGGKPSQSPSQEAAGEAVLGAKTKATSLVDAVNS
DAAKPSQPGEGLKKPVLPATPKPQSAKPSGTPISPAPVPSTLPSASSALAGDQPSSTAFIPL
ISTRVSLRKTRQPPERIASGAITKGVVLDSTEALCLAISRNSEQMASHSAVLEAGKNLYTF
CVSYVDSIQQMRNKFAFREAINKLENNLRELQICPATAGSGPAATQDFSKLLSSVKEISDI
VQR

2. **Write about PTM involved in P53355 and comment on the residues involved in it.**

Aim: To determine the Post Translational Modifications involved in **P53355** and to determine the residues involved in PTM.

Introduction:

Post-translational modification (PTM) is the chemical modification of a protein after its translation. It is one of the later steps in protein biosynthesis, and thus gene expression, for many proteins. After translation, the posttranslational modification of amino acids extends the range of functions of the protein by attaching it to other biochemical functional groups (such as acetate, phosphate, various lipids and carbohydrates), changing the chemical nature of an amino acid (e.g.citrullination), or making structural changes (e.g. formation of disulfide bridges).

Also, enzymes may remove amino acids from the amino end of the protein, or cut the peptide chain in the middle. Also, most nascent polypeptides start with the amino acid methionine because the "start" codon onmRNA also codes for this amino acid. This amino acid is usually taken off during post-translational modification.

Other modifications, like phosphorylation, are part of common mechanisms for controlling the behavior of a protein, for instance activating or inactivating an enzyme.

Method:

1. Open Uniprot Database *www.unitprot.org*
2. Enter the protein Id **P53355** in search tab and click on Find
3. Click on the protein name displayed on the result page.

Result and inference:

P53355 is a Death-associated protein kinase 1.
PTM involved is Phosphoprotein Ubl conjugation and PTM is Ubiquitinated by the BCR(KLHL20) E3 ubiquitin ligase complex, leading to its degradation by the proteasome.
The residues involved are :

- Phosphoprotein : Target amino acid is usually serine, threonine or tyrosine residues

Ubl conjugation : Protein which is posttranslationally modified by the attachment of at least one ubiquitin-like modifier protein, such as ubiquitin, SUMO, APG12, URM1 or RUB1. Ubiquitin, for example, is linked through a thioester bond between its C-terminus and the epsilon group of a lysine residue present on either another ubiquitin-like modifier protein or a target protein.

3. **Retrieve any one FASTA sequence of GABA transaminase in Human, mouse, pig and chick.**
Aim: To retrieve any one FASTA sequence of GABA transaminase in Human, mouse, pig and chick
Introduction:
4-aminobutyrate aminotransferase (or GABA transaminase) is an enzyme which catalyzes the conversion of 4-aminobutanoic acid (GABA) and 2-oxoglutarate into succinic semialdehyde and glutamate.

Method:

1. Open NCBI http://www.ncbi.nlm.nih.gov/
2. Choose the Protein Database and enter GABA transaminase in the search box
3. Click on Advanced Search Tab and Choose the Organism option from the drop down menu.
4. Enter Homo sapiens and the results are displayed.
5. The above steps can be repeated by entering the Organism name as Sus scrofa(Pig), Mus musculus(Mouse) and Gallus gallus(Chick).

Result and inference:

Human

>gi|188536080|ref|NP_001120920.1| 4-aminobutyrate aminotransferase, mitochondrial precursor [Homo sapiens]
MASMLLAQRLACSFQHSYRLLVPGSRHISQAAAKVDVEFDYDGPLMKTEVPGP
RSQELMKQLNIIQNAEA
VHFFCNYEESRGNYLVDVDGNRMLDLYSQISSVPIGYSHPALLKLIQQPQNASMF
VNRPALGILPPENFV
EKLRQSLLSVAPKGMSQLITMACGSCSNENALKTIFMWYRSKERGQRGFSQEEL
ETCMINQAPGCPDYSILSFMGAFHGRTMGCLATTHSKAIHKIDIPSFDWPIAPFPR
LKYPLEEFVKENQQEEARCLEEVEDLIVKYRKKKKTVAGIIVEPIQSEGGDNHAS
DDFFRKLRDIARKHGCAFLVDEVQTGGGCTGKFWAHEHWGLDDPADVMTFSK
KMMTGGFFHKEEFRPNAPYRIFNTWLGDPSKNLLLAEVINIIKREDLLNNAAHAG
KALLTGLLDLQARYPQFISRVRGRGTFCSFDTPDDSIRNKLILIARNKGVVLGGCG
DKSIRFRPTLVFRDHHAHLFLNIFSDILADFK

Pig

>gi|47523600|ref|NP_999428.1| 4-aminobutyrate aminotransferase, mitochondrial [Sus scrofa]
MASVLLTRRLACSFRHNHRLLVPGWRHISQAAAKVDVEFDYDGPLMKTEVPGP
RSRELMKQLNIIQNAEAVHFFCNYEESRGNYLVDVDGNRMLDLYSQISSIPIGYS
HPALVKLVQQPQNVSTFINRPALGILPPENFVEKLRESLLSVAPKGMSQLITMAC
GSCSNENAFKTIFMWYRSKERGQSAFSKEELETCMINQAPGCPDYSILSFMGAFH
GRTMGCLATTHSKAIHKIDIPSFDWPIAPFPRLKYPLEEFVKENQQEEARCLEEVE
DLIVKYRKKKKTVAGIIVEPIQSEGGDNHASDDFFRKLRDISRKHGCAFLVDEVQ
TGGGSTGKFWAHEHWGLDDPADVMTFSKKMMTGGFFHKEEFRPNAPYRIFNT
WLGDPSKNLLLAEVINIIKREDLLSNAAHAGKVLLTGLLDLQARYPQFISRVRGR
GTFCSFDTPDESIRNKLNSIARNKGVMLGGCGDKSIRFRPTLVFRDHHAHLFLN
IFSDILADFKDLQARYPQFISRVRGRGTFCSFDTPDESIRNKLISIARNKGVMLGGC
GDKSIRFRPTLVFRDHHAHLFLNIFSDILADFK

Mouse

>gi|283483966|ref|NP_001164449.1| 4-aminobutyrate aminotransferase, mitochondrial isoform 2 precursor [Mus musculus]
MAFLLITRRLACSSQKNLHLFIPGSRYISQAAAKVDIEFDYDGPLMKTEVPGPRSK
ELMKQLNTIQNAEAVHFFCNYEESRGNYLVDVDGNRMLDLYSQISSVPIGYNHP
ALAKLVQQPQNASTFINRPALGILPPENFVDKLQESLMSVAPRGMSQLITMACGS
CSNENAFKTIFMWYRSKERGQRGFSKEELETCMVNQSPGCPDYSILSFMGAFHG
RTMGCLATTHSKAIHKIDIPSFDWPIAPFPRLKYPLEEFTTDNQQEEARCLEEVED
LIVKYRKKKRTVAGIIVEPIQSEGGDNHASDDFFRKLRDIARKPYRIFNTWLGDPS
KNLLLAEVINIIKREDLLNNVARVGKTLLTGLLDLQAQYPQFISRVRGRGTFCSFD
TPDEAIRNKLILIARNKGVVLGGCGDKSIRFRPTLVFRDHHAHLFLSIFSGILADFK

Chick

It is absent in chick.

4. **Find out the number of entries in SWISSPROT for Serine kinase in PIG.**
   Aim: To determine the number of entries in SWISSPROT for Serine kinase in PIG.
   Introduction:
   Serine/threonine protein kinases (EC 2.7.11.1) phosphorylate the OH group of serine or threonine (which have similar sidechains). Serine/Threonine Kinase receptors plays a role in the regulation of cell proliferation, programmed cell death (apoptosis), cell differentiation, and embryonic development.
   Method:
   1. Open the following url- http://www.ebi.ac.uk/uniprot/
   2. Enter the query as follows:
      Serine kinase AND organism:"Sus scrofa [9823]"

   Result and inference:

   Number of entries = 325

5. **Comment on the secondary structure information about P68871 AND P24071.**
   Aim: To determine the Secondary structure of P68871 AND P24071
   Introduction:
   Proteins are an important class of biological macromolecules present in all organisms. Proteins are polymers of amino acids.
   Levels of Protein Structure-
   The primary structure refers to amino acid sequence of the polypeptide chain. The primary structure is held together by covalent or peptide bonds, which are made during the process of protein biosynthesis or translation. The two ends of the polypeptide chain are referred to as the carboxyl terminus (C-terminus) and the amino terminus (N-terminus) based on the nature of the free group on each extremity.
   Secondary structure refers to highly regular local sub-structures. Two main types of secondary structure, the alpha helix and the beta strand. These secondary structures are defined by patterns of hydrogen bonds between the main-chain peptide groups. They have a regular geometry, being constrained to specific values of the dihedral angles ψ and φ on the Ramachandran plot. Both the alpha helix and the beta-sheet represent a way of saturating all the hydrogen bond donors and acceptors in the peptide backbone. Some parts of the protein are ordered but do not form any regular structures. They should not be confused with random coil, an unfolded polypeptide chain lacking any fixed three-dimensional structure. Several sequential secondary structures may form a "supersecondary unit".
   Tertiary structure refers to three-dimensional structure of a single protein molecule. The alpha-helices and beta-sheets are folded into a compact globule. The folding is driven by the *non-specific* hydrophobic interactions (the burial of hydrophobic residues from

water), but the structure is stable only when the parts of a protein domain are locked into place by *specific* tertiary interactions, such as salt bridges, hydrogen bonds, and the tight packing of side chains and disulfide bonds.
Quaternary structure is a larger assembly of several protein molecules or polypeptide chains, usually called subunits in this context. The quaternary structure is stabilized by the same non-covalent interactions and disulfide bonds as the tertiary structure.
Complexes of two or more polypeptides (i.e. multiple subunits) are called multimers. Specifically it would be called a dimer if it contains two subunits, a trimer if it contains three sub-units, and a tetramer if it contains four subunits. The subunits are frequently related to one another by symmetry operations, such as a 2-fold axis in a dimer.
Multimers made up of identical subunits are referred to with a prefix of "homo-" (e.g. a homotetramer) and those made up of different subunits are referred to with a prefix of "hetero-" (e.g. a heterotetramer, such as the two alpha and two beta chains of hemoglobin).

Method:

1. Open Uniprot http://www.uniprot.org/
2. Enter the protein ID.

Result and inference:

P68871: Hemoglobin subunit beta

Gene Name: HBB

Organism: Homo sapiens

Sequence length: 147aa

Function: Involved in oxygen transport from the lung to the various peripheral tissues and is specific to the red blood cells.

P68871 is hemoglobin subunit beta with helix turn helix conformation.

Helix	6 – 16	11
Turn	21 – 23	3
Helix	24 – 35	12
Helix	37 – 42	6
Helix	44 – 46	3
Helix	52 – 57	6
Helix	59 – 77	19
Turn	78 – 80	3
Helix	82 – 94	13
Turn	95 – 97	3
Helix	102 – 119	18
Helix	120 – 122	3
Helix	125 – 142	18
Helix	144 – 146	3

Secondary structure of Hemoglobin subunit beta is rich in alpha helices.

P24071: Immunoglobulin alpha Fc receptor

Gene Name: FCAR

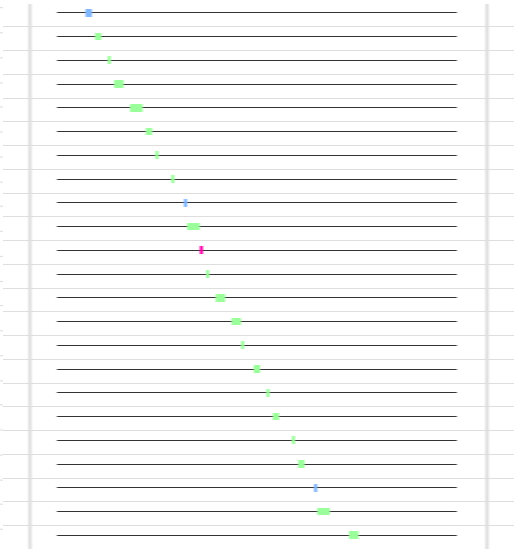
Organism: Homo sapiens

Sequence length: 287aa

Function: Binds to the Fc region of immunoglobulins alpha. Mediates several functions including cytokine production

P24071 is Immunoglobulin alpha Fc receptor with helix strand turn conformation..its rich in beta sheet

Helix	23 – 25	3
Beta strand	31 – 35	5
Beta strand	37 – 40	4
Beta strand	45 – 49	5
Beta strand	56 – 64	9
Beta strand	67 – 70	4
Beta strand	72 – 74	3
Beta strand	84 – 87	4
Helix	92 – 94	3
Beta strand	96 – 104	9
Turn	105 – 107	3
Beta strand	108 – 111	4
Beta strand	115 – 120	6
Beta strand	127 – 132	6
Beta strand	134 – 136	3
Beta strand	143 – 147	5
Beta strand	149 – 151	3
Beta strand	154 – 160	7
Beta strand	168 – 171	4
Beta strand	173 – 179	7
Helix	185 – 187	3
Beta strand	189 – 196	8
Beta strand	211 – 215	5



6. Find the approximate region on 2D gel where **Q8N423** is found.

Aim: To find the approximate region on 2D gel where **Q8N423** is found

Introduction: **Compute pI / Mw Tool**

This tool calculates the estimated pI and Mw of a specified Swiss-Prot/TrEMBL entry or a user-entered AA sequence. These parameters are useful if you want to know the approximate region of a 2-D gel where a protein may be found. Protein pI is calculated using pK values of amino acids. Prediction of protein pI for highly basic proteins is yet to be studied and it is possible that current Compute pI/Mw predictions may not be adequate for this purpose.

The buffer capacity of a protein will affect the accuracy of its predicted pI, with poor buffer capacity leading to greater error in prediction. Because of this, pI predictions for small proteins can be problematic.

Protein Mw is calculated by the addition of average isotopic masses of amino acids in the protein and the average isotopic mass of one water molecule.

This program does not account for the effects of post-translational modifications, thus modified proteins on a 2-D gel may migrate to a position quite different to that predicted. Protein glycosylation in particular can affect protein migration in both pI and Mw dimensions. In addition to the standard one-letter-codes for the 20 amino acids, the 2 non-standard amino acids (Selenocysteine and Pyrrolysine), the characters B, Z and X are accepted:

B Asx Aspartic acid or Asparagine

Z Glx Glutamine or Glutamic acid

X Xaa any amino acid

Method:

1. Open the Compute pI/Mw tool http://web.expasy.org/compute_pi/
2. Enter accession numbers (Q8N423) into the text field, and select the "click here to compute pI/Mw" button.
3. To calculate the Isoelectric point of the entire protein enters the N-Terminal as 1 and C-Terminal as 598 since the length of the protein query is 598 amino acids.

Result and inference:

LIRB2_HUMAN (Q8N423)

Leukocyte immunoglobulin-like receptor subfamily B member 2 precursor (LIR-2) (Leukocyte immunoglobulin-like receptor 2) (CD85 antigen-like family member D) (Immunoglobulin-like transcript 4) (ILT-4) (Monocyte/macrophage immunoglobulin-like receptor 10) (MIR-10) (CD85d antigen)
Organism: Homo sapiens (Human)

Sequence Length: 598 residues

Molecular weight (average): 65038.65

Theoretical pI: 6.79

6. Find the presence of super secondary structure if any in **Q9H6F5.**
   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 86
   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

8. Find the number of proteins which are having isoelectric point value between 5 and 5.5. Comment on the result.
   Aim: To find the number of proteins which are having isoelectric point value between 5 and 5. 5

   Introduction:
   TagIdent is a tool which allows the generation of a list of proteins close to a given pI and Mw, the identification of proteins by matching a short sequence tag of up to 6 amino acids against proteins in the UniProt Knowledgebase (Swiss-Prot and TrEMBL) databases close to a given pI and Mw and the identification of proteins by their mass, if this mass has been determined by mass spectrometric techniques for one or more species and with an optional keyword. When searching in UniProtKB/Swiss-Prot, TagIdent removes signal sequences and/or propeptides (as documented in the UniProtKB/Swiss-

Prot feature table (FT lines)) before computing pI and Mw for each of the resulting chains.
The annotation in UniProtKB/TrEMBL is done automatically; it is incomplete and not always correct. Thus information on UniProtKB/TrEMBL FT lines is not used to process UniProtKB/TrEMBL proteins into mature chains or peptides (i.e. pI and Mw are always computed for the whole sequence), and the use of a keyword is not allowed for searches in UniProtKB/TrEMBL.

Method:

1. Open the TagIdent tool http://web.expasy.org/tagident/
2. Enter the Query name (Optional), the pI-range, the Mw in Daltons.
3. Enter the Mw in percent as 20% (default)
4. Check for proteins with with cysteines in reduced form (-SH)
5. Choose database as Uniprot/SwissProt
6. Click on Start TagIdent

Result and inference:

TagIdent tool is a Identify proteins with isoelectric point (pI), molecular weight (Mw) and sequence tag, or generate a list of proteins close to a given pI and Mw.
And number of proteins found in the specified pI/Mw ranges(5 and 5.5) are **96933**

9. Find the disease with which **PPE protein is involved?**

Aim: To find the disease with which the PPE protein is involved.

Introduction:

The PE/PPE protein has been widely speculated that these proteins may play a role in evasion of host immune responses, possibly via antigenic variation. Emerging data increasingly supports a role for the PE/PPE proteins at multiple levels of the infectious process.

Method:

1. Open Uniprot http://www.uniprot.org/
2. Enter PPE protein in the Query box.

Result and inference:

Tuberculosis involves PPE protein

10. **Write about** NAGK gene present in Homo sapiens**.**

Aim: To retrieve information about NAGK gene present in Homo sapiens

Introduction:

UniProt is to provide the scientific community with a comprehensive, high quality and freely accessible resource of protein sequence and functional information. UniProt is comprised of four components, each optimised for different uses. The **UniProt Knowledgebase (UniProtKB)** is the central access point for extensive curated protein information, including function, classification, and cross-reference. It consists of two sections: **UniProtKB/Swiss-Prot** which is manually annotated and is reviewed and **UniProtKB/TrEMBL** which is automatically annotated and is not reviewed. The **UniProt Reference Clusters (UniRef)** databases provide clustered sets of sequences from the UniProtKB and selected UniProt Archive records to obtain complete coverage of sequence space at several resolutions while hiding redundant sequences. The **UniProt Archive (UniParc)** is a comprehensive repository, used to keep track of sequences and their identifiers. The **UniProt Metagenomic and Environmental Sequences (UniMES)**database is a repository specifically developed for metagenomic and environmental data.

Method:

1. Open Uniprot http://www.uniprot.org/
2. Enter PPE protein in the Query box.

Result and inference:

NAGK: N-acetylglucosamine kinase [Homo *sapiens*]

N-acetylglucosamine kinase (NAGK; EC 2.7.1.59) converts endogenous N-acetylglucosamine (GlcNAc), a major component of complex carbohydrates, from lysosomal degradation or nutritional sources into GlcNAc 6-phosphate. NAGK belongs to the group of N-acetylhexosamine kinases and is a prominent salvage enzyme of amino sugar metabolism in mammals.30 PubMed Neighbors

Experiment 3: Sequence similarity searching using 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**

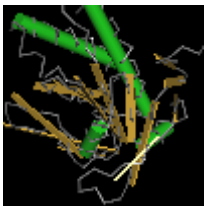
Introduction:

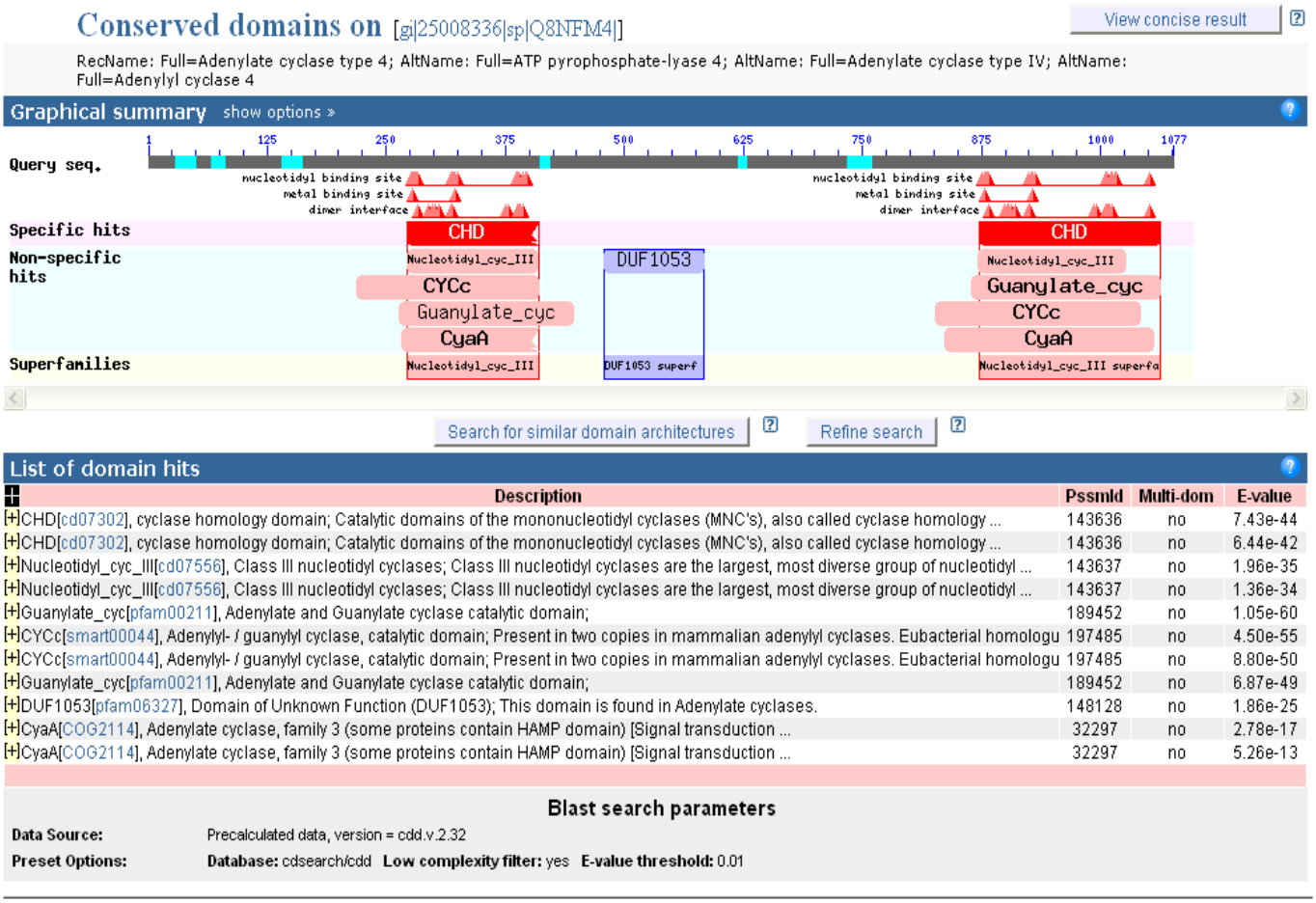
Conserved domains (CD) in proteins play a crucial role in protein interactions, DNA binding, enzyme activity, and other important cellular processes. With recently released gene number predictions in the human genome being less than many previous predictions, interactions among these domains may prove 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.

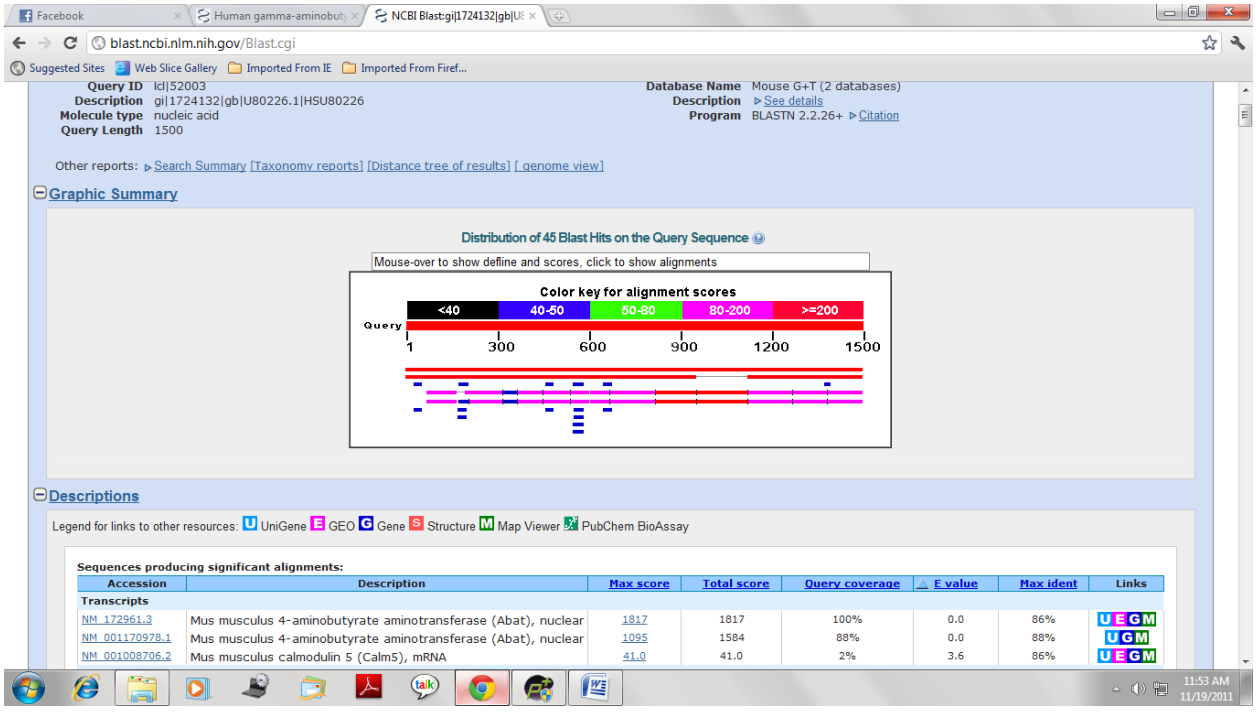
Introduction:

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.

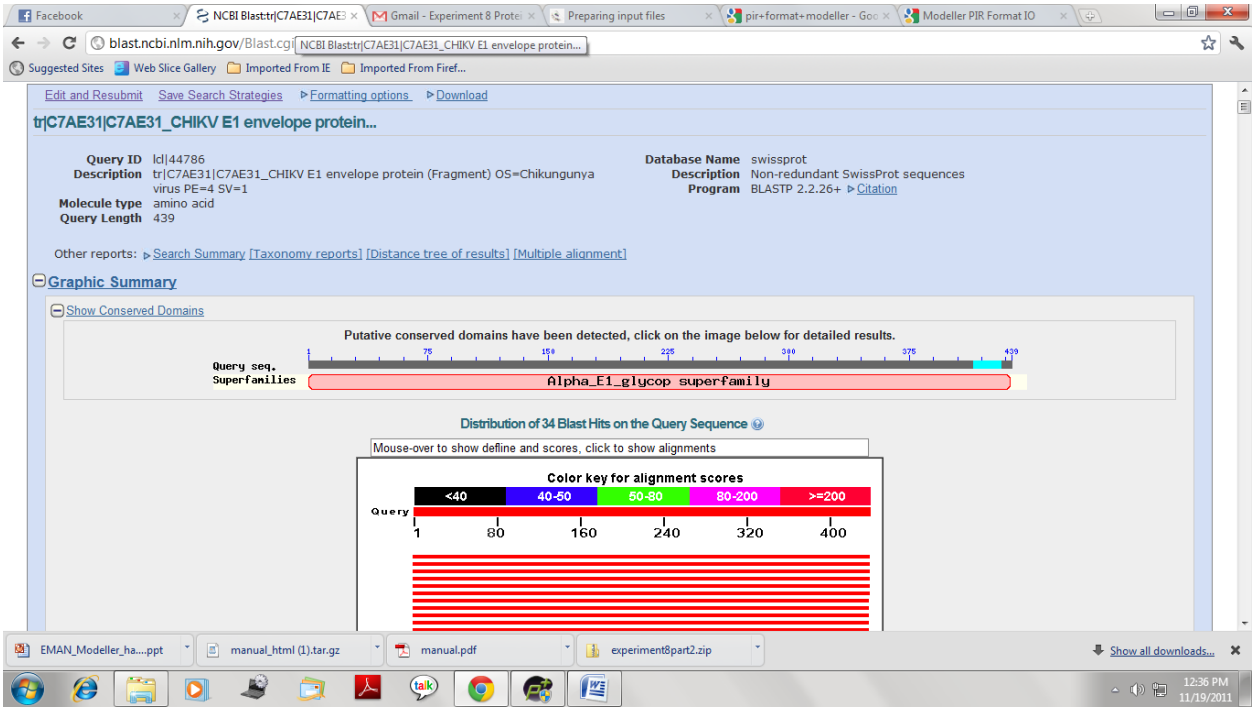
Introduction:

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:

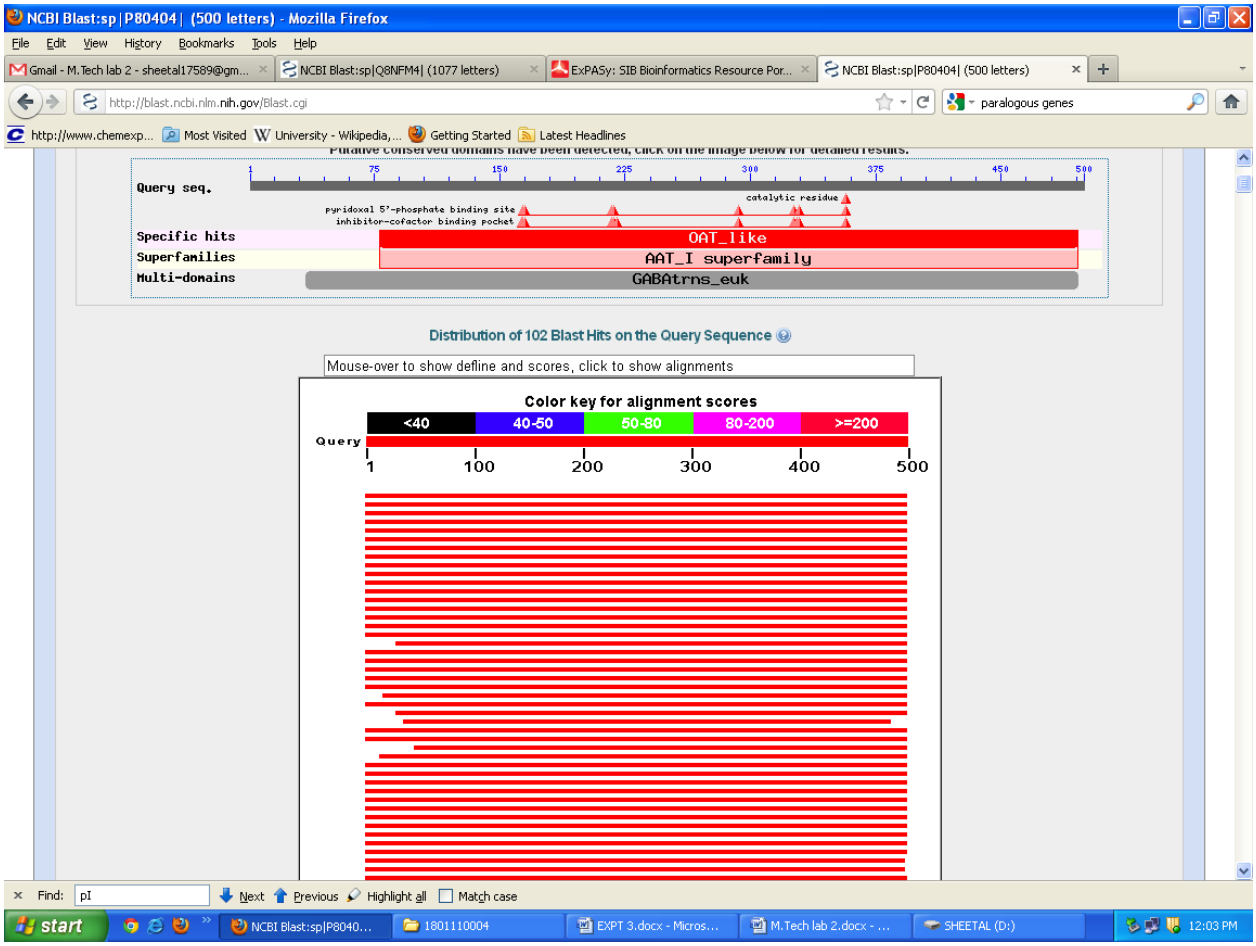
Paralogous means genes that have arisen from a common ancestor and are present in the same genome. Paralgous may or may not have the same function. Paralogous proteins are proteins that have arisen by gene duplication. The group of paralogous proteins that are descended from a common ancestor by gene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

5. Find whether the given pattern is present in the following protein. Also find its homologous proteins present in SWISPROT database possessing the similar pattern.

Aim: To find whether the given pattern is present in the following protein. Also to find its homologous proteins present in SWISPROT database possessing the similar pattern.

Introduction:

By filling in the "regular expression" box on the PSI-blast page, you can execute a PHI-blast search. PHI-blast enforces the presence of a motif in addition to the usual PSI-blast criteria for matching. Regular expressions can be used to confine the results to a formally defined family. The syntax for patterns in PHI-BLAST follows the conventions of PROSITE.

Pattern:

[LIVMFYWCS]-[LIVMFYWCAH]-x-D-[ED]-[IVA]-x(2,3)-[GAT]-[LIVMFAGCYN]-x(0,1)-[RSACLIH]-x-[GSADEHRM]-x(10,16)-[DH]-[LIVMFCAG]-[LIVMFYSTAR]-x(2)-[GSA]-K-x(2,3)-[GSTADNV]-[GSAC]

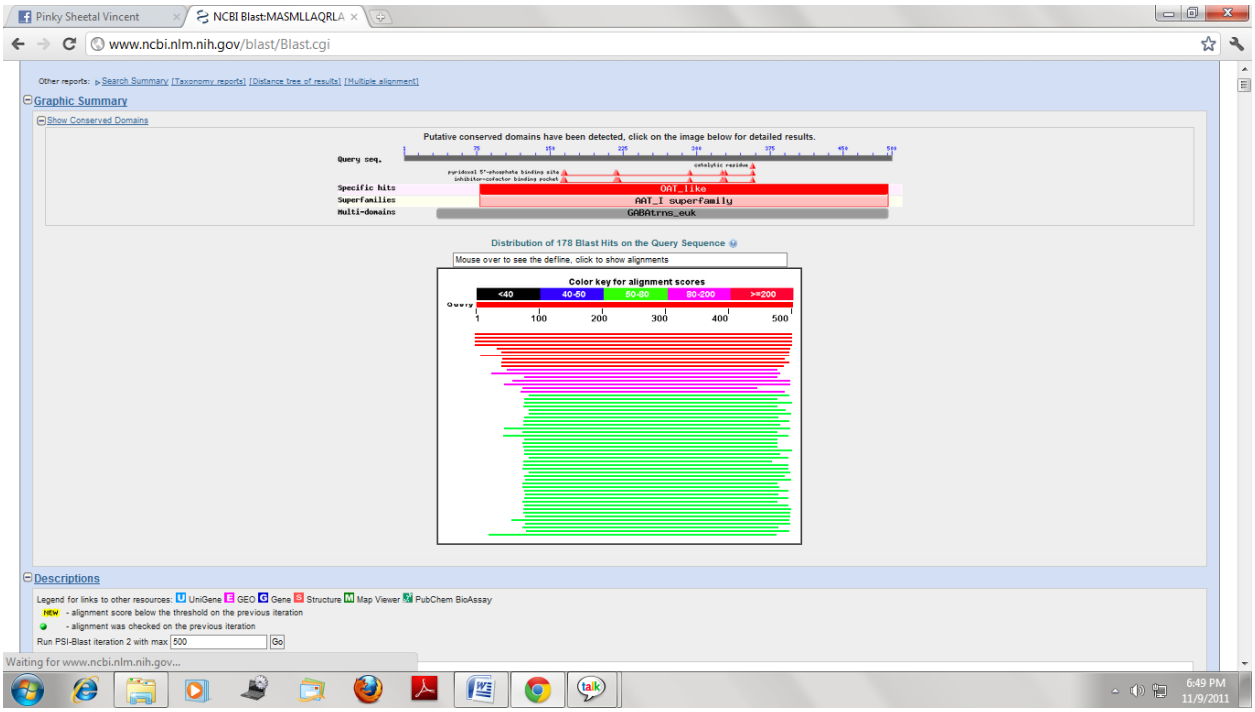
Protein:

>
MASMLLAQRLACSFQHSYRLLVPGSRHISQAAAKVDVEFDYDGPLMKTEVPGPRSQELMK
QLNIIQNAEAVHFFCNYEESRGNYLVDVDGNRMLDLYSQISSVPIGYSHPALLKLIQQPQ
NASMFVNRPALGILPPENFVEKLRQSLLSVAPKGMSQLITMACGSCSNENALKTIFMWYR
SKERGQRGFSQEELETCMINQAPGCPDYSILSFMGAFHGRTMGCLATTHSKAIHKIDIPS
FDWPIAPFPRLKYPLEEFVKENQQEEARCLEEVEDLIVKYRKKKKTVAGIIVEPIQSEGG
DNHASDDFFRKLRDIARKHGCAFLVDEVQTGGGCTGKFWAHEHWGLDDPADVMTFSKKMM
TGGFFHKEEFRPNAPYRIFNTWLGDPSKNLLLAEVINIIKREDLLNNAAHAGKALLTGLL
DLQARYPQFISRVRGRGTFCSFDTPDDSIRNKLILIARNKGVVLGGCGDKSIRFRPTLVF
RDHHAHLFLNIFSDILADFK

Method:

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

PHI-BLAST Results:



OAT_like[cd00610], Acetyl ornithine aminotransferase family. This family belongs to pyridoxal phosphate (PLP)

6. Identify the given sequence and also find the similar sequences present in SWISSPROT database for the following query.

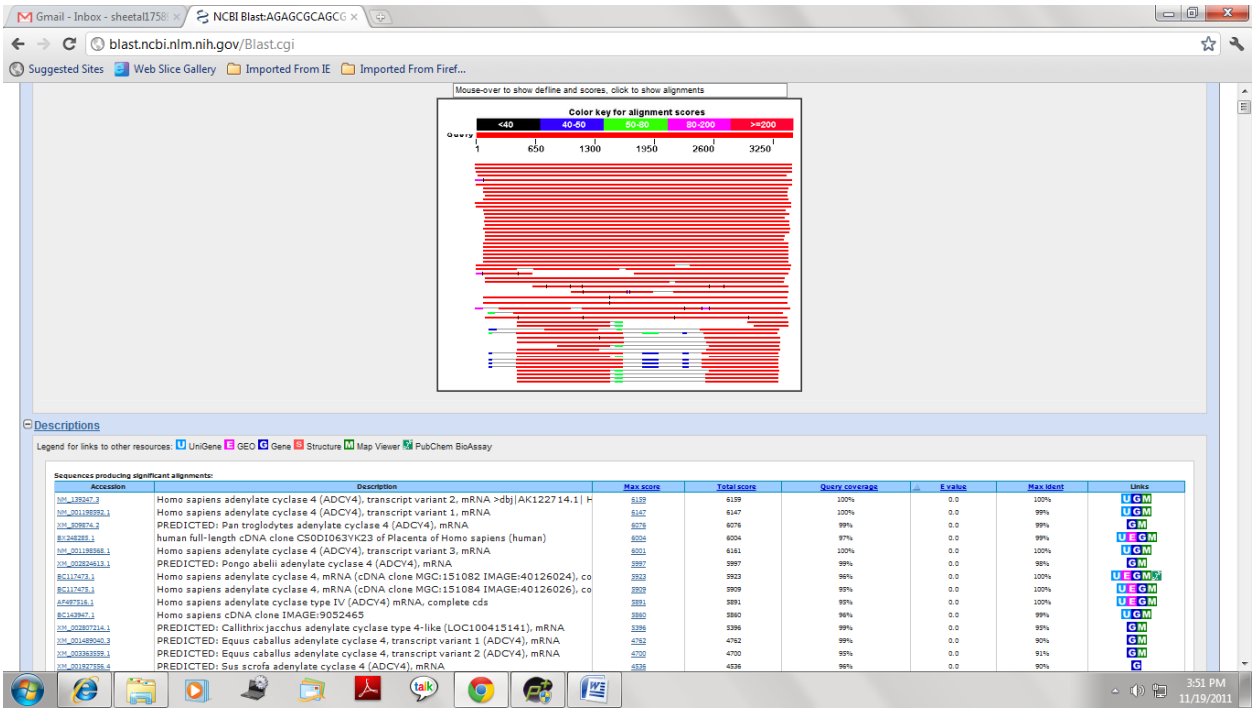
>
AGAGCGCAGCGGCGAGCGTGACTCCGCCATCAGGTCCCCGGCTCCCTCCCCGGACCTAGCCCACTCCGCT
GCGCCAGCGCCGCGGGCACCCCGGGGCTCGGGGCTGGGGAGATCATGGCCCGCCTCTTCAGCCCCCGGCC
GCCCCCCAGCGAAGACCTCTTCTACGAGACCTACTACAGCCTGAGCCAGCAGTACCCGCTGCTGCTGCTG
CTGCTGGGGATCGTGCTCTGTGCGCTCGCGGCGCTGCTCGCAGTGGCCTGGGCCAGCGGCAGGGAGCTGA
CCTCAGACCCGAGCTTCCTGACCACTGTGCTGTGCGCGCTGGGCGGCTTCTCGCTGCTGCTGGGCCTCGC
TTCCCGGGAGCAGCGACTGCAGCGCTGGACGCGTCCCCTGTCCGGCTTGGTATGGGTCGCGCTGCTAGCG
CTAGGCCACGCCTTCCTGTTCACCGGGGGCGTGGTGAGCGCCTGGGACCAGGTGTCCTATTTTCTCTTCG
TCATCTTCACGGCGTATGCCATGCTGCCCTTGGGCATGCGGGACGCCGCCGTCGCGGGCCTCGCCTCCTC
ACTCTCGCATCTGCTGGTCCTCGGGCTGTATCTTGGGCCACAGCCGGACTCACGGCCTGCACTGCTGCCG
CAGTTGGCAGCAAACGCAGTGCTGTTCCTGTGCGGGAACGTGGCAGGAGTGTACCACAAGGCGCTGATGG
AGCGCGCCCTGCGGGCCACGTTCCGGGAGGCACTCAGCTCCCTGCACTCACGCCGGCGGCTGGACACCGA
GAAGAAGCACCAGGAACACCTTCTCTTGTCCATCCTTCCTGCCTACCTGGCCCGAGAGATGAAGGCAGAG
ATCATGGCACGGCTGCAGGCAGGACAGGGGTCACGGCCAGAGAGCACTAACAATTTCCACAGCCTCTATG
TCAAGAGGCACCAGGGAGTCAGCGTGCTGTATGCTGACATCGTGGGCTTCACGCGGCTGGCCAGCGAGTG
TTCCCCTAAGGAGCTGGTGCTCATGCTCAATGAGCTCTTTGGCAAGTTCGACCAGATTGCCAAGGAGCAT
GAATGCATGCGGATCAAGATCCTGGGGGACTGTTACTACTGTGTCTCTGGGCTGCCACTCTCACTGCCAG
ACCATGCCATCAACTGCGTGCGCATGGGCCTGGACATGTGCCGGGCCATCAGGAAACTGCGGGCAGCCAC
TGGCGTGGACATCAACATGCGTGTGGGCGTGCACTCAGGCAGCGTACTGTGTGGAGTCATCGGGCTGCAG
AAGTGGCAGTACGACGTTTGGTCACATGATGTCACACTGGCTAACCACATGGAGGCAGGCGGTGTACCAG
GGCGAGTGCACATCACAGGGGCTACCCTGGCCCTGCTGGCAGGGGCTTATGCTGTGGAGGACGCAGGCAT
GGAGCATCGGGACCCCTACCTTCGGGAGCTAGGGGAGCCTACCTATCTGGTCATCGATCCACGGGCAGAG
GAGGAGGATGAGAAGGGCACTGCAGGAGGCTTGCTGTCCTCGCTTGAGGGCCTCAAGATGCGTCCATCAC
TGCTGATGACCCGTTACCTGGAGTCCTGGGGCGCAGCCAAGCCTTTTGCCCACCTGAGCCACGGAGACAG
CCCTGTGTCCACCTCCACCCCTCTCCCGGAGAAGACCCTGGCTTCCTTCAGCACCCAGTGGAGCCTGGAT
CGGAGCCGTACCCCCCGGGGACTAGATGATGAACTGGACACCGGGGATGCCAAGTTCTTCCAGGTCATTG
AGCAGCTCAACTCGCAGAAACAGTGGAAGCAGTCGAAGGACTTCAACCCACTGACACTGTACTTCAGAGA
GAAGGAGATGGAGAAAGAGTACCGACTCTCTGCAATCCCCGCCTTCAAATACTATGAAGCCTGCACCTTC
CTGGTTTTTCTCTCCAACTTCATCATCCAGATGCTAGTGACAAACAGGCCCCCAGCTCTGGCCATCACGT
ATAGCATCACCTTCCTCCTCTTCCTCCTCATCCTTTTTGTCTGCTTCTCAGAGGACCTGATGAGGTGTGT
CCTGAAAGGCCCCAAGATGCTGCACTGGCTGCCTGCACTGTCTGGCCTGGTGGCCACACGACCAGGACTG
AGAATAGCCTTGGGCACCGCCACCATCCTCCTTGTCTTTGCCATGGCCATTACCAGCCTGTTCTTCTTCC
CAACATCATCAGACTGCCCTTTCCAAGCTCCCAATGTGTCCTCCATGATTTCCAACCTCTCCTGGGAGCT
CCCTGGGTCTCTGCCTCTCATCAGTGTCCCATACTCCATGCACTGCTGCACGCTGGGCTTCCTCTCCTGC
TCCCTCTTTCTGCACATGAGCTTCGAGCTGAAGCTGCTGCTGCTCCTGCTGTGGCTGGCGGCATCCTGCT
CCCTCTTCCTGCACTCCCATGCCTGGCTGTCGGAATGCCTCATCGTCCGCCTCTATCTGGGCCCCTTGGA
CTCCAGGCCCGGAGTGCTGAAGGAGCCCAAACTGATGGGTGCTATCTCCTTCTTCATCTTCTTCTTCACC
CTCCTTGTCCTGGCTCGCCAGAATGAGTACTACTGCCGCCTGGACTTCCTGTGGAAGAAGAAGCTGAGGC
AGGAGAGGGAGGAGACAGAGACGATGGAGAACCTGACTCGGCTGCTCTTGGAGAACGTGCTCCCTGCACA
CGTGGCCCCCCAGTTCATTGGCCAGAACCGGCGCAACGAGGATCTCTACCACCAGTCCTATGAATGCGTT
TGTGTCCTCTTCGCCTCAGTCCCAGACTTCAAGGAGTTCTACTCTGAATCCAACATCAATCATGAGGGCC
TAGAGTGTCTGAGGCTGCTCAATGAGATAATTGCTGATTTTGATGAGCTGCTCTCCAAGCCCAAGTTCAG
TGGGGTGGAGAAGATCAAGACCATCGGCAGCACCTACATGGCAGCCACAGGCTTAAATGCCACCTCTGGA
CAGGATGCACAACAGGATGCTGAACGGAGCTGCAGCCACCTTGGCACTATGGTGGAATTTGCCGTGGCCC
TGGGGTCTAAGCTGGACGTCATCAACAAGCATTCATTCAACAACTTCCGCCTGCGAGTGGGGTTGAACCA
TGGACCCGTAGTAGCTGGAGTTATTGGGGCCCAGAAGCCGCAATATGACATTTGGGGCAACACAGTGAAC

GTGGCCAGCCGCATGGAGAGTACAGGAGTCCTTGGCAAAATCCAAGTGACTGAGGAGACAGCATGGGCCC
TACAGTCCCTGGGCTACACCTGCTACAGCCGGGGTGTCATCAAGGTGAAAGGCAAAGGGCAGCTCTGCAC
CTACTTCCTGAACACAGACTTGACACGAACTGGACCTCCTTCAGCTACCCTAGGCTGAGATTGCACTCGC
CTTCTAAGAACCTCAATAAAGAGACTCTGGGGTGTCTGGAGCCCATTGATGTCTG

Method:

1. Run the query in Blastn to identify the sequence.
2. Then, run Blastx to determine the similar proteins.

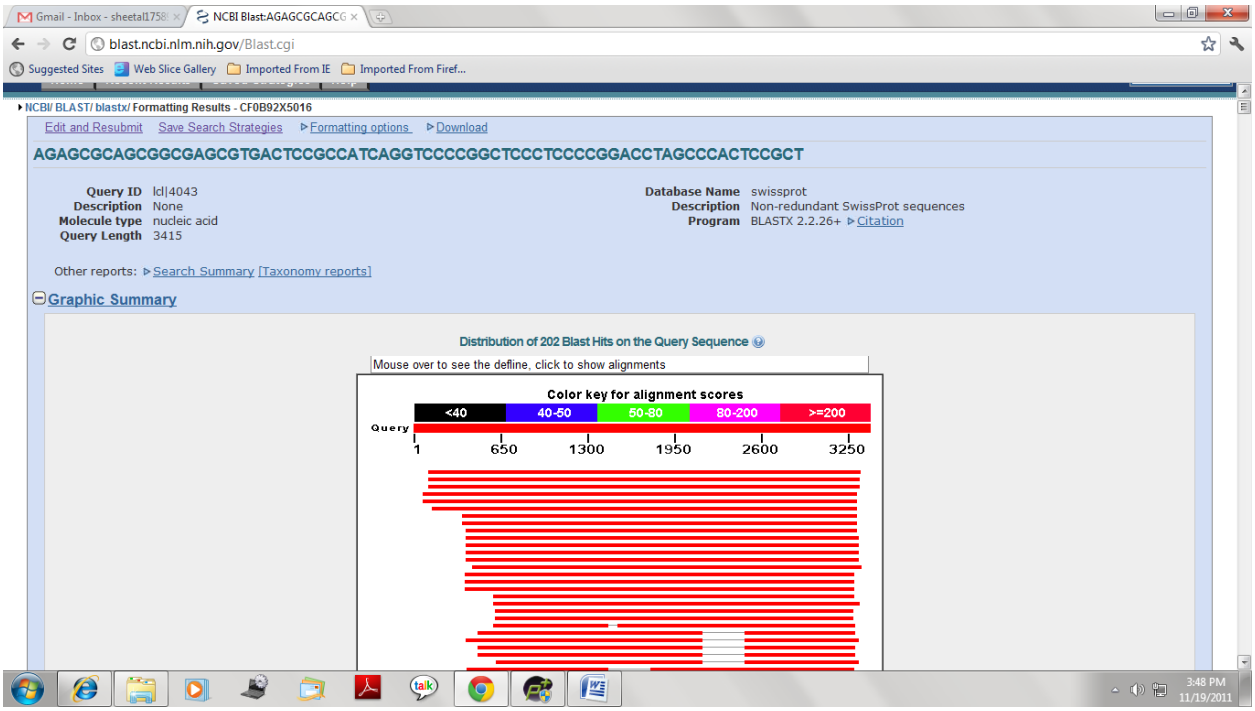
Results and inference:



Blastn results-

Homo sapiens adenylate cyclase 4 (ADCY4), transcript variant 2, mRNA
Score = 6159 bits (6830), Expect = 0.0
Identities = 3415/3415 (100%), Gaps = 0/3415 (0%)
Strand=Plus/Plus

Blastx results-



Similar protein:
sp|Q8NFM4.1|ADCY4_HUMAN
Score = 1926 bits (4990),Expect = 0.0
Identities = 1077/1077 (100%), Positives = 1077/1077 (100%), Gaps = 0/1077 (0%)
NP_640340.2
adenylate cyclase type 4 [Homo sapiens]

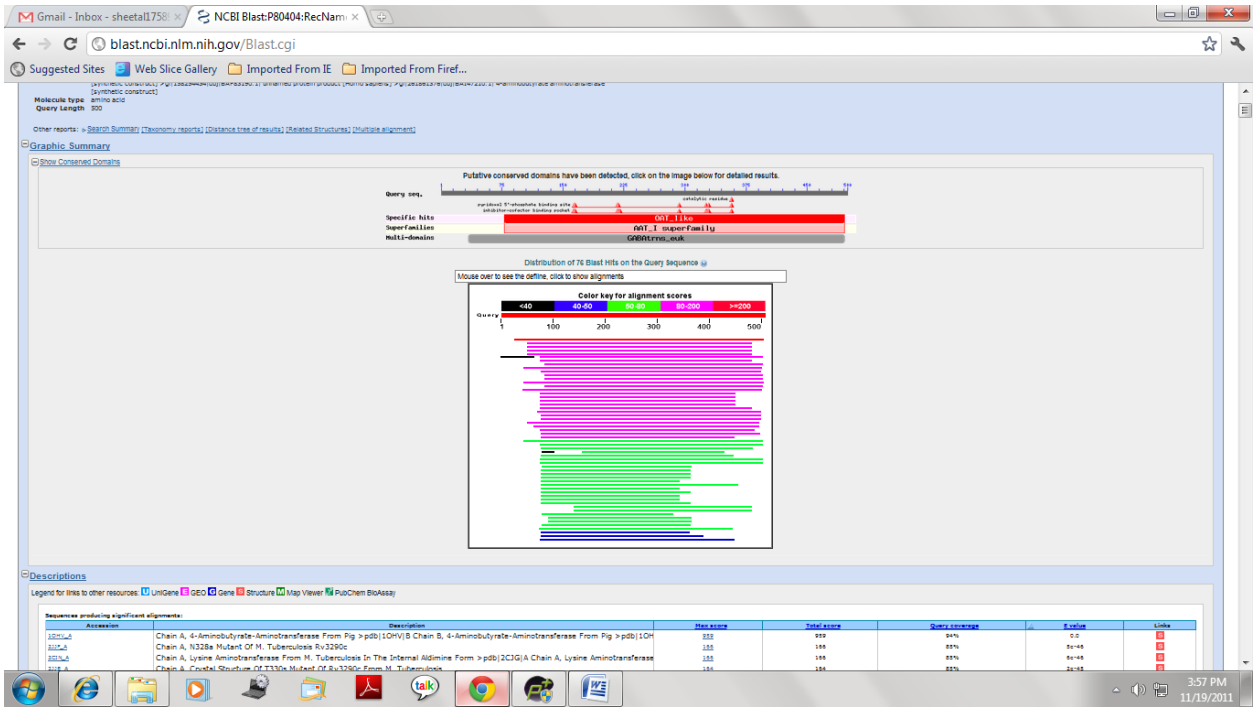
7. Find the structurally solved homologous proteins for **P80404**. Comment on the results.

Aim: To find Struturally solved homologus proteins for **P80404**

Method:

1. Run Blastp for the query against PDB.
2. Observe results.

Results and inference:



Structurally Similar protein:

pdb|1OHV|A

Chain A, 4-Aminobutyrate-Aminotransferase From Pig

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%)

Experiment 4: 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:

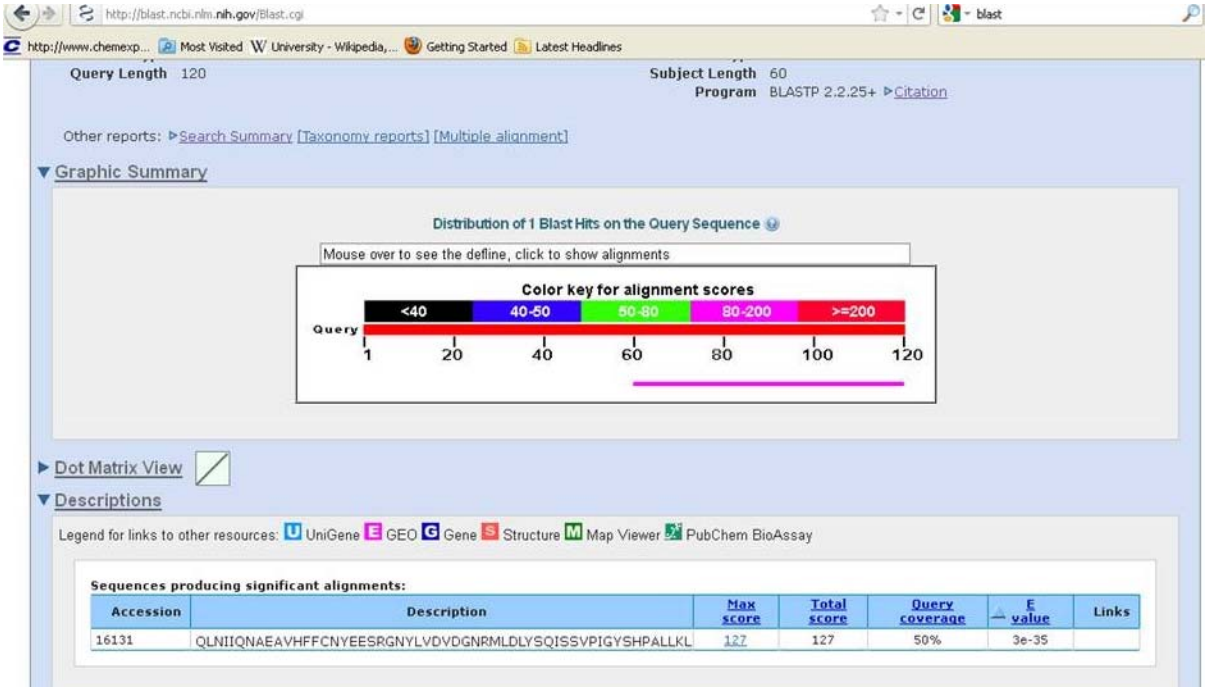
1. Enter the Given Sequences in Blastp.

MASMLLAQRLACSFQHSYRLLVPGSRHISQAAAKVDVEFDYDGPLMKTEVPGPRSQELMKQLNIIQNAEAVHFFCNYEESRGNYLV
DVDGNRMLDLYSQISSVPIGYSHPALLKLIQQPQNASMFVNRPALGILPPENFVEKLRQSLLSVAPKGMSQLITMACGSCSNENALK
TIFMWYR

QLNIIQNAEAVHFFCNYEESRGNYLVDVDGNRMLDLYSQISSVPIGYSHPALLKLIQQPQ
NASMFVNRPALGILPPENFVEKLRQSLLSVAPKGMSQLITMACGSCSNENALKTIFMWYR

2. Run Blast with following algorithmic parameters-
   Matrix: BLOSUM 62
   Gap costs: Existence 11 extension 1

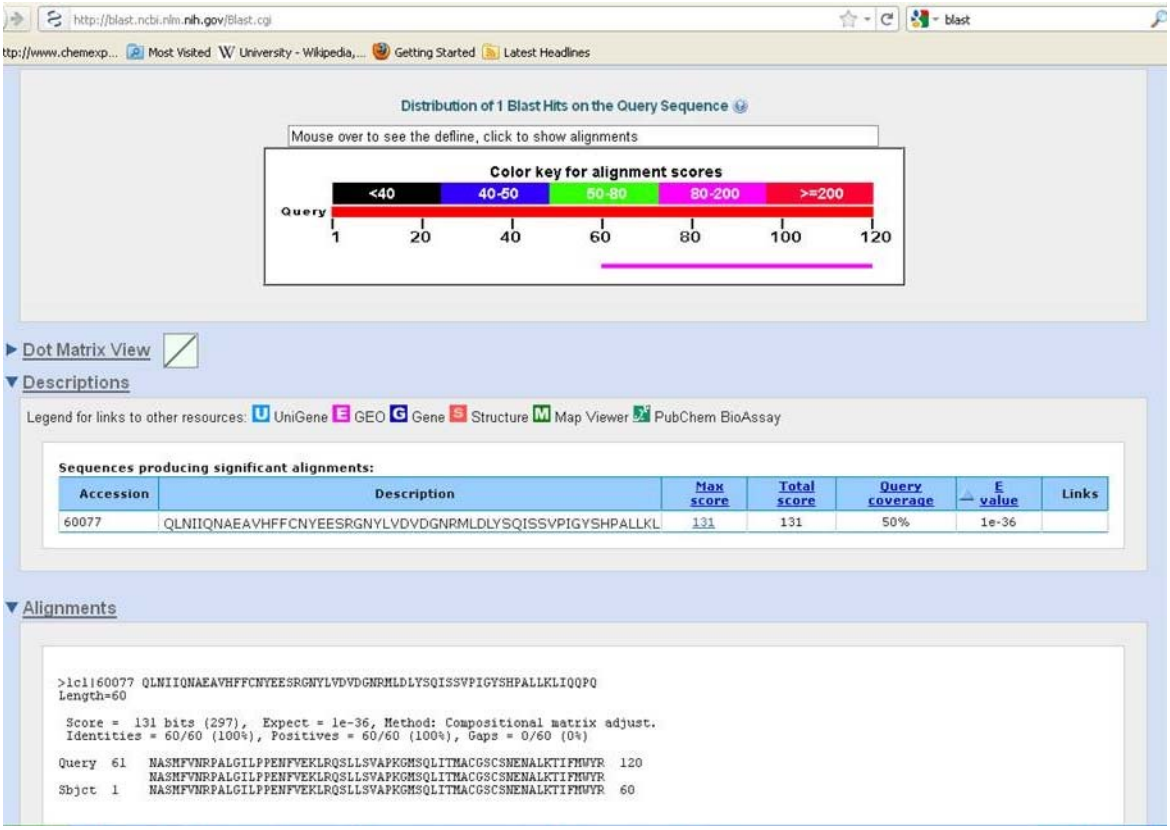
Result and inference:



>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%)

There is significant difference in the score and Expect value between the two BLOSUM scoring matrices.

2. Perform the local alignment between following sequences. Comment on the result.

Aim: Perform the local alignment between following sequences

MASMLLAQRLACSFQHSYRLLVPGSRHISQAAAKVDVEFDYDGPLMKTEVPGPRSQELMK
QLNIIQNAEAVHFFCNYEESRGNYLVDVDGNRMLDLYSQISSVPIGYSHPALLKLIQQPQ
NASMFVNRPALGILPPENFVEKLRQSLLSVAPKGMSQLITMACGSCSNENALKTIFMWYR

QLNIIQNAEAVHFFCNYEESRGNYLVDVDGNRCGSCSNENALKTIF

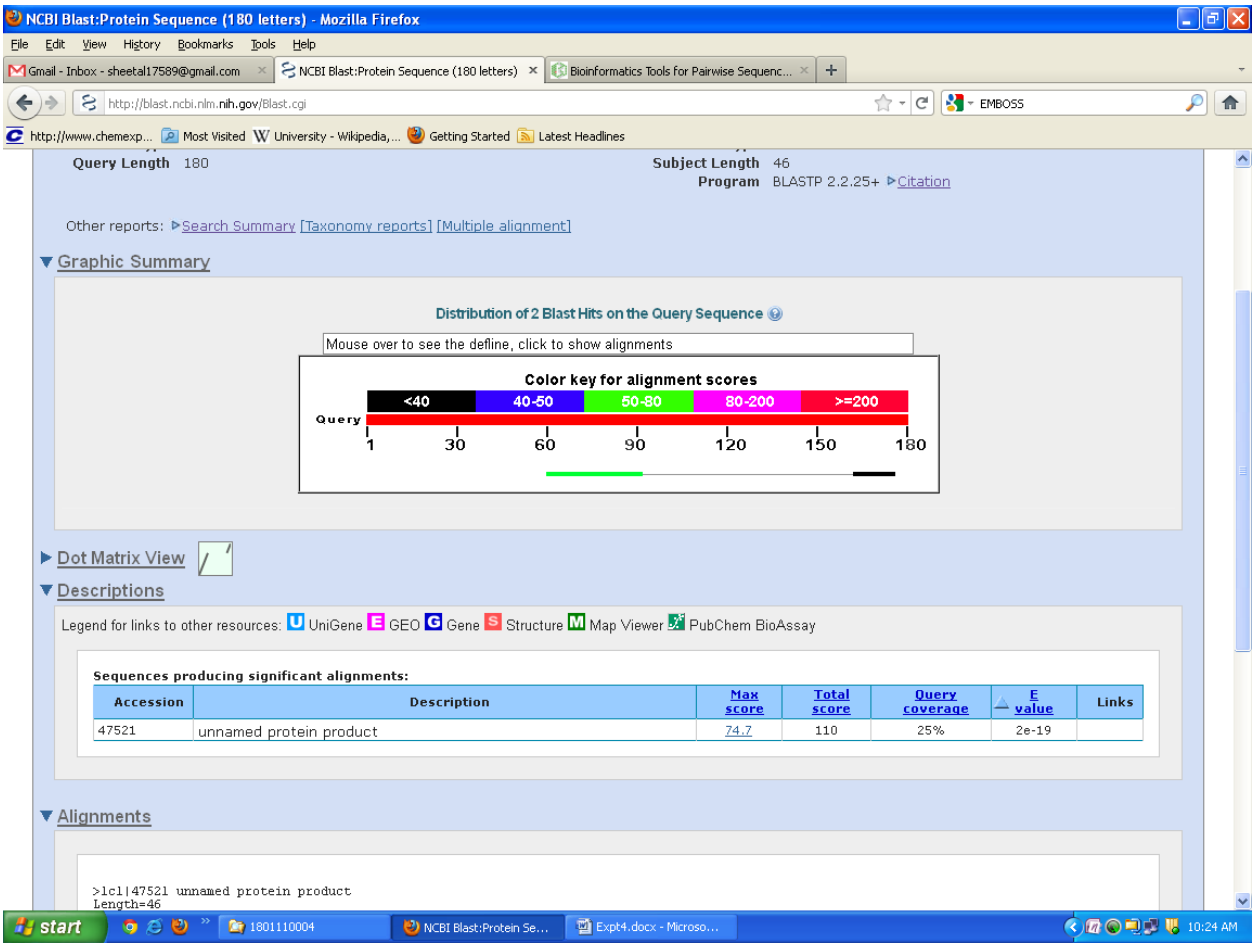
Introduction:

Local Alignment is an alignment that searches for segments of the two sequences that match well. There is no attempt to force entire sequences into an alignment, just those parts that appear to have good similarity, according to some criterion.

Method:

1. Enter the 2 sequences in Blastp with BLOSUM62 as the matrix.
2. Observe results.

Result and inference:



>lcl|47521 unnamed protein product
Length=46
Score = 74.7 bits (182), Expect = 2e-19, Method: Composition-based stats.
Identities = 32/32 (100%), Positives = 32/32 (100%), Gaps = 0/32 (0%)

Query 61 QLNIIQNAEAVHFFCNYEESRGNYLVDVDGNR 92
QLNIIQNAEAVHFFCNYEESRGNYLVDVDGNR
Sbjct 1 QLNIIQNAEAVHFFCNYEESRGNYLVDVDGNR 32

Score = 35.4 bits (80), Expect = 1e-07, Method: Composition-based stats.
Identities = 14/14 (100%), Positives = 14/14 (100%), Gaps = 0/14 (0%)

Query 163 CGSCSNENALKTIF 176

CGSCSNENALKTIF
Sbjct 33 CGSCSNENALKTIF 46

In the second sequence first 32 residues is found in the first query sequence from position 61 to 91 and the remaining are found in position 163 to 176.

3. Obtain the global alignment between the following sequences. Comment on the result.
QLNIIQNAEAVHFFCNYEESRGNYLVDVDGNRMLDLYSQISSVPIGYSHPALLKLIQQPQ
NASMFVNRPALGILPPENFVEKLRQSLLSVAPKGMSQLITMACGSCSNENALKTIFMWYR

QLNIIQNAEAVHFFCNYEESRGNYLYSQISSVPASMFVNRPALGILPPENFVSCSNENALKTIFMWY

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

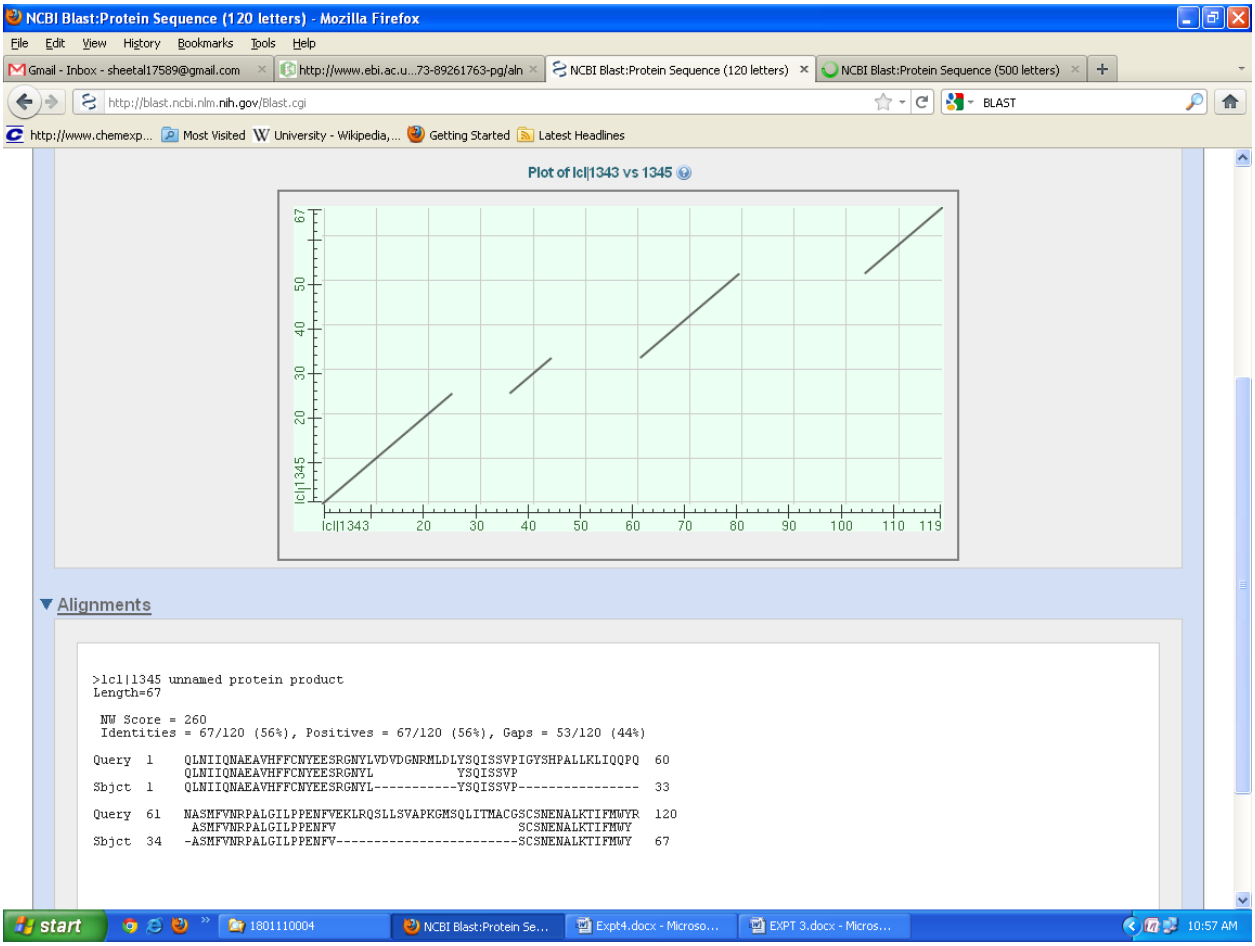
Method:

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%)



There is significant similar with significant score.

4. Compare the local and global alignments between the give n sequences. Comment on the results.

MASMLLAQRLACSFQHSYRLLVPGSRHISQAAAKVDVEFDYDGPLMKTEVPGPRSQELMKQLNIIQNAEAVHFFCNYEESRGNYLV
DVDGNRMLDLYSQISSVPIGYSHPALLKLIQQPQNASMFVNRPALGILPPENFVEKLRQSLLSVAPKGMSQLITMACGSCSNENALK
TIFMWYRMASMLLAQRLACSFQHSYRLLVPGSRHISQAALVDVDGNRMLDLYSQISSVPIGYSHPALLKLIQQPQNASMFVNRPAL
GILPPENFVQLITMACGSCSNENALKTIFMWYR

Aim: To compare the local and global alignments between the given sequences

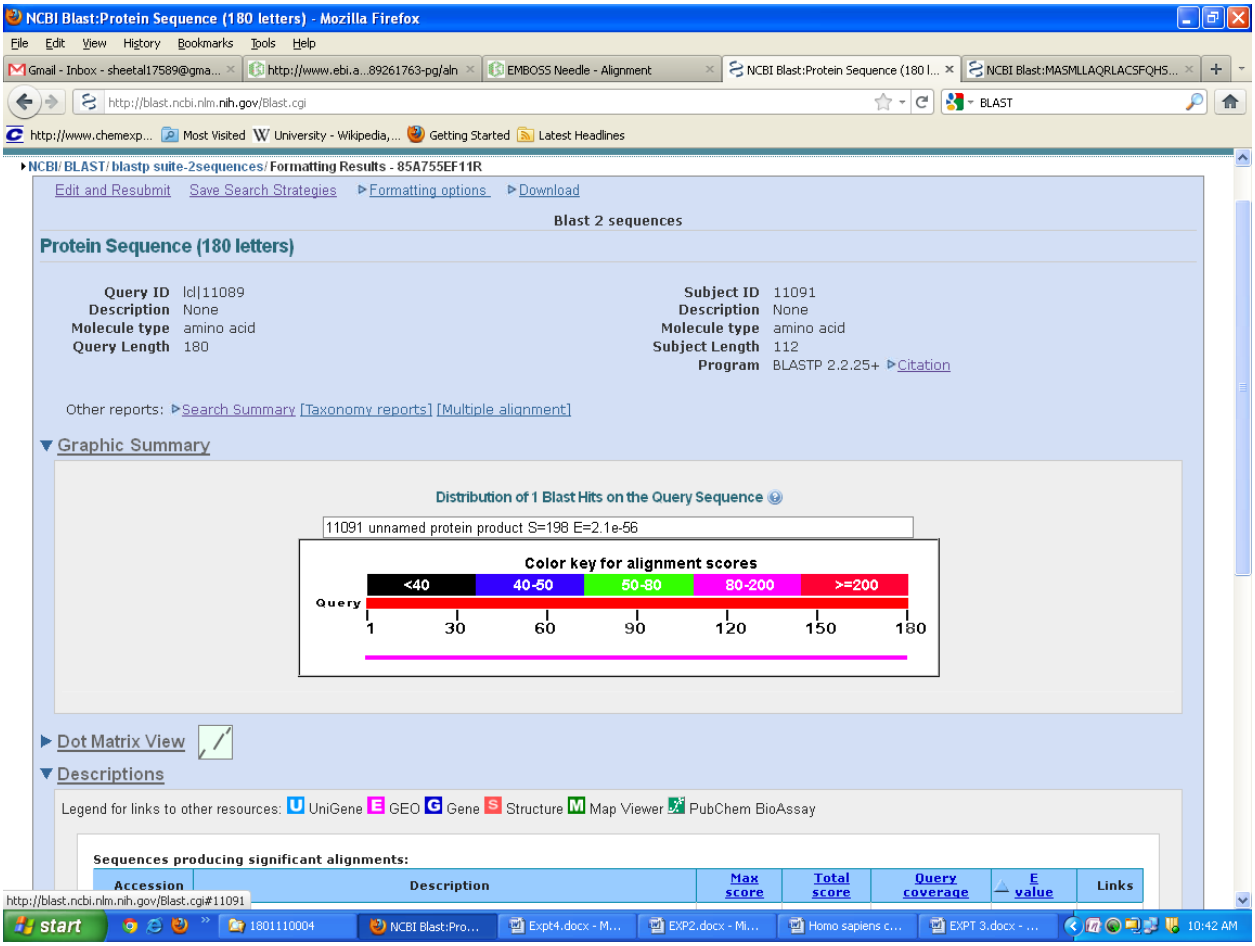
Introduction:
Global alignments, which attempt to align every residue in every sequence, are most useful when the sequences in the query set are similar and of roughly equal size. (This does not mean global alignments cannot end in gaps.) A general global alignment technique is the Needleman–Wunsch algorithm, which is based on dynamic programming. Local alignments are more useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context. The Smith–Waterman algorithm is a general local alignment method also based on dynamic programming. With sufficiently similar sequences, there is no difference between local and global alignments.

Method:

1. For local alignment, run the two query sequences in Blastp.
2. For global alignment, run the two query sequences in Needleman-Wunsch **Global Sequence Alignment Tool.**
3.

Result and inference:

Local Alignment:



>lcl|11091 unnamed protein product
Length=112

Score = 198 bits (504), Expect = 2e-56, Method: Compositional matrix adjust.
Identities = 112/180 (62%), Positives = 112/180 (62%), Gaps = 68/180 (38%)

Global Alignment:

NW Score = 487
Identities = 112/180 (62%), Positives = 112/180 (62%), Gaps = 68/180 (38%)
Difference: Scores are different.

Similar: Identities, Positives and Gap scores.

5. Perform the alignment using Needleman Wunsch algorithm between **P80404** and **P80147**.

   Aim: To perform the alignment using Needleman Wunsch algorithm between **P80404** and **P80147**

   Introduction:
   The **Needleman–Wunsch algorithm** performs a global alignment on two sequences (called *A* and *B* here). It is commonly used in bioinformatics to align protein or nucleotide sequences.

   Method:
   1. Retrieve the sequences from NCBI and run the Needleman-Wunsch **Global Sequence Alignment Tool.**
   2. Observe results.

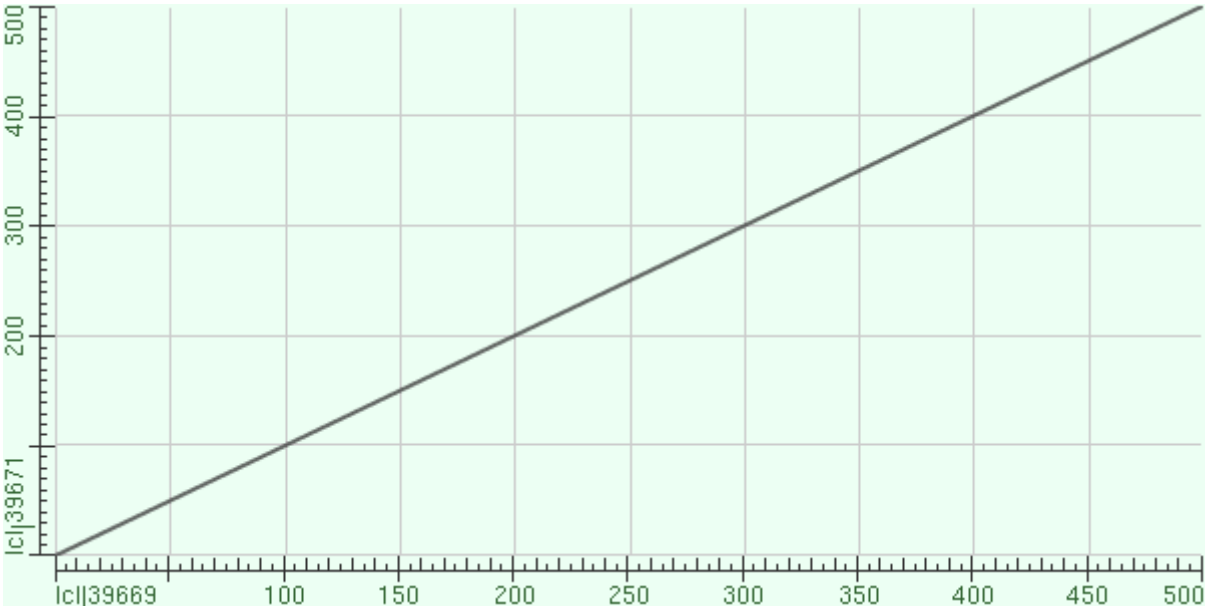
   Results and inference:

P80404: 4-aminobutyrate aminotransferase HUMAN

P80147: 4-aminobutyrate aminotransferase PIG

NW Score = 2536
Identities = 474/500 (95%), Positives = 490/500 (98%), Gaps = 0/500 (0%)
Query ID
lcl|39669
Description
gi|48429239|sp|P80404.3|GABT_HUMAN

Subject ID
39671
Description
gi|120968|sp|P80147.2|GABT_PIG


500
400
300
200
lcl|39671
lcll39669
100
150
200
250
300
350
400
450
500

Experiment 5: 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.

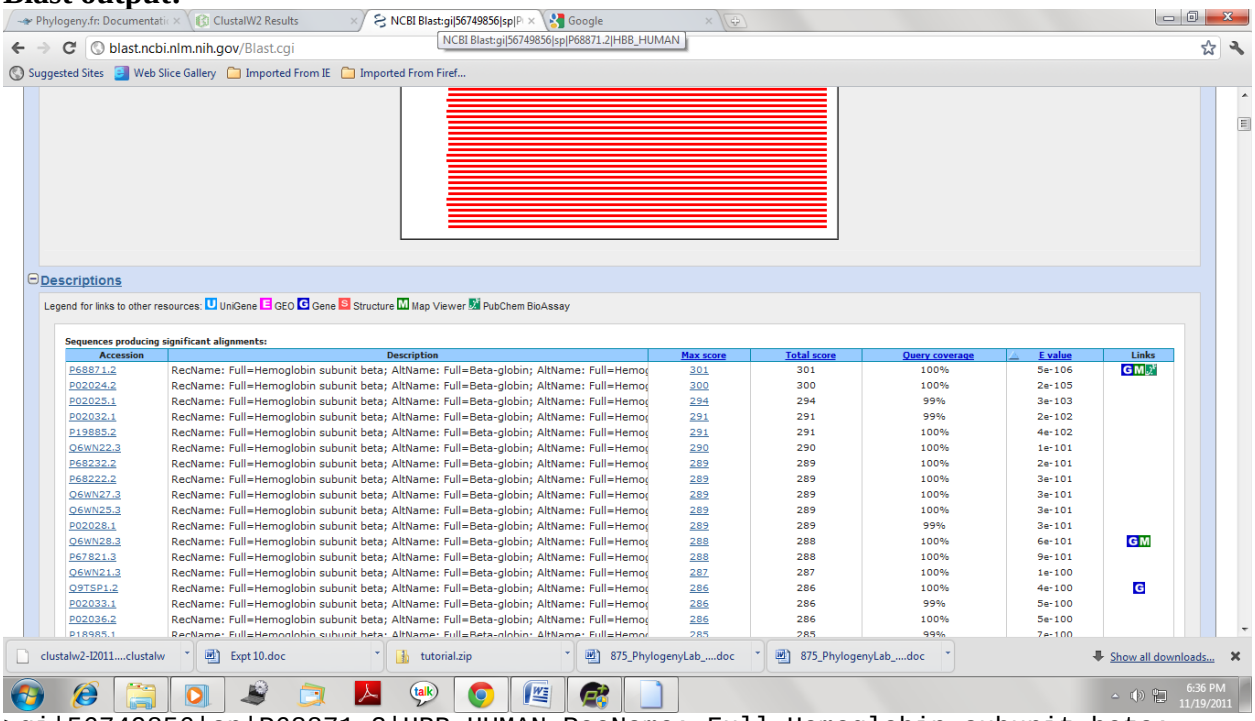
The "A la carte mode" offers the possibility of running and testing more alignment and phylogeny programs, such as MUSCLE, ClustalW, T-Coffee, PhyML,BioNJ, TNT.

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

Results and inference:

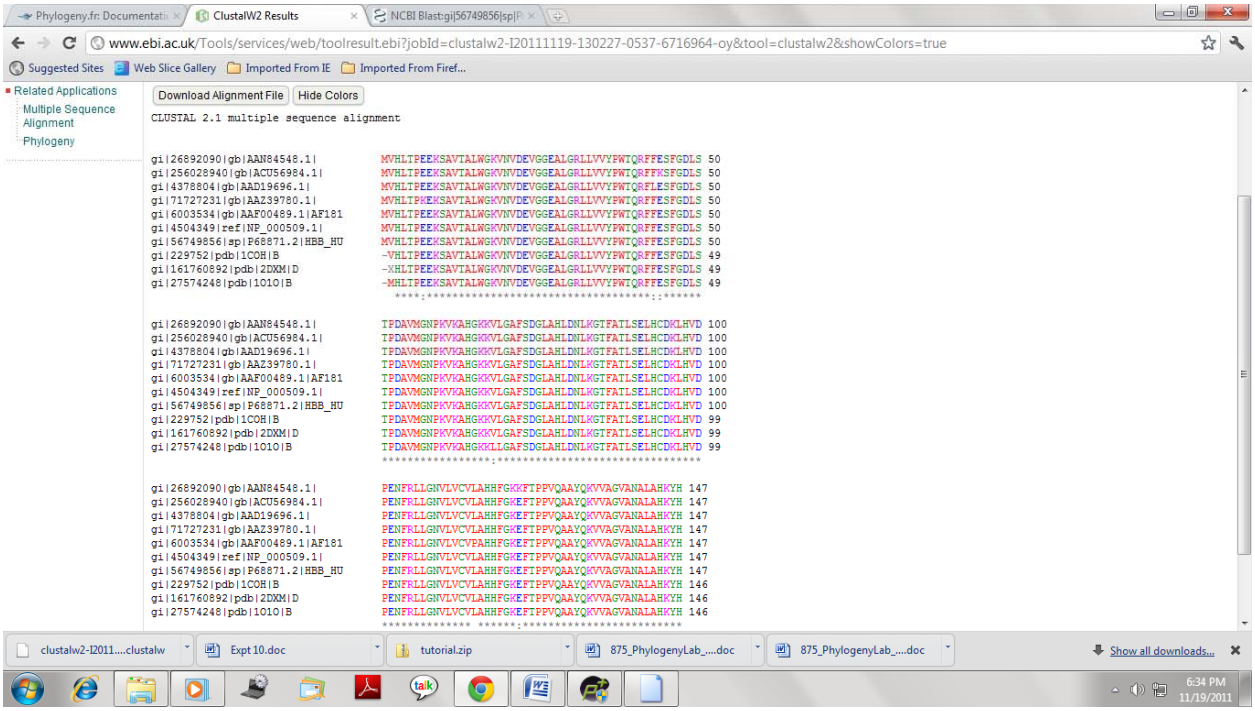
Blast output:




```
>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
```

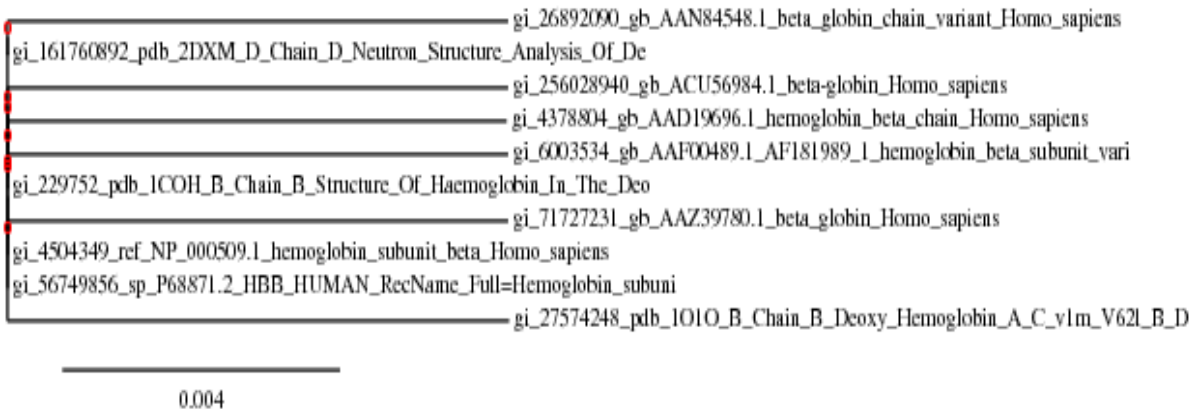
>gi|229752|pdb|1COH|B Chain B, Structure Of Haemoglobin In The Deoxy Quaternary State With Ligand Bound At The Alpha Haems
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGA
FSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANA
LAHKYH
>gi|6003534|gb|AAF00489.1|AF181989_1 hemoglobin beta subunit variant [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVPAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
>gi|71727231|gb|AAZ39780.1| 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|1O1O|B Chain B, Deoxy Hemoglobin (A,C:v1m,V62l; B,D:v1m,V67l)
MHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKLLGA
FSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANA
LAHKYH

Clustaw Alignment:

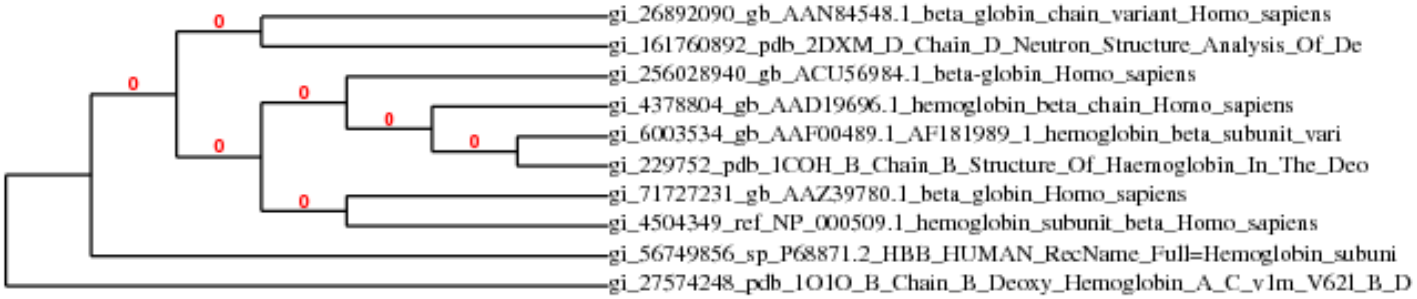


Red: small, hydrophobic, aromatic, not Y.
Blue: acidic. Magenta: basic.
Green: hydroxyl, amine, amide, basic.
Gray: others.
"*": identical.
":": conserved substitutions (same colour group).
".": semi-conserved substitution (similar shapes)
For the first 50 aa of the query,
Conserved regions:
HLTP
Semi-conserved: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

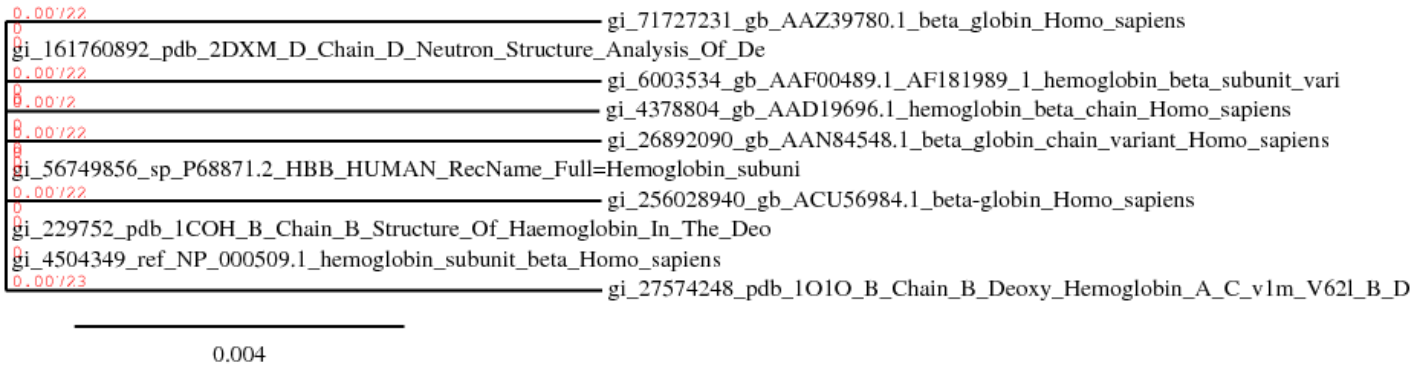
MA: ClustalW

Align curation: GBlocks

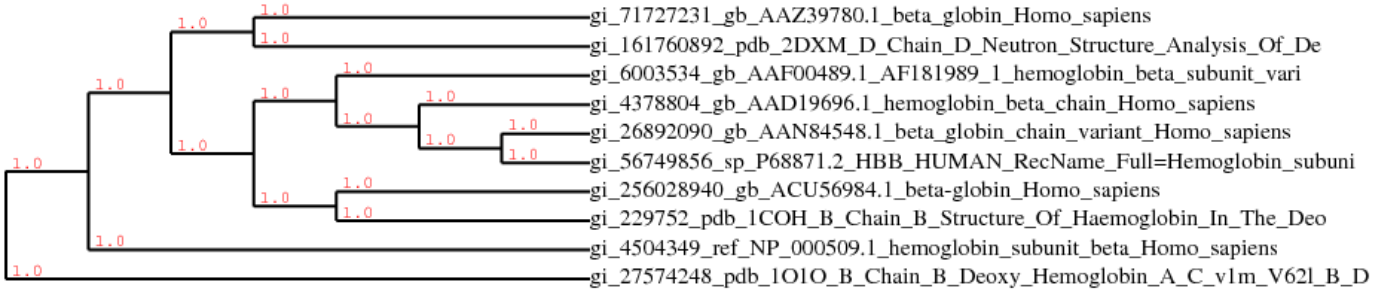
Contruction of tree: PhyML (http://www.phylogeny.fr/version2_cgi/one_task.cgi?task_type=phyml)

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.

Experiment 6: Gene Prediction

Aim:

Identify the Genes present if any in the given genomic sequence NC_010456.

Introduction:

Under the probabilistic model of gene structural and compositional properties used by GENSCAN, each possible "parse" (gene structure description) which is compatible with the sequence is assigned a probability. The default output of the program is simply the "optimal" (highest probability) parse of the sequence. The exons in this optimal parse are referred to as "optimal exons" and the translation products of the corresponding "optimal genes" are printed as GENSCAN predicted peptides. Of course, the optimal parse does not always correspond to the actual (biological) parse of the sequence, that is, the actual set of exons/genes present. In addition, there may be more than one parse which can be considered "correct", for example, in the case of a gene which is alternatively transcribed, translated or spliced. For both of these reasons, it may be of interest to consider "suboptimal" ("near-optimal") exons as well, i.e. exons which have reasonably high probability but are not present in the optimal parse. Specifically, for every potential exon E in the sequence, the probability P(E) is defined as the sum of the probabilities under the model of all possible "parses" (gene structures) which contain the exact exon E in the correct reading frame.

Given a probability cutoff C, suboptimal exons are those potential exons with P(E) > C which are not present in the optimal parse.

Suboptimal exons have a variety of potential uses. First, suboptimal exons sometimes correspond to real exons which were missed for whatever reason by the optimal parse of the sequence. Second, regions of a prediction which contain multiple overlapping and/or incompatible optimal and suboptimal exons may in some cases indicate alternatively spliced regions of a gene (Burge & Karlin, in preparation). The probability cutoff C used to determine which potential exons qualify as suboptimal exons can be set to any of a range of values between 0.01 and 1.00. The default value on the web page is 1.00, meaning that no suboptimal exons are printed. For most applications, a cutoff value of about 0.10 is recommended. Setting the value much lower than 0.10 will often lead to an explosion in the number of suboptimal exons, most of which will probably not be useful. On the other hand, if the value is set much higher than 0.10, then potentially interesting suboptimal exons may be missed. Gene: is aa locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and or other functional sequence regions

Exon: is a nucleic acid sequence that is represented in the mature form of an RNA molecule either after portions of a precursor RNA (introns) have been removed by cis-splicing or when two or more precursor RNA molecules have been ligated by trans-splicing. The mature RNA

molecule can be a messenger RNA or a functional form of a non-coding RNA such as rRNA or tRNA.

Intron: is any nucleotide sequence within a gene that is removed by RNA splicing to generate the final mature RNA product of a gene. The term *intron* refers to both the DNA sequence within a gene, and the corresponding sequence in RNA transcripts.

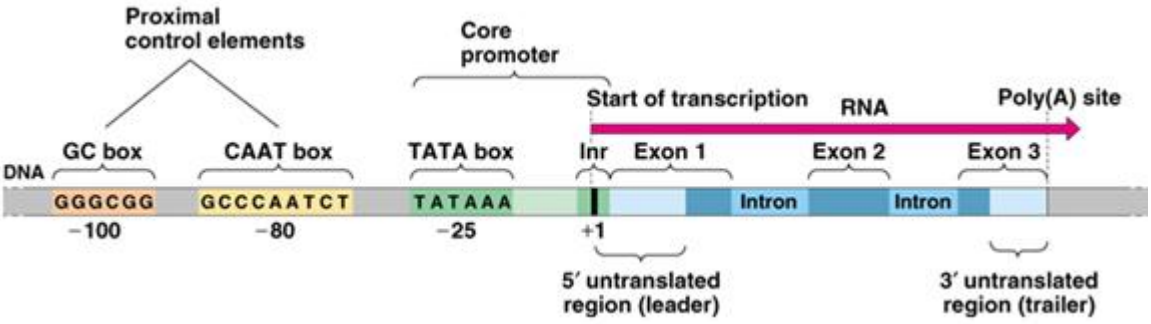
Intergenic region: (IGR) is a stretch of DNA sequences located between clusters of genes that contain few or no genes. Occasionally some intergenic DNA acts to control genes nearby, but most of it has no currently known function. It is one of the DNA sequences collectively referred to as junk DNA, though it is only one phenomenon labeled such and in scientific studies today, the term is less used. In humans, intergenic regions comprise a large percentage of the genome.

Isochore: is a large region of DNA (greater than 300 KB) with a high degree uniformity in G-C and C-G (collectively GC) which tends to have more genes, higher local melting or denaturation temperatures, and different flexibility. Overall, isochores are largely homogeneous in GC content in contrast to the heterogeneity of the entire genome.

Exon-intron boundary: is defined as the 5' splice site

Intron-exon boundary: is defined as the 3' splice site

Eukaryotic gene structure:



Method:

1. Open http://genes.mit.edu/GENSCAN.html
2. Enter the sequence ID .
3. Choose Organisms: Vertebrate; Arabidopsis; Maize and Suboptimal exon cut-off as 1.00
4. Run Genscan.

Results and inference:

NC_010456:

Sus scrofa breed mixed chromosome 14, Sscrofa10.2
DNA; linear; Length: 153,851,969 nt
Replicon Type: chromosome
Replicon Name: 14

GENSCAN Output

GENSCAN 1.0 Date run: 28-Oct-111 Time: 01:35:09

Sequence /tmp/10_28_11-01:35:07.fasta : 137437 bp : 45.48% C+G : Isochore 2 (43 - 51 C+G%)

Parameter matrix: HumanIso.smat

Predicted genes/exons:

Gn.Ex Type S .Begin ...End .Len Fr Ph I/Ac Do/T CodRg P.... Tscr...

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

1.07 Intr - 980 825 156 2 0 88 96 209 0.992 21.91

1.06 Intr - 1478 1355 124 1 1 84 30 53 0.728 -0.31

1.05 Intr - 2257 2195 63 0 0 133 77 129 0.999 14.13

1.04 Intr - 2574 2453 122 0 2 92 91 180 0.994 17.99

1.03 Intr - 2983 2795 189 1 0 131 60 163 0.994 17.58

1.02 Intr - 6253 6140 114 1 0 59 73 299 0.418 26.14

1.01 Init - 6680 6606 75 2 0 95 83 165 0.977 17.79

1.00 Prom - 10426 10387 40 -10.25

2.00 Prom + 10680 10719 40 -3.16

2.01 Init + 17975 18118 144 1 0 65 48 81 0.517 1.93

2.02 Intr + 20230 20515 286 0 1 93 26 413 0.039 32.51

2.03 Intr + 23199 23242 44 1 2 48 100 5 0.017 -4.34

2.04 Intr + 24382 25089 708 0 0 106 50 345 0.633 24.14

2.05 Term + 33832 33918 87 0 0 54 52 115 0.147 2.16

2.06 PlyA + 34674 34679 6 1.05

3.04 PlyA - 37228 37223 6 1.05

3.03 Term - 37465 37293 173 2 2 87 46 156 0.757 9.29

3.02 Intr - 38310 38303 8 2 2 71 113 0 0.456 -5.82

3.01 Init - 39366 39257 110 0 2 72 76 104 0.927 7.29

3.00 Prom - 44781 44742 40 -3.76

4.00 Prom + 50694 50733 40 -4.06

4.01 Init + 52037 52103 67 1 1 94 80 27 0.022 3.89

4.02 Intr + 57912 58061 150 1 0 60 36 117 0.370 3.83

4.03 Intr + 58710 58955 246 1 0 97 3 208 0.532 10.63

4.04 Term + 60919 61070 152 2 2 114 48 89 0.575 5.57

4.05 PlyA + 62517 62522 6 1.05

5.04 PlyA - 62625 62620 6 1.05

5.03 Term - 63618 63452 167 1 2 74 47 85 0.518 1.08

5.02 Intr - 64008 63860 149 2 2 123 0 116 0.696 6.18

5.01 Init - 64213 64209 5 1 2 79 98 0 0.707 -0.43

5.00 Prom - 64967 64928 40 -7.76

6.00 Prom + 65831 65870 40 -4.26

6.01 Init + 69724 69814 91 0 1 73 25 113 0.643 4.15

6.02 Term + 72265 72371 107 2 2 104 53 70 0.837 3.67

6.03 PlyA + 73213 73218 6 1.05

7.00 Prom + 77758 77797 40 -3.86

7.01 Init + 77872 78087 216 0 0 55 27 156 0.494 4.94

7.02 Term + 82159 82311 153 0 0 105 48 130 0.795 8.62

7.03 PlyA + 83120 83125 6 1.05

8.00 Prom + 89284 89323 40 -6.56

8.01 Init + 91007 91040 34 1 1 99 72 49 0.095 2.70

8.02 Intr + 92146 92246 101 2 2 97 41 54 0.114 1.43

8.03 Term + 97656 97808 153 2 0 122 53 124 0.863 10.22

8.04 PlyA + 99067 99072 6 1.05

9.04 PlyA - 100863 100858 6 1.05

9.03 Term - 104728 104589 140 2 2 93 53 104 0.853 5.53

9.02 Intr - 108675 108554 122 2 2 113 30 64 0.410 3.24

9.01 Init - 110760 110687 74 0 2 64 62 14 0.086 -3.06

9.00 Prom - 115534 115495 40 -2.56

10.04 PlyA - 115558 115553 6 1.05

10.03 Term - 116039 115888 152 0 2 136 40 175 0.748 15.57

10.02 Intr - 123513 123397 117 1 0 94 64 14 0.113 0.04

10.01 Init - 125080 125023 58 1 1 99 52 62 0.181 3.40

10.00 Prom - 126832 126793 40 -6.86

11.04 PlyA - 126944 126939 6 1.05

11.03 Term - 127975 127808 168 1 0 78 52 122 0.700 5.48

11.02 Intr - 130349 130138 212 0 2 104 3 163 0.417 8.03

11.01 Intr - 135637 135458 180 2 0 122 -15 111 0.288 4.04

Suboptimal exons with probability > 0.010

Exnum Type S .Begin ...End .Len Fr Ph B/Ac Do/T CodRg P.... Tscr..

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

S.001 Intr - 1550 1355 196 1 1 34 30 99 0.245 -1.88

S.002 Intr - 3469 3322 148 0 1 73 3 65 0.071 -3.26

S.003 Intr - 3407 3328 80 0 2 67 3 66 0.013 -4.73

S.004 Intr - 3469 3328 142 0 1 73 3 89 0.408 -0.74

S.005 Intr - 3514 3328 187 0 1 57 3 60 0.011 -5.91

S.006 Intr - 4004 3726 279 0 0 90 70 86 0.026 4.45

S.007 Intr - 4004 3814 191 0 2 90 -67 167 0.024 0.70

S.008 Intr - 4004 3821 184 0 1 90 -24 176 0.477 5.45

S.009 Intr - 4251 4112 140 2 2 118 -57 58 0.020 -5.59

S.010 Intr - 4251 4151 101 2 2 118 -57 70 0.037 -4.75

S.011 Intr - 4870 4766 105 0 0 125 39 31 0.057 1.23

S.012 Intr - 6253 6127 127 1 1 59 9 284 0.046 17.85

S.013 Intr - 6253 6136 118 1 1 59 2 316 0.224 20.14

S.014 Intr - 6294 6136 159 1 0 86 2 331 0.011 24.26

S.015 Intr - 6361 6136 226 1 1 84 2 277 0.015 16.26

S.016 Intr - 6294 6140 155 1 2 86 73 314 0.012 29.49

S.017 Intr - 6361 6140 222 1 0 84 73 260 0.032 22.42

S.018 Intr - 6253 6144 110 1 2 59 36 293 0.072 20.28

S.019 Intr - 6253 6148 106 1 1 59 24 288 0.124 19.29

S.020 Intr - 6361 6148 214 1 1 84 24 249 0.017 16.39

S.021 Intr - 6361 6349 13 1 1 84 57 -5 0.017 -9.65

S.022 Intr - 6520 6431 90 1 0 60 50 6 0.017 -5.81

S.023 Intr - 6520 6481 40 1 1 60 40 4 0.019 -9.40

S.024 Intr - 6790 6606 185 2 2 26 83 185 0.012 11.31

S.025 Init + 8639 8703 65 1 2 78 40 37 0.015 -1.48

S.026 Term - 9407 9351 57 2 0 75 45 62 0.013 -1.71

S.027 Init + 9569 9600 32 1 2 56 75 28 0.011 -2.49

S.028 Init + 10873 10940 68 0 2 96 17 -3 0.014 -6.06

S.029 Init + 10873 10964 92 0 2 96 52 -18 0.036 -4.64

S.030 Term - 12023 11915 109 1 1 125 46 48 0.181 2.28

S.031 Term - 12023 12003 21 2 0 125 49 15 0.051 -0.39

S.032 Sngl - 12283 12089 195 1 0 99 43 92 0.086 0.89

S.033 Init - 12283 12170 114 1 0 99 46 45 0.061 1.55

S.034 Init - 12283 12216 68 1 2 99 -6 44 0.011 -3.42

S.035 Init - 12283 12226 58 1 1 99 52 62 0.021 3.40

S.036 Intr - 12772 12709 64 2 1 38 97 47 0.129 -1.62

S.037 Init - 12926 12709 218 2 2 56 97 13 0.012 -2.83

S.038 Init + 12859 12890 32 0 2 63 72 -6 0.018 -5.77

S.039 Init + 13995 14005 11 2 2 37 116 2 0.259 -2.27

S.040 Term + 14074 14230 157 1 1 93 48 108 0.337 4.71

S.041 Init - 15253 15220 34 1 1 95 100 14 0.101 3.33

S.042 Init - 15666 15609 58 0 1 53 89 52 0.070 3.28

S.043 Init + 15967 15977 11 0 2 55 73 14 0.012 -3.99

S.044 Term + 16162 16297 136 1 1 131 37 78 0.051 4.49

S.045 Init - 17245 17214 32 1 2 94 52 51 0.011 1.32

S.046 Init + 17244 17293 50 2 2 80 75 37 0.011 2.02

S.047 Init + 17975 18209 235 1 1 65 -23 166 0.096 1.51

S.048 Sngl + 17975 18232 258 1 0 65 55 150 0.102 4.54

S.049 Term - 18191 18097 95 0 2 61 41 96 0.028 0.19

S.050 Intr + 18393 18626 234 2 0 42 55 57 0.016 -4.31

S.051 Intr + 18415 18626 212 2 2 59 55 34 0.015 -4.17

S.052 Intr + 18460 18603 144 0 0 32 -30 137 0.034 -3.22

S.053 Intr + 18460 18639 180 0 0 32 43 91 0.164 -0.94

S.054 Init - 18581 18467 115 2 1 39 62 73 0.021 0.45

S.055 Intr + 18479 18603 125 0 2 22 -30 141 0.017 -4.00

S.056 Intr + 18479 18639 161 0 2 22 43 95 0.073 -1.89

S.057 Intr + 19429 19455 27 0 0 82 81 -1 0.068 -3.09

S.058 Term + 20230 20592 363 0 0 93 42 422 0.620 32.67

S.059 Sngl + 20239 20592 354 0 0 101 42 399 0.286 32.65

S.060 Sngl + 20257 20592 336 0 0 95 42 374 0.033 29.53

S.061 Init + 23883 23939 57 2 0 96 10 84 0.899 2.04

S.062 Intr + 24382 24673 292 0 1 106 94 180 0.041 17.11

S.063 Intr + 24382 25018 637 0 1 106 26 296 0.027 16.87

S.064 Intr + 24382 25034 653 0 2 106 -1 318 0.026 15.41

S.065 Intr + 24382 25257 876 0 0 106 67 294 0.049 20.63

S.066 Term + 24382 25392 1011 0 0 106 45 291 0.045 18.71

S.067 Intr + 24394 25089 696 0 0 63 50 350 0.047 20.39

S.068 Init + 24415 25089 675 0 0 21 50 363 0.022 20.97

S.069 Init + 24433 25089 657 0 0 52 50 334 0.027 21.27

S.070 Intr + 25189 25257 69 0 0 83 67 -14 0.027 -4.62

S.071 Intr + 25189 25311 123 0 0 83 40 -10 0.014 -5.52

S.072 Term + 25837 25881 45 0 0 112 44 -6 0.047 -5.39

S.073 Term + 25837 25927 91 1 1 112 52 64 0.033 2.39

S.074 Term + 26677 26781 105 0 0 67 41 45 0.118 -3.89

S.075 Intr + 26948 27074 127 1 1 54 16 15 0.010 -8.85

S.076 Term + 26948 27100 153 1 0 54 53 29 0.026 -6.08

S.077 Intr + 27462 27492 31 2 1 107 94 -9 0.138 -0.70

S.078 Term + 27462 27608 147 2 0 107 40 22 0.173 -2.80

S.079 Intr + 27462 27492 31 1 1 107 94 14 0.099 1.09

S.080 Term + 27872 27901 30 1 0 102 39 -11 0.010 -6.75

S.081 Term + 27973 27978 6 0 0 86 49 0 0.015 -6.23

S.082 Term + 28765 28790 26 2 2 121 50 17 0.085 -0.41

S.083 Intr + 28765 28912 148 0 1 121 55 -7 0.039 -0.89

S.084 Init + 29535 29701 167 2 2 53 42 37 0.012 -5.28

S.085 Intr + 29595 29701 107 2 2 62 42 40 0.114 -4.19

S.086 Init + 29601 29625 25 2 1 103 62 15 0.015 0.10

S.087 Init + 29601 29701 101 2 2 103 42 28 0.314 -0.57

S.088 Term + 30337 30431 95 2 2 17 46 104 0.016 -2.91

S.089 Term + 31672 31727 56 2 2 120 42 21 0.018 -1.58

S.090 Term + 31672 31816 145 1 1 120 40 99 0.520 5.68

S.091 Term + 31747 31877 131 2 2 12 41 123 0.017 -1.66

S.092 Term + 31793 31877 85 2 1 59 41 106 0.011 0.13

S.093 Init + 32034 32110 77 2 2 37 73 57 0.084 -0.34

S.094 Init + 32085 32110 26 2 2 63 73 30 0.028 -1.95

S.095 Term - 32889 32666 224 1 2 128 39 80 0.129 4.18

S.096 Intr + 33681 33759 79 0 1 114 69 22 0.123 2.45

S.097 Term + 33681 33819 139 0 1 114 43 70 0.027 2.54

S.098 Term - 33914 33745 170 0 2 29 42 143 0.023 1.84

S.099 Sngl - 33945 33745 201 0 0 28 42 141 0.039 -1.31

S.100 Term - 33971 33745 227 0 2 82 42 135 0.245 5.24

S.101 Init - 33884 33821 64 2 1 56 42 28 0.018 -3.63

S.102 Init - 33890 33821 70 2 1 59 42 48 0.081 -1.45

S.103 Init - 33890 33857 34 2 1 59 20 53 0.013 -4.13

S.104 Init - 34146 34134 13 0 1 29 93 -12 0.062 -6.07

S.105 Term + 34344 34470 127 0 1 106 43 74 0.028 2.46

S.106 Init - 36881 36824 58 2 1 101 52 42 0.046 1.40

S.107 Init - 36881 36848 34 2 1 101 39 88 0.164 3.24

S.108 Term - 37465 37354 112 0 1 87 40 89 0.194 2.03

S.109 Init - 38251 38046 206 1 2 37 75 112 0.015 3.12

S.110 Init - 38321 38303 19 2 1 16 113 3 0.014 -4.14

S.111 Intr - 39013 38964 50 0 2 81 65 -21 0.244 -6.78

S.112 Intr - 39013 38991 23 0 2 81 4 18 0.024 -10.11

S.113 Term - 41813 41652 162 2 0 112 46 37 0.696 -0.16

S.114 Init - 42360 42277 84 0 0 34 111 14 0.031 -0.87

S.115 Intr - 42396 42277 120 0 0 47 111 44 0.067 3.29

S.116 Intr - 42410 42277 134 0 2 102 111 17 0.604 5.76

S.117 Term - 42410 42362 49 1 1 102 50 60 0.024 0.38

S.118 Init - 42504 42498 7 0 1 5 88 0 0.021 -7.19

S.119 Init - 42510 42498 13 0 1 32 88 4 0.072 -5.44

S.120 Init - 42767 42698 70 2 1 57 28 7 0.022 -7.12

S.121 Init - 42767 42734 34 2 1 57 20 40 0.040 -6.29

S.122 Term - 44120 43976 145 1 1 113 40 30 0.021 -1.92

S.123 Init - 45486 45371 116 0 2 69 91 46 0.051 0.71

S.124 Term - 46065 45945 121 2 1 96 47 56 0.027 0.25

S.125 Init - 46807 46719 89 1 2 68 64 15 0.015 -2.79

S.126 Init - 46807 46732 76 1 1 68 65 17 0.108 -1.35

S.127 Init - 46867 46732 136 1 1 20 65 50 0.019 -3.80

S.128 Init - 47321 47274 48 2 0 74 36 35 0.011 -2.15

S.129 Intr - 49601 49408 194 0 2 64 72 27 0.016 -2.09

S.130 Init - 49411 49408 4 1 1 83 72 0 0.016 -1.94

S.131 Intr - 49601 49530 72 0 0 64 41 45 0.030 -3.22

S.132 Term + 49878 49985 108 2 0 92 43 69 0.011 1.31

S.133 Term - 50616 50492 125 1 2 -10 47 99 0.015 -5.45

S.134 Init - 50611 50498 114 1 0 69 40 96 0.015 3.12

S.135 Term - 51170 51130 41 0 2 110 46 -13 0.012 -5.95

S.136 Term - 51745 51539 207 1 0 111 41 135 0.606 8.44

S.137 Intr - 51745 51548 198 1 0 111 -21 144 0.011 5.35

S.138 Intr - 51745 51553 193 1 1 111 -3 145 0.077 6.77

S.139 Intr - 51745 51592 154 1 1 111 33 121 0.266 8.55

S.140 Intr - 51991 51881 111 1 0 18 89 25 0.035 -3.82

S.141 Init - 52038 51898 141 0 0 67 3 91 0.379 -1.37

S.142 Intr - 52072 51898 175 0 1 35 3 115 0.045 -2.36

S.143 Init - 52725 52622 104 0 2 101 6 58 0.030 -1.39

S.144 Init - 52748 52743 6 2 0 78 86 10 0.430 0.11

S.145 Init - 52830 52828 3 0 0 50 101 0 0.068 -2.50

S.146 Init - 53189 53122 68 2 2 64 44 39 0.012 -2.35

S.147 Init - 53163 53122 42 0 0 54 44 25 0.011 -4.87

S.148 Term + 53879 54045 167 0 2 117 46 50 0.016 1.78

S.149 Init + 54490 54552 63 0 0 64 94 8 0.022 -1.20

S.150 Init - 55464 55399 66 0 0 87 72 -21 0.015 -2.62

S.151 Init + 56970 57018 49 2 1 78 78 10 0.351 0.27

S.152 Intr + 57373 57465 93 2 0 -31 61 128 0.080 -1.86

S.153 Intr + 57375 57465 91 2 1 -17 61 128 0.018 -0.83

S.154 Init + 57462 57465 4 2 1 67 61 0 0.012 -4.64

S.155 Init + 57917 58061 145 1 1 55 36 129 0.583 4.71

S.156 Intr + 58710 58895 186 1 0 97 -90 236 0.028 6.36

S.157 Intr + 58710 58900 191 1 2 97 -39 233 0.016 10.80

S.158 Term + 58710 59008 299 1 2 97 42 197 0.400 11.43

S.159 Term - 60529 60386 144 1 0 88 40 54 0.032 -1.49

S.160 Term - 61066 60982 85 0 1 37 42 73 0.014 -5.27

S.161 Term - 61701 61561 141 0 0 101 38 49 0.120 -0.87

S.162 Term - 61701 61614 88 2 1 101 43 24 0.057 -3.67

S.163 Term - 61701 61688 14 1 2 101 45 -10 0.023 -5.64

S.164 Intr - 61855 61804 52 1 1 78 87 46 0.034 1.98

S.165 Intr - 62237 62162 76 1 1 15 88 56 0.139 -2.18

S.166 Term - 63268 63125 144 1 0 43 44 85 0.056 -2.49

S.167 Term - 63383 63125 259 1 1 47 44 65 0.048 -6.98

S.168 Intr - 63383 63314 70 1 1 47 30 41 0.018 -6.62

S.169 Intr - 63618 63468 151 1 1 74 85 71 0.326 4.62

S.170 Term - 64008 63750 259 2 1 123 28 110 0.033 3.52

S.171 Intr - 64008 63824 185 2 2 123 -29 126 0.137 3.83

S.172 Intr - 64008 63828 181 2 1 123 -6 124 0.075 6.37

S.173 Init - 64844 64786 59 2 2 74 2 45 0.027 -4.72

S.174 Term - 65615 65428 188 0 2 31 42 98 0.036 -2.85

S.175 Init - 65592 65516 77 0 2 91 6 57 0.089 -1.64

S.176 Intr - 65615 65516 100 0 1 31 6 88 0.023 -5.93

S.177 Sngl + 66020 66142 123 1 0 60 41 129 0.016 -2.98

S.178 Init - 66120 66047 74 0 2 99 -1 48 0.024 -2.21

S.179 Init - 66120 66051 70 0 1 99 28 66 0.058 3.16

S.180 Init - 66120 66053 68 0 2 99 -6 67 0.064 -0.80

S.181 Init - 66120 66087 34 0 1 99 20 93 0.020 1.65

S.182 Init + 66467 66491 25 1 1 95 100 8 0.263 2.44

S.183 Term + 66989 67098 110 0 2 117 43 18 0.093 -1.13

S.184 Term + 67065 67195 131 1 2 41 41 69 0.011 -4.16

S.185 Init + 67501 67606 106 0 1 66 22 120 0.157 1.58

S.186 Sngl + 67501 67650 150 0 0 66 42 139 0.026 -0.73

S.187 Term + 67888 67988 101 2 2 44 39 82 0.034 -2.81

S.188 Term + 67903 67988 86 2 2 72 39 84 0.193 -0.28

S.189 Intr + 69734 69814 81 0 0 24 25 116 0.021 -1.49

S.190 Init + 69739 69814 76 0 1 39 25 120 0.150 2.05

S.191 Term - 70816 70682 135 1 0 66 47 103 0.011 2.02

S.192 Term - 73395 73388 8 1 2 110 44 0 0.012 -4.17

S.193 Term - 74642 74617 26 0 2 88 47 -6 0.014 -6.31

S.194 Term - 74995 74876 120 1 0 79 38 106 0.101 3.17

S.195 Intr - 74995 74911 85 1 1 79 77 93 0.047 6.69

S.196 Init - 75497 75495 3 2 0 92 48 0 0.019 -3.60

S.197 Init + 75532 75565 34 0 1 64 20 80 0.013 -3.00

S.198 Init + 75532 75601 70 0 1 64 83 32 0.271 1.37

S.199 Init + 75538 75601 64 0 1 58 83 42 0.149 0.50

S.200 Init - 75970 75794 177 1 0 44 25 86 0.029 -2.84

S.201 Init - 75828 75799 30 0 0 59 73 32 0.037 -2.56

S.202 Intr + 76718 76822 105 0 0 123 21 59 0.355 2.89

S.203 Term + 76718 76830 113 0 2 123 41 49 0.074 2.42

S.204 Intr - 76916 76867 50 0 2 61 99 18 0.026 -1.40

S.205 Init - 77012 76962 51 2 0 96 50 18 0.043 -0.04

S.206 Term + 77749 78035 287 2 2 43 41 182 0.134 4.57

S.207 Init + 77872 77954 83 0 2 55 0 70 0.011 -4.61

S.208 Sngl + 77872 78096 225 0 0 55 41 148 0.010 1.99

S.209 Intr + 77900 78087 188 0 2 60 27 123 0.244 2.81

S.210 Term + 77900 78096 197 0 2 60 41 115 0.017 1.57

S.211 Intr + 78200 78349 150 1 0 8 -6 115 0.015 -5.54

S.212 Term + 78200 78358 159 1 0 8 41 119 0.016 -2.86

S.213 Intr + 78395 78504 110 1 2 58 64 -32 0.017 -9.52

S.214 Intr + 78635 78821 187 2 1 126 -24 112 0.015 3.49

S.215 Intr + 78635 78828 194 2 2 126 10 122 0.018 7.34

S.216 Term + 78635 78658 24 1 0 126 51 27 0.237 1.02

S.217 Init - 78739 78655 85 1 1 34 96 90 0.034 5.38

S.218 Init + 80172 80183 12 2 0 99 52 -5 0.032 -2.68

S.219 Init + 80172 80229 58 2 1 99 42 30 0.025 0.55

S.220 Init + 80172 80285 114 2 0 99 9 31 0.020 -3.36

S.221 Init + 81163 81237 75 0 0 33 67 18 0.021 -4.70

S.222 Init + 81589 81594 6 0 0 41 81 17 0.076 -3.82

S.223 Init + 81589 81633 45 0 0 41 83 -7 0.029 -5.21

S.224 Init + 81814 81847 34 0 1 87 56 26 0.057 -0.65

S.225 Init + 81814 81852 39 0 0 87 22 25 0.071 -4.00

S.226 Init + 81829 81847 19 0 1 72 56 16 0.012 -2.89

S.227 Init + 81845 81847 3 1 0 78 56 0 0.062 -4.20

S.228 Term + 82159 82370 212 2 2 105 52 89 0.093 4.46

S.229 Term - 83509 83379 131 2 2 44 41 88 0.046 -1.96

S.230 Sngl - 83567 83379 189 2 0 39 41 141 0.014 -1.18

S.231 Init - 83864 83807 58 2 1 26 100 38 0.014 0.27

S.232 Init - 83870 83807 64 2 1 39 100 37 0.028 1.31

S.233 Init + 84017 84088 72 1 0 59 53 39 0.040 -1.43

S.234 Term - 84480 84361 120 0 0 52 41 93 0.011 -0.53

S.235 Intr + 85518 85626 109 2 1 31 98 48 0.034 -0.04

S.236 Term - 85844 85747 98 0 2 95 46 4 0.029 -4.97

S.237 Init - 86797 86725 73 1 1 89 66 30 0.041 2.13

S.238 Init + 86965 87127 163 0 1 67 63 1 0.036 -4.61

S.239 Term + 87242 87334 93 1 0 67 43 103 0.012 1.53

S.240 Init - 87309 87262 48 0 0 51 51 58 0.015 -0.55

S.241 Init + 87993 88053 61 2 1 17 41 42 0.019 -6.19

S.242 Term + 88285 88442 158 2 2 114 46 79 0.168 4.40

S.243 Term - 89632 89475 158 2 2 -51 41 170 0.125 -3.50

S.244 Term - 89559 89528 32 1 2 55 53 21 0.014 -6.68

S.245 Term - 89679 89645 35 1 2 125 50 2 0.321 -2.15

S.246 Init + 91375 91389 15 0 0 64 98 7 0.014 -0.37

S.247 Term - 91532 91493 40 1 1 111 43 83 0.075 2.76

S.248 Term - 91532 91522 11 0 2 111 46 -3 0.136 -3.84

S.249 Init - 91707 91574 134 0 2 60 -17 131 0.029 -1.19

S.250 Init - 91707 91604 104 0 2 60 -29 130 0.013 -2.40

S.251 Init - 91707 91632 76 0 1 60 11 133 0.036 3.44

S.252 Init - 91707 91638 70 0 1 60 54 127 0.455 7.10

S.253 Intr - 91767 91638 130 0 1 84 54 94 0.032 5.87

S.254 Intr - 91907 91638 270 0 0 32 54 123 0.038 0.81

S.255 Intr - 91925 91638 288 0 0 34 54 135 0.081 1.92

S.256 Intr - 91767 91638 130 1 1 84 54 -6 0.017 -4.65

S.257 Init + 91743 91788 46 2 1 64 70 30 0.041 -0.35

S.258 Intr + 91857 91946 90 1 0 63 94 -17 0.011 -3.63

S.259 Term + 92146 92273 128 2 2 97 48 54 0.024 0.74

S.260 Init + 92212 92342 131 0 2 28 79 97 0.056 2.42

S.261 Init - 92323 92296 28 1 1 56 109 62 0.040 2.90

S.262 Init - 92329 92296 34 1 1 59 109 74 0.139 4.70

S.263 Term + 92611 92707 97 1 1 124 52 32 0.055 0.64

S.264 Init - 92830 92726 105 1 0 57 75 37 0.024 -0.48

S.265 Init - 92830 92789 42 1 0 57 45 53 0.010 -1.68

S.266 Init + 95183 95224 42 1 0 50 115 -18 0.076 -2.32

S.267 Init + 95216 95224 9 1 0 62 115 0 0.586 0.60

S.268 Intr + 95438 95509 72 1 0 47 45 27 0.012 -6.10

S.269 Term - 95915 95784 132 2 0 17 41 140 0.019 0.29

S.270 Init + 95806 95863 58 0 1 99 52 57 0.019 2.77

S.271 Init + 95886 95918 33 2 0 42 9 95 0.016 -5.03

S.272 Init + 96345 96347 3 2 0 73 53 0 0.032 -5.00

S.273 Intr + 96897 96976 80 0 2 136 70 -10 0.011 1.19

S.274 Intr + 97656 97792 137 2 2 122 -8 134 0.044 6.57

S.275 Term + 97731 97861 131 1 2 35 41 133 0.021 1.64

S.276 Term + 98158 98333 176 2 2 115 40 57 0.020 1.52

S.277 Term + 98158 98344 187 1 1 115 46 33 0.030 -1.24

S.278 Init - 99668 99603 66 2 0 83 94 17 0.010 3.03

S.279 Intr - 99686 99603 84 2 0 82 94 45 0.018 4.52

S.280 Intr - 100052 99979 74 0 2 71 47 29 0.013 -3.77

S.281 Init - 101216 101159 58 2 1 99 52 63 0.028 3.42

S.282 Term - 104728 104635 94 0 1 93 53 35 0.017 -2.20

S.283 Term + 104827 104929 103 1 1 119 37 97 0.024 5.45

S.284 Init - 104998 104971 28 1 1 66 20 48 0.024 -5.28

S.285 Init - 105004 104971 34 1 1 61 20 61 0.063 -3.92

S.286 Init - 105854 105787 68 2 2 90 96 20 0.011 1.80

S.287 Init - 105796 105787 10 1 1 67 96 -5 0.230 -1.00

S.288 Intr - 105894 105787 108 1 0 72 96 -26 0.053 -3.14

S.289 Init - 106227 106212 16 0 1 86 45 0 0.025 -3.87

S.290 Init - 106502 106445 58 2 1 99 52 -10 0.022 -3.91

S.291 Init - 106965 106874 92 0 2 106 94 3 0.023 2.67

S.292 Term - 107477 107387 91 1 1 80 44 75 0.048 -0.51

S.293 Term - 107477 107428 50 0 2 80 47 79 0.088 0.47

S.294 Init - 107711 107687 25 2 1 90 91 -18 0.093 -1.48

S.295 Intr - 107725 107687 39 2 0 80 91 6 0.107 -1.70

S.296 Intr - 107898 107687 212 0 2 83 91 60 0.033 3.51

S.297 Intr - 107914 107687 228 0 0 85 91 55 0.017 2.48

S.298 Term - 108675 108522 154 2 1 113 48 67 0.024 2.59

S.299 Init - 109297 109287 11 1 2 66 65 0 0.078 -4.66

S.300 Init - 109327 109287 41 1 2 36 65 8 0.015 -7.04

S.301 Init - 109345 109287 59 1 2 46 65 4 0.048 -5.32

S.302 Init - 109458 109400 59 0 2 29 21 58 0.010 -7.53

S.303 Intr - 109559 109400 160 0 1 64 21 88 0.146 -1.25

S.304 Init - 110356 110235 122 1 2 42 82 -8 0.011 -6.33

S.305 Init - 110301 110271 31 0 1 36 86 31 0.100 -2.37

S.306 Init - 110356 110271 86 1 2 42 86 -3 0.033 -4.80

S.307 Init - 110837 110816 22 2 1 33 68 25 0.011 -5.02

S.308 Init - 110918 110816 103 2 1 67 68 -2 0.024 -3.87

S.309 Init - 111102 110964 139 0 1 82 16 30 0.015 -4.29

S.310 Init - 111153 110964 190 0 1 41 16 79 0.011 -4.74

S.311 Init - 111675 111616 60 0 0 83 58 20 0.029 -0.53

S.312 Init + 112407 112440 34 2 1 99 20 68 0.013 -0.76

S.313 Init + 112407 112464 58 2 1 99 52 62 0.231 3.40

S.314 Init + 112407 112474 68 2 2 99 -6 44 0.012 -3.42

S.315 Init + 112407 112476 70 2 1 99 28 43 0.032 0.55

S.316 Init + 112407 112522 116 2 2 99 2 69 0.066 -0.98

S.317 Sngl + 112407 112526 120 2 0 99 47 89 0.018 -1.71

S.318 Init + 112706 112881 176 1 2 59 41 57 0.019 -2.84

S.319 Init + 112787 112881 95 1 2 65 41 83 0.324 1.18

S.320 Intr + 113982 114113 132 0 0 111 0 100 0.074 3.26

S.321 Term + 113982 114126 145 0 1 111 40 96 0.363 4.48

S.322 Intr + 113982 114113 132 1 0 111 0 59 0.026 0.02

S.323 Term + 113982 114151 170 1 2 111 46 58 0.222 1.94

S.324 Init + 114309 114366 58 2 1 99 52 62 0.160 3.40

S.325 Init + 114309 114376 68 2 2 99 25 44 0.012 -0.32

S.326 Sngl + 114309 114452 144 2 0 99 41 92 0.011 -0.56

S.327 Intr + 114896 114985 90 0 0 122 9 26 0.011 -1.93

S.328 Intr + 114896 114993 98 0 2 122 95 18 0.124 5.63

S.329 Intr + 114896 114997 102 0 0 122 5 32 0.014 -1.55

S.330 Term + 114896 115065 170 0 2 122 52 24 0.022 0.24

S.331 Intr + 115201 115238 38 0 2 12 97 24 0.070 -7.24

S.332 Init + 115216 115238 23 0 2 38 97 30 0.018 -1.92

S.333 Intr + 115577 115754 178 2 1 116 21 117 0.030 7.72

S.334 Term + 115577 115769 193 2 1 116 49 116 0.145 7.39

S.335 Term + 115577 115710 134 0 2 116 53 76 0.072 5.15

S.336 Term + 115577 115666 90 1 0 116 43 35 0.040 -0.48

S.337 Intr - 116618 116557 62 1 2 15 20 72 0.011 -8.52

S.338 Init + 116752 116807 56 0 2 78 89 -16 0.048 -1.74

S.339 Init - 116836 116791 46 1 1 68 47 5 0.199 -4.64

S.340 Init - 116879 116795 85 2 1 77 55 -39 0.036 -7.07

S.341 Init - 116894 116795 100 2 1 58 55 3 0.050 -6.58

S.342 Init - 116897 116795 103 2 1 47 55 13 0.029 -7.36

S.343 Init - 116903 116795 109 2 1 58 55 13 0.071 -6.04

S.344 Init - 116937 116886 52 0 1 47 24 38 0.109 -5.38

S.345 Init - 116937 116901 37 0 1 47 50 14 0.059 -6.28

S.346 Init - 116923 116901 23 1 2 85 50 43 0.012 -0.52

S.347 Init - 118741 118711 31 1 1 51 72 -4 0.012 -5.69

S.348 Term + 119006 119153 148 2 1 122 40 122 0.102 8.17

S.349 Intr + 119006 119036 31 1 1 122 91 -12 0.019 0.20

S.350 Init + 119728 119761 34 0 1 99 20 74 0.147 0.63

S.351 Term - 120211 120036 176 2 2 125 41 96 0.356 6.52

S.352 Init + 120348 120387 40 2 1 39 94 21 0.038 -1.85

S.353 Init - 120779 120740 40 2 1 72 47 12 0.032 -3.85

S.354 Init - 120779 120746 34 2 1 72 35 16 0.014 -5.11

S.355 Init - 120901 120886 16 1 1 61 91 6 0.183 -1.21

S.356 Intr + 121000 121090 91 2 1 123 46 31 0.061 1.45

S.357 Term + 121000 121127 128 2 2 123 41 47 0.137 1.94

S.358 Init - 121603 121546 58 1 1 99 52 -8 0.026 -3.14

S.359 Init + 121961 121994 34 1 1 99 20 63 0.018 -1.30

S.360 Init + 121961 122018 58 1 1 99 52 45 0.133 1.62

S.361 Init + 121961 122030 70 1 1 99 28 26 0.014 -1.63

S.362 Sngl + 122695 122982 288 0 0 53 48 154 0.278 3.50

S.363 Term + 122747 122982 236 0 2 -11 48 183 0.027 0.88

S.364 Term + 122793 122982 190 0 1 20 48 178 0.061 3.92

S.365 Term + 122843 122982 140 0 2 -6 48 178 0.084 2.53

S.366 Init + 122848 122950 103 0 1 49 -64 155 0.012 -3.19

S.367 Sngl + 122848 122982 135 0 0 49 48 163 0.092 1.86

S.368 Init + 124034 124115 82 1 1 50 40 56 0.069 -1.86

S.369 Term + 124610 124790 181 2 1 18 52 178 0.020 4.28

S.370 Term + 124615 124790 176 2 2 7 52 180 0.143 4.22

S.371 Init - 125080 125011 70 1 1 99 28 43 0.025 0.55

S.372 Init - 125080 125047 34 1 1 99 20 68 0.010 -0.76

S.373 Term - 127598 127514 85 1 1 39 41 111 0.077 -1.37

S.374 Intr - 127975 127812 164 1 2 78 38 109 0.080 3.67

S.375 Init - 128490 128272 219 0 0 40 18 105 0.076 -2.67

S.376 Init - 128790 128677 114 0 0 72 27 79 0.133 0.41

S.377 Intr - 128873 128677 197 0 2 96 27 99 0.038 3.73

S.378 Sngl - 130332 130123 210 0 0 33 52 168 0.021 3.15

S.379 Term - 130349 130123 227 0 2 104 52 171 0.330 12.04

S.380 Init - 130332 130138 195 0 0 33 3 160 0.111 1.20

S.381 Init - 130623 130614 10 0 1 53 13 0 0.017 -10.44

S.382 Init - 130629 130614 16 0 1 58 13 2 0.025 -9.86

S.383 Init - 130623 130620 4 0 1 53 23 0 0.025 -9.84

S.384 Init - 130629 130620 10 0 1 58 23 15 0.072 -8.33

S.385 Intr - 131846 131735 112 2 1 40 86 53 0.183 0.25

S.386 Intr - 132096 132076 21 0 0 74 98 0 0.025 -2.66

S.387 Intr - 132096 132076 21 1 0 74 98 10 0.084 -1.88

S.388 Intr - 132158 132076 83 1 2 88 98 -22 0.041 -1.92

S.389 Init - 132216 132210 7 0 1 65 47 0 0.040 -5.31

S.390 Init - 132833 132797 37 2 1 53 55 5 0.015 -6.12

S.391 Intr - 133494 133416 79 2 1 60 36 47 0.016 -3.75

S.392 Init - 133821 133788 34 0 1 82 70 -8 0.017 -4.65

S.393 Init - 133928 133857 72 2 0 88 48 1 0.058 -2.83

S.394 Init - 133928 133891 38 2 2 88 42 38 0.047 -1.31

S.395 Init - 133928 133908 21 2 0 88 37 15 0.028 -3.88

S.396 Term - 135637 135378 260 2 2 122 32 111 0.150 4.61

S.397 Intr - 135637 135450 188 2 2 122 -46 102 0.015 -0.39

S.398 Term - 135637 135502 136 0 1 122 40 89 0.309 4.99

S.399 Intr - 135637 135607 31 0 1 122 66 -2 0.017 -0.87

S.400 Term - 136467 136359 109 2 1 107 41 86 0.140 3.78

S.401 Init - 136719 136664 56 0 2 91 31 9 0.021 -3.74

S.402 Init - 137019 136925 95 0 2 53 101 15 0.170 -0.68

S.403 Init - 137019 136950 70 0 1 53 28 34 0.016 -4.72

S.404 Init - 137019 136962 58 0 1 53 42 36 0.035 -3.58

Predicted peptide sequence(s):

>/tmp/10_28_11-01:35:07.fasta|GENSCAN_predicted_peptide_1|281_aa

MEQYCILGRIGEGAHGIVFKAKHVETGEIVALKKVALRRLEDGIPNQALREIKALQEIED

SQYVVQLKAVFPHGAGFVLAFEYMLSDLAEVLRHAQRPLAQAQVKSYLQMLLKGVAF
CHA

NNIVHRDLKPANLLISASGQLKIADFGLARVFSPDGSRLYTHQVATRWYRAPELLYGAR
Q

YNQGVDLWAVGCILGELLNGSPLFPGENDIEQLCCVLRILGTPSPQVWPEITELPDYNKI

SFKEQAPVPLEEVLPDASPQALDLLGRFLLYPPLQRIAASQ

Function: Cyclin-dependent kinase
Score = 30.0 bits (66), Expect = 2.6, Method: Composition-based stats.
Identities = 11/23 (48%), Positives = 19/23 (83%), Gaps = 0/23 (0%)

>/tmp/10_28_11-01:35:07.fasta|GENSCAN_predicted_peptide_2|422_aa

MARPDPSHICDLHHGSWQCQILNPLSEARNRTCILMFPSRIQFHCARMSPKMAQSPKMA
Q

SFKMAQSPKMAQSPKMAQSPKMAQSFKMAQSPKMAQSFKMAQSPKMAQSPKMAQSP
KMAQ

SFKMAQSFKMAQSPNKAGGAREPAGPIKRLLLCSWRAKVWFQNRRARISMKNAQKMK
KPH

VVPGPTQGEVGAPVQRGNGDPVPGSSQGGLYAPVPGSSQGGLYVPVPGSSQGGLYAPV
PE

SNQDGLYAPIPGPNQGKFCAPSFSCQQGPLAAQRDASHFWNPEDIPGQAIFLGYFGNRG
V

NIALPNTEIPAEEPSGNPNCSFFSGFSPTFLTTSQQPFSWAEAGSADLGGLGMPLAGNQA

LQDWRQHPPSGEQQSWWNQQPSPPPLPTAVLEPQLQGQILNPLREVRDRIRILMDTSRV
G

YR

>/tmp/10_28_11-01:35:07.fasta|GENSCAN_predicted_peptide_3|96_aa

MAERREKKNEKKEKKKKKKKRGRGRERRGVEIGSLKSSKGPQPRHMEVPRLGVELELQ
LLLAYATAAALRDLSLICDIHHNSQQYWVLNPLMETRD

Function:
Probable serine/threonine-protein kinase mkcC
Length=891

Score = 30.0 bits (66), Expect = 2.6, Method: Composition-based stats.
Identities = 11/23 (48%), Positives = 19/23 (83%), Gaps = 0/23 (0%)

>/tmp/10_28_11-01:35:07.fasta|GENSCAN_predicted_peptide_4|204_aa

MLPPVMLPEPQGLGSSDEAWEQSGMWKFPSQEWNLCHSSNLSHCSDKARSLICGATGE
LLELVMSENDGRDVGLYLRRMEVPRRGVELELLAYAMATATQDPSHVFDLHHSSWQR
WILNPLSQARDQIRVLVDTSCVRYLCATTGAPLFLFFIVWLHPQHMEVPRLGVQSELHL
LATPQPQQRGIQAVSGTYTTVHGNAGSLTH

>/tmp/10_28_11-01:35:07.fasta|GENSCAN_predicted_peptide_5|106_aa

MGPAPEAYGSSGLGVKSELQLPAYATAKATPDPSHVCHLQHSSGHPLNGTRAHLKVSQ
PQYNPVTTCDHQLPFGTVQNEGHPQRWMPPPGGALAWVTPYQARYAPY

>/tmp/10_28_11-01:35:07.fasta|GENSCAN_predicted_peptide_6|65_aa

MSSGLMHPGLLYVDRKCLSEEQLLSATVGTWLLPWYMEVPRLGVESELQPPAYTTATA
MQDLSCI

>/tmp/10_28_11-01:35:07.fasta|GENSCAN_predicted_peptide_7|122_aa

MGLYTPWLPELASGNTRLGVELTQQLPAYTTATAMQDPSCVCDLPHISWQCQILSPLSK
ARDRTRNLVDTSQGCTCGIMEVLRLRIQSELQLLATATSTAMVDPGCICKLQYSSQQCR
SLTH

>/tmp/10_28_11-01:35:07.fasta|GENSCAN_predicted_peptide_8|95_aa

MRLQFRSLPLLGPHPQHMEVPRLGLTLELQLPGYATATTMWDPSHDYTCAMWRFPGY
GSNWSCSRRPTPQPQQHQIRAASANYTTSHGNTGSLTH

>/tmp/10_28_11-01:35:07.fasta|GENSCAN_predicted_peptide_9|111_aa

MSAAAPAWERGNSTDLTLAKRAEARPTLMVHGGSQARGQIGAVAPKPQSQPQQSQIQT
VAPTYTRGPHLQHMEVPRMGRIRGAAASPRHSHSNAESERSLRPTSQLMAAPH

>/tmp/10_28_11-01:35:07.fasta|GENSCAN_predicted_peptide_10|108_aa

MRLRVRSLPLLSGLTIRRCRIHWQALVSVGLLKAESSVLAGNEESALRKQMFPKSQTRR
PHWQHMEVSGLGVELELQLLAYTTAKATTTPDLSRICYLHPHLVATPDP

>/tmp/10_28_11-01:35:07.fasta|GENSCAN_predicted_peptide_11|186_aa

XPHPLHMQVPRLGITLDLQLPTTTTAPDPSHIGNPCCSLWQCKILNPLSKARDRIHILMN
RLHLWHMEVPRLGVESELQLPAYTAATATPDLSRIFYLYCSLWQHLILNLLSEATDGTR
NGTRILMNATRPNLQHMEVPRRGVKLELQLLAYPTATETPDPSHICSLHCSSRQCCILNP
LREAEE

Experiment 7: 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 310 helix and π 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
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKA
SEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSK
HPGDFGADAQGAMNKALELFRKDMASNYKELGFQG

>2
MDPKQTTLLCLVLCLGQRIQAQEGDFPMPFISAKSSPVIPLDGSVKIQCQAIREAYLTQL
MIIKNSTYREIGRRLKFWNETDPEFVIDHMDANKAGRYQCQYRIGHYRFRYSDTLELVV
TGLYGKPFLSADRGLVLMPGENISLTCSSAHIPFDRFSLAKEGELSLPQHQSGEHPANFSL
GPVDLNVSGIYRCYGWYNRSPYLWSFPSNALELVVTDSIHQDYTTQNLIRMAVAGLVL
VALLAILVENWHSHTALNKEASADVAEPSWSQQMCQPGLTFARTPSVCK

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:

Output for seq 1 and 2:

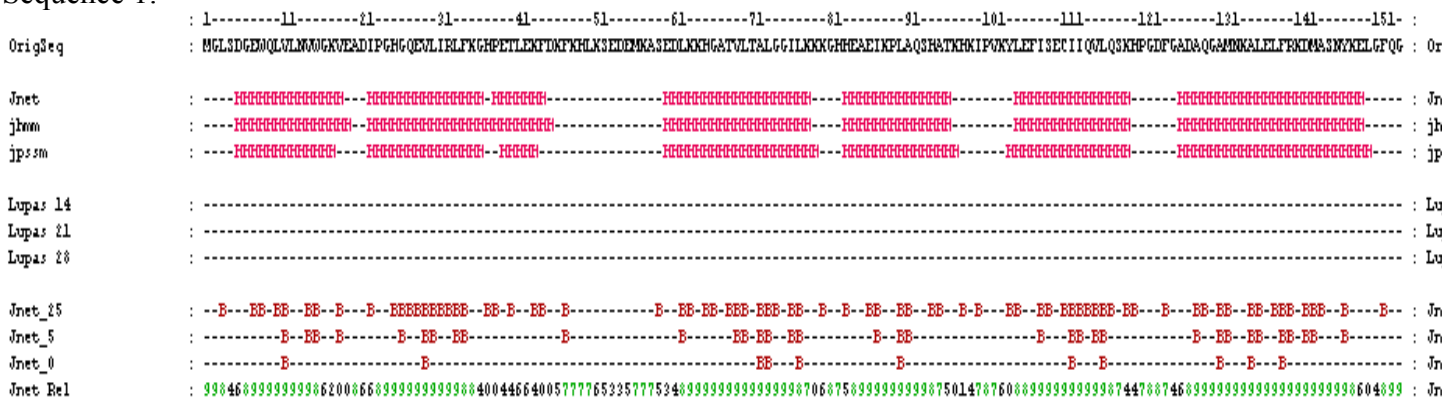
Colour code for alignment:
Blue - Complete identity at a position
Shades of red - The more red a position is, the higher the level of conservation of chemical properties of the amino acids
Jnet - Final secondary structure prediction for query
jalign - Jnet alignment prediction
jhmm - Jnet hmm profile 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.

Sequence 1:



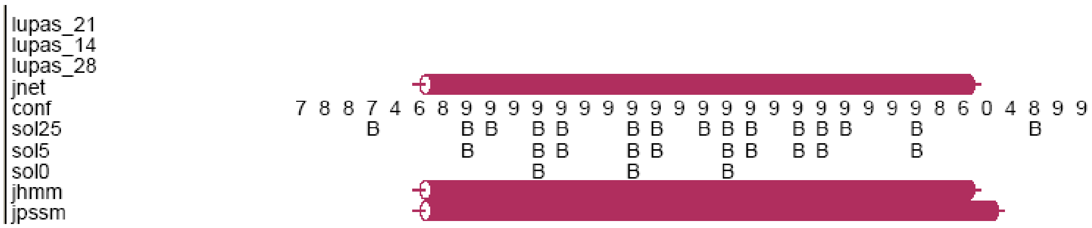
Jnet : ----HHHHHHHHHHHHHHH---HHHHHHHHHHHHHHHH-HHHHHHHH----------
-----HHHHHHHHHHHHHHHHHHHH----HHHHHHHHHHHHHHH--------
HHHHHHHHHHHHHHHH------HHHHHHHHHHHHHHHHHHHHHHHHH----- : Jnet
jhmm : ----HHHHHHHHHHHHHHHH--HHHHHHHHHHHHHHHHHHHHHHHHH---
----------HHHHHHHHHHHHHHHHHHH----HHHHHHHHHHHHHHH--------
HHHHHHHHHHHHHHHH------HHHHHHHHHHHHHHHHHHHHHHHHH----- : jhmm
jpssm : ----HHHHHHHHHHHHHH----HHHHHHHHHHHHHHHH--HHHHH----------
---HHHHHHHHHHHHHHHHHHHHH---HHHHHHHHHHHHHHHH------
HHHHHHHHHHHHHHHH------HHHHHHHHHHHHHHHHHHHHHHHHH---- : jpssm

Lupas 14 : ----------

Lupas 21 : ----------

Lupas 28 : ----------

Jnet_25 : --B---BB-BB--BB--B---B--BBBBBBBBBB--BB-B--BB--B----------B--BB-
BB-BBB-BBB-BB--B--B--BB--BB--BB--B-B---BB--BB-BBBBBBB-BB---B---BB-BB--BB-
BBB-BBB--B----B--
Jnet_5 : ----------B--BB--B-------B--BB--BB----------B----------B------BB-BB--
BB---------B--BB----------B---BB-BB----------B--BB--BB--BB-BB---B-------
Jnet_0 : ----------B----------B--BB---B----------
B--------------------B---B----------B---B---B---------
Jnet Rel :
998468999999998620086689999999999884004466400577776533577753489999999999999870687589999999998750147876088999999999987447887468999999999999999999998604899



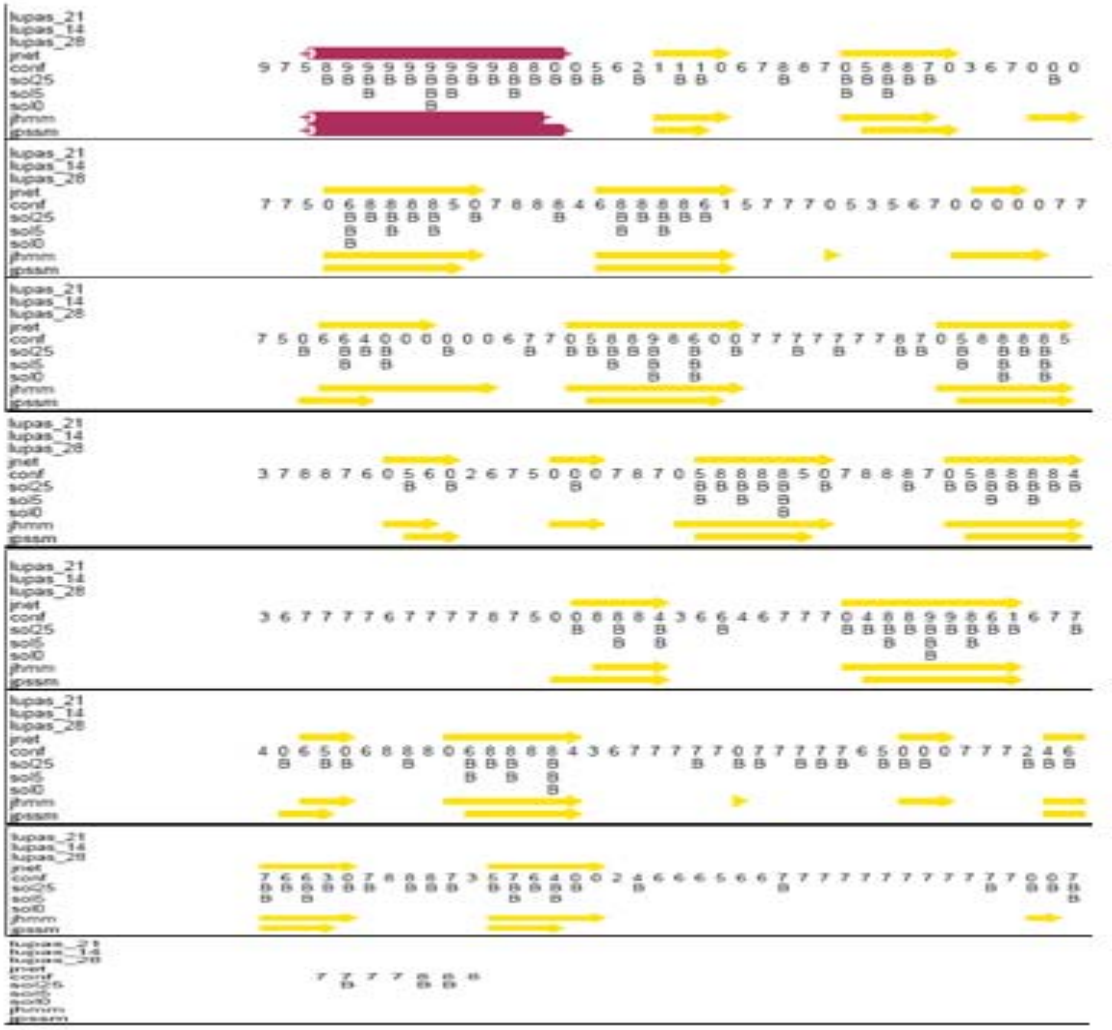
Inference:
Proline residues position is identical in template and query sequence. Histidine is highly conserved.
Less than 50% coils are predicted.

Sequence 2:
Jnet : --HHHHHHHHHHHHH----EEEE-----EEEEEE---------EEEEEEEE-----EEEEEEE-----------EEE------EEEEEE------EEEEEEEEE---------EEEEEEE------EEEE----EEE----EEEEEEE-----EEEEEEE---------------EEEEE--------EEEEEEEEE-----EEE----EEEEEEE---------------EEE----EEEEEEE------EEEEEE------------------------------ : Jnet
jhmm : --HHHHHHHHHHHH-----EEEE-----EEEEE----EEE---EEEEEEEE-----EEEEEEE----E-----EEEEE-----EEEEEEEEE---EEEEEEEEE---------EEEEEEE------EEE-----EEE---EEEEEEEE-----EEEEEEE----------------EEEE--------EEEEEEEEE-----EEE----EEEEEEE-------E-------EEE----EEEEEEE------EEEEEE--------------------EE-------- : jhmm
jpssm : --HHHHHHHHHHHHHH----EEE-------EEEEE---------EEEEEEE------EEEEEEE-------------------EEEE----------EEEEEEE-----------EEEEEE-------EEE-----------EEEEEE-------EEEEEE--------------EEEEEE---------EEEEEEEE----EEE------EEEEEE----------------------EEEEEE-------EEEE------------------------------- : jpssm

Lupas 14 : -- : Lupas 14
Lupas 21 : -- : Lupas 21
Lupas 28 : -- : Lupas 28

Jnet_25 : ---BBBBBBBBBBBBBBB-B-BB---B--BBBBB-----B-----BBBBB-B---B--BBBBB--------------------B-BBB--B---B-BBBBB-B-B--B-B--BB-B-BBB--------B-B-----B-----BBBBB-B---B-BBBBBBB---------------B-B-B--B-----BBBBBBBBB--B-B-BB--B--BBB-B------B-BB-BBB-BBB----BBBBBBBBB-BBB-BBBBB--B------B---------B-BBB-B--BB- : Jnet_25
Jnet_5 : -----B--BB--B---------------B-B-------------B-B-B--------B-B-----------------------B-B----------B-B-B------------B-B-B----------------------B-B-B---------B-B-------------------B-B----------B-B-B---------------B-B-B-------------------------B-B---------B-B------------------------B------- : Jnet_5
Jnet_0 : --------B----------------------------------B---B-B--------------B-B-------------------------B---B---------------------B--- : Jnet_0

Jnet Rel :
97589999999988005621110678870588703670007750688885078884688886157770535670000077750664000000677058898600777777787058888537887605602675000787058888507888705888843677776777787500888436646777048899861677406506888068888436777770777776500077724676630788873576400246665667777777777770077777888 : Jnet Rel



Inference:
Less than 50% Strands is predicted

2. Predict the secondary structure composition of O13837.

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 :

O13837: 4-aminobutyrate aminotransferase

>gi|6016100|sp|O13837.1|GABAT_SCHPO RecName: Full=4-aminobutyrate aminotransferase; AltName: Full=GABA aminotransferase; Short=GABA-AT; AltName: Full=Gamma-amino-N-butyrate transaminase; Short=GABA transaminase

MSSTATVTESTHFFPNEPQGPSIKTETIPGPKGKAAAEEMSKYHDISAVKFPVDYEKSIGN
YLVDLDGNVLLDVYSQIATIPIGYNNPTLLKAAKSDEVATILMNRPALGNYPPKEWARV
AYEGAIKYAPKGQKYVYFQMSGSDANEIAYKLAMLHHFNNKPRPTGDYTAEENESCLN
NAAPGSPEVAVLSFRHSFHGRLFGSLSTTRSKPVHKLGMPAFPWPQADFPALKYPLEEH
VEENAKEEQRCIDQVEQILTNHHCPVVACIIEPIQSEGGDNHASPDFFHKLQATLKKHDV
KFIVDEVQTGVGSTGTLWAHEQWNLPYPPDMVTFSKKFQAAGIFYHDLALRPHAYQHF
NTWMGDPFRAVQSRYILQEIQDKDLLNNVKSVGDFLYAGLEELARKHPGKINNLRGKG
KGTFIAWDCESPAARDKFCADMRINGVNIGGCGVAAIRLRPMLVFQKHHAQILLKKIDE
LI

Structure Prediction:

Jnet : ----------HHHHHHHHHHHH---------EEEEE---EEEEE----
EEEHHH--HHHHH-----HHHHHHHHHHHHHH----------HHHHHHHHHHHHHHH------EEEE-
----HHHHHHHHHHHHHHHHH----------HHHHHHHHH---------EEEEEE----------
---E----------HHHHHHHHHHHHHHHH-----EEEEEE----------HHHHHHHHHHHHHH---
EEEEE----------HHHHHHHHH---HHHHHHH----------HHHHHHHHHHHHHH-
---HHHHHHHHHHHHHHHHHHHHHH----------EEEEEEEE---HHHHHHHHHHHHH--
EEEEE----EEEEE------HHHHHHHHHHHH-- : Jnet
jhmm : ----------HHHHHHHHHH----------EEEEE---EEEEE----
HHHHHH--HHHHH-----HHHHHHHHHHHH----------HHHHHHHHHHHHHHH------
EEEE-----HHHHHHHHHHHHHHHH----------HHHHHHHH----------EEEEEE-------HHHH------
----------EE----------HHHHHHHHHHHH-----EEEEEE----------
HHHHHHHHHHHHH---EEEEEHHHH----------EEEEE------HHHHHHH----------
HHHHHHHHHHHHHHH---HHHHHHHHHHHHHHHHHHHHH----------EEEEEEEEE-----
HHHHHHHHHH--EEEEE----EEEEE------HHHHHHHHHHHH-- : jhmm
jpssm : ----------HHHHHHHHHHHH---------EEEEE---EEEEE----
EEEEE---HHHHH------HHHHHHHHHHHH----------HHHHHHHHHHHHH------EEEEE---
-HHHHHHHHHHHHHHHH----------HHHHHHHHH---------EEEEEE----------E----------
-EE----------HHHHHHHHHHHHHHHHH-----EEEEEEE----------HHHHHHHHHHHHH---
EEEEE----------HHHHHHHHHH----HHHHHH----------HHHHHHHHHHH---
-HHHHHHHHHHHHHHHHHHHHHH----EEE-------EEEEEEE---HHHHHHHHHHHHH---
EEEE----EEEE-------HHHHHHHHHHHH-- : jpssm

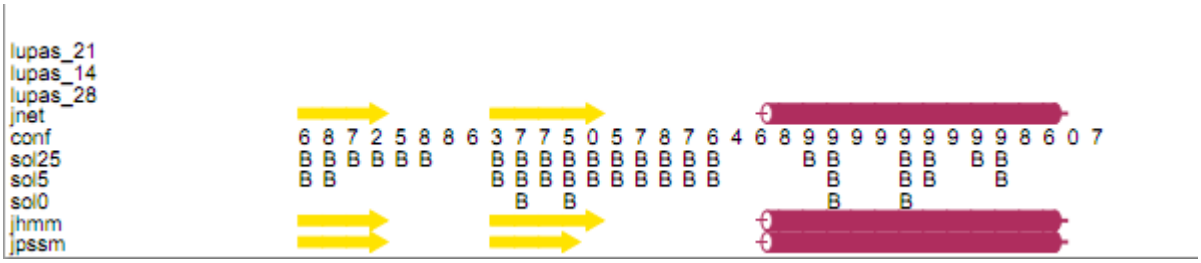
Lupas 14 : ----------

---------- : Lupas 14
Lupas 21 : ----------

---------- : Lupas 21
Lupas 28 : ----------

---------- : Lupas 28

Jnet_25 : ----B-------BBB-----B-B-B--B-B--B--BB--B--BBBBBB--BBBBB--B--
BBBBB-----BBBBBBBBBBBBBBBB--BB-BBB-BB--BBBBBBBBBBBBBBB-BB--BB-
BBB-BB---B-BBBBBBBBBBBBBBBBBBBBBBB---------B--BBBBBBB-BB-----B-
BBBBBBBBBBBBBBBBBBBBB---B---B-BBB-BBBBBBBBBBB----B--BB-BB--BB--B--
BB------BBBBBBBBBBBBBBBBB---BB--BB-BB--B-
BBBBBBBBBBBBBBBBBBBBBBBBB-B-BBBBBBBBBBBBBBBBBBBBBB-B--
BBBBBBBBBBBBBBBBBBBB-BB----BB--B--BB--BB--B--BB--B--BB-
BBBBBBBBBBBBBB----BB--BB-BBB-BBBBBBBBB--BBBBBBBBBB---BB--BB-BB---- :
Jnet_25
Jnet_5 : ------------------------------B--B----------B----BBB--B--B-B-------
BBBBBBBBBBBBBB-----B---B---B---B-BBBBBB-------BB--BB--B----------BBBB-BBB-
BBBBBB-BB----------------------B----------BBBB-------BBBBBBBB-----------B--B--B-BB--B--
----------B--B---B--B-------BBBBBBBBBBB--B-------BB--B---B----BBBBBB-B-BBB-BBB--
BBBB-------B-BBBBB--BBBBBBBBBBBBB----B-B-BB-BB-BBBBBBB-BBB--B--------B-----
-B---B--B-------B----B-B-BBBBBB-------B--BB--B---BBBB------BBBBBBBBBB----B--BB--B--
-- : Jnet_5
Jnet_0 : --B--------BBBB----B--------
----------------B-----------------B-----------B------B---BB-BB----------------------------------BB-----
------BB---BBBBB-B----------------B---B------
BBB--B--B---BB---B-----------B-B-B-------B----------------------B-BBB--B-----------B------B-----
-----------------B--------------B---B-------------B-B----------B--B-------- : Jnet_0
Jnet Rel :
998654677777776777777642577877507899999874036787887458985584689954884000000123
33414888703899999998703665677666677006899999999874478883577606775689999999998
6104677877765178888870267877872588860577777700000367777777777777770000157777
7777775000099999999998615888548888600577777776368999999998745864788500007777777
7654224677777500000000005000454400046764567777870001678999987308866899999999
99999998743784000268880058898850650099999998872570687258863775057876468999999
998607 : Jnet Rel



Inference:
P: blue
H: red
C: highly conserved.
Structure: Consists of both alpha helices and extended beta sheets

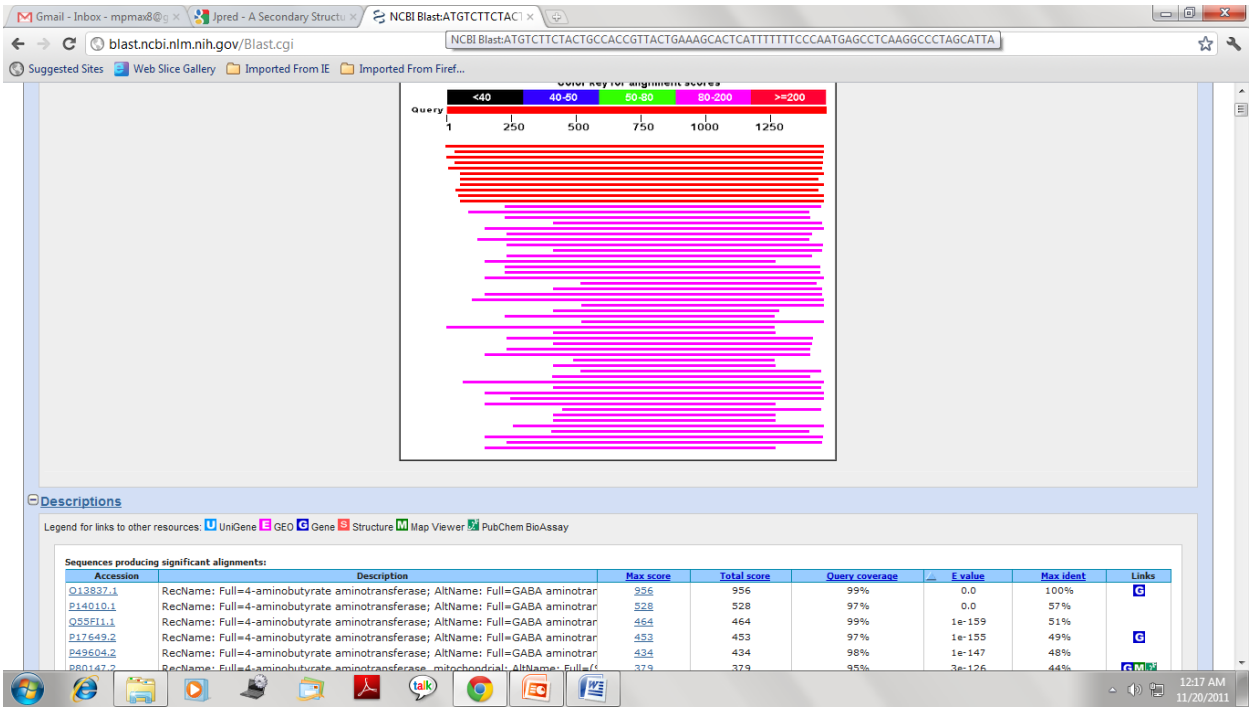
3. Find the secondary structure of the given sequence and compare with the output of 2.

Method:

1. Run Blastx to determine protein.
2. Predict Secondary Structure.
3. Go to http://www.compbio.dundee.ac.uk/www-jpred/
4. Paste the sequence and click on make prediction
5. Wait for the software to predict the structure
6. Once Job is done . Save the output.

>
ATGTCTTCTACTGCCACCGTTACTGAAAGCACTCATTTTTTTCCCAATGAGCCTCAAGGCCCTAGCATTA
AGACCGAAACTATTCCCGGTCCCAAAGGTAAGGCCGCTGCTGAAGAAATGTCCAAATACCACGACATCAG
CGCTGTCAAGTTTCCTGTAGACTATGAAAAGTCCATTGGTAACTATCTTGTCGACTTGGATGGTAACGTT
CTCTTGGATGTTTACTCTCAAATCGCTACTATCCCCATTGGCTACAACAATCCTACTCTCCTCAAGGCTG
CCAAGTCGGACGAAGTCGCTACCATTTTAATGAACCGTCCTGCTTTGGGAAATTACCCTCCTAAGGAATG
GGCTCGTGTCGCTTATGAGGGTGCCATCAAATATGCCCCCAAGGGTCAAAAGTATGTTTACTTTCAAATG
AGTGGAAGTGATGCCAACGAGATTGCTTACAAGCTTGCTATGCTTCATCATTTCAACAACAAGCCTAGAC
CTACTGGTGATTACACTGCTGAAGAGAACGAGAGCTGCTTAAACAACGCTGCTCCTGGATCTCCCGAAGT
TGCTGTTCTCTCTTTCCGTCACTCTTTCCACGGACGTCTCTTTGGTTCTCTTTCCACTACTCGCTCCAAG
CCTGTTCACAAGCTTGGTATGCCTGCTTTCCCATGGCCTCAAGCTGATTTCCCTGCTTTGAAGTATCCTT
TGGAAGAGCACGTCGAAGAGAATGCAAAGGAGGAGCAACGCTGCATTGACCAGGTCGAGCAAATTTTAAC
TAACCACCATTGCCCTGTCGTTGCCTGTATCATTGAGCCCATTCAATCTGAGGGTGGTGACAACCATGCC
TCTCCTGACTTTTTCCACAAGCTTCAAGCTACTTTGAAGAAGCATGATGTCAAGTTTATCGTCGATGAAG
TCCAAACTGGTGTCGGCTCTACCGGTACTTTATGGGCTCACGAGCAATGGAATTTACCCTATCCTCCTGA
CATGGTTACCTTTTCCAAGAAATTCCAGGCTGCCGGTATTTTCTATCATGATTTGGCTCTTCGTCCTCAT
GCTTATCAGCACTTCAATACTTGGATGGGTGACCCATTCCGTGCTGTTCAATCTAGATATATTCTTCAAG
AAATTCAAGACAAGGATCTCCTTAATAACGTCAAGTCTGTTGGCGATTTCTTGTATGCTGGACTTGAAGA
GCTTGCTCGTAAGCACCCTGGCAAAATCAACAACCTCCGCGGTAAGGGAAAGGGTACTTTTATCGCTTGG
GATTGTGAGTCTCCTGCAGCCCGTGACAAATTCTGTGCTGACATGAGAATTAATGGTGTCAACATTGGTG
GCTGTGGTGTAGCTGCTATTCGTCTTCGTCCTATGCTTGTATTCCAAAAGCACCATGCTCAAATCCTTCT
CAAGAAGATTGACGAATTGATTTA

Results:



Inference:

Schizosaccharomyces pombe chromosome I, complete replicon
Length=5579133

Features in this part of subject sequence:
4-aminobutyrate aminotransferase (GABA transaminase)

Score = 2632 bits (1425), Expect = 0.0
Identities = 1425/1425 (100%), Gaps = 0/1425 (0%)
Strand=Plus/Plus

Corresponding protein: O13837: 4-aminobutyrate aminotransferase

>gi|6016100|sp|O13837.1|GABAT_SCHPO RecName: Full=4-aminobutyrate aminotransferase; AltName: Full=GABA aminotransferase; Short=GABA-AT; AltName: Full=Gamma-amino-N-butyrate transaminase; Short=GABA transaminase

Secondary structures are the same.

Experiment 8: 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

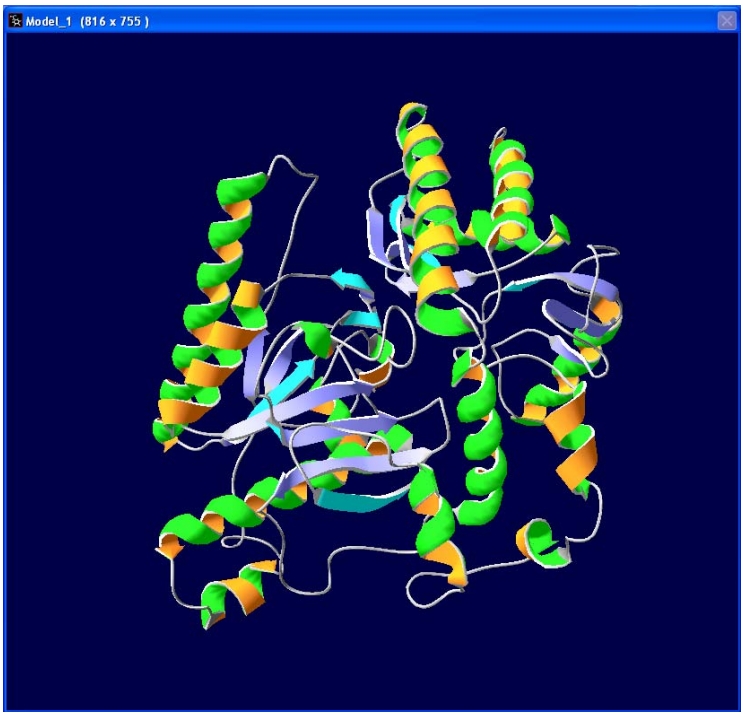
MASMLLAQRLACSFQHSYRLLVPGSRHISQAAAKVDVEFDYDGPLMKTEVPGPRSQEL
MKQLNIIQNAEAVHFFCNYEESRGNYLVDVDGNRMLDLYSQISSVPIGYSHPALLKLIQQ
PQNASMFVNRPALGILPPENFVEKLRQSLLSVAPKGMSQLITMACGSCSNENALKTIFM
WYRSKERGQRGFSQEELETCMINQAPGCPDYSILSFMGAFHGRTMGCLATTHSKAIHKI
DIPSFDWPIAPFPRLKYPLEEFVKENQQEEARCLEEVEDLIVKYRKKKKTVAGIIVEPIQSE
GGDNHASDDFFRKLRDIARKHGCAFLVDEVQTGGGCTGKFWAHEHWGLDDPADVMT
FSKKMMTGGFFHKEEFRPNAPYRIFNTWLGDPSKNLLLAEVINIIKREDLLNNAAHAGK
ALLTGLLDLQARYPQFISRVRGRGTFCSFDTPDDSIRNKLILIARNKGVVLGGCGDKSIRF
RPTLVFRDHHAHLFLNIFSDILADFK

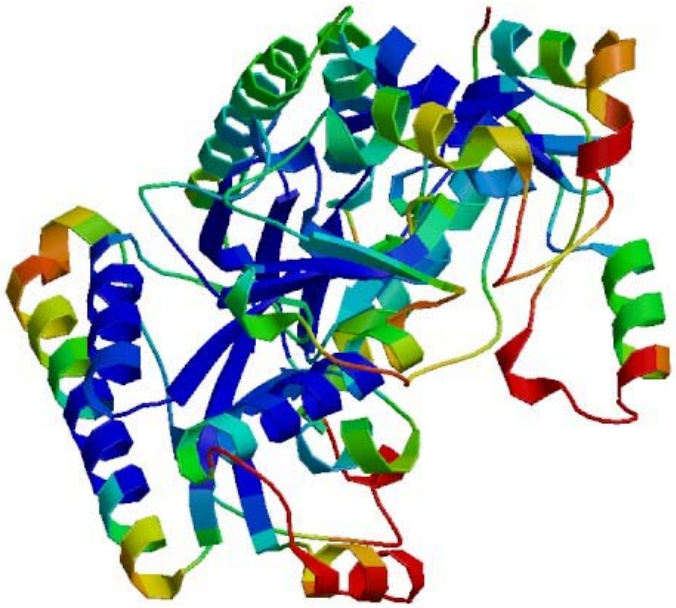
Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value
1OHV_A	Chain A, 4-Aminobutyrate-Aminotransferase From Pig >pdb|1OHV|B Chair	959	959	94%	0.0
2JJF_A	Chain A, N328a Mutant Of M. Tuberculosis Rv3290c	166	166	85%	5e-46
2CIN_A	Chain A, Lysine Aminotransferase From M. Tuberculosis In The Internal Al	166	166	85%	6e-46

```
>pdb|1OHV|A S Chain A, 4-Aminobutyrate-Aminotransferase From Pig
 pdb|1OHV|B S Chain B, 4-Aminobutyrate-Aminotransferase From Pig
 pdb|1OHV|C S Chain C, 4-Aminobutyrate-Aminotransferase From Pig
  ▶9 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   DGNRMLDLYSQISSIPIGYSHPALVKLVQQPQNVSTFINRPALGILPPENFVEKLRESLL  120

Query  149  SVAPKGMSQLITMACGSCSNENALKTIFMWYRSKERGQRGFSQEELETCMINQAPGCPDY  208
            SVAPKGMSQLITMACGSCSNENA KTIFMWYRSKERGQ  FS+EELETCMINQAPGCPDY
Sbjct  121  SVAPKGMSQLITMACGSCSNENAFKTIFMWYRSKERGQSAFSKEELETCMINQAPGCPDY  180

Query  209  SILSFMGAFHGRTMGCLATTHSKAIHKIDIPSFDWPIAPFPRLKYPLEEFVKENQQEEAR  268
            SILSFMGAFHGRTMGCLATTHSKAIHKIDIPSFDWPIAPFPRLKYPLEEFVKENQQEEAR
Sbjct  181  SILSFMGAFHGRTMGCLATTHSKAIHKIDIPSFDWPIAPFPRLKYPLEEFVKENQQEEAR  240

Query  269  CLEEVEDLIVKYRKKKKTVAGIIVEPIQSEGGDNHASDDFFRKLRDIARKHGCAFLVDEV  328
            CLEEVEDLIVKYRKKKKTVAGIIVEPIQSEGGDNHASDDFFRKLRDI+RKHGCAFLVDEV
Sbjct  241  CLEEVEDLIVKYRKKKKTVAGIIVEPIQSEGGDNHASDDFFRKLRDISRKHGCAFLVDEV  300

Query  329  QTGGGCTGKFWAHEHWGLDDPADVMTFSKKMMTGGFFHKEEFRPNAPYRIFNTWLGDPSK  388
            QTGGG TGKFWAHEHWGLDDPADVMTFSKKMMTGGFFHKEEFRPNAPYRIFNTWLGDPSK
Sbjct  301  QTGGGSTGKFWAHEHWGLDDPADVMTFSKKMMTGGFFHKEEFRPNAPYRIFNTWLGDPSK  360

Query  389  NLLLAEVINIIKREDLLNNAAHAGKALLTGLLDLQARYPQFISRVRGRGTFCSFDTPDDS  448
            NLLLAEVINIIKREDLL+NAAHAGK LLTGLLDLQARYPQFISRVRGRGTFCSFDTPD+S
Sbjct  361  NLLLAEVINIIKREDLLSNAAHAGKVLLTGLLDLQARYPQFISRVRGRGTFCSFDTPDES  420

Query  449  IRNKLILIARNKGVVLGGCGDKSIRFRPTLVFRDHHAHLFLNIFSDILADFK  500
            IRNKLI IARNKGV+LGGCGDKSIRFRPTLVFRDHHAHLFLNIFSDILADFK
Sbjct  421  IRNKLISIARNKGVMLGGCGDKSIRFRPTLVFRDHHAHLFLNIFSDILADFK  472
```

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





Energy minimization score: -26789.707

RMSD: 0.07A

Quality information:
QMEAN Z-Score: -1.129

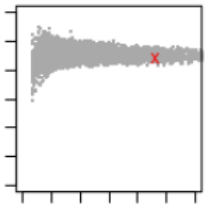
Ligand information:

Global Model Quality Estimation:

QMEAN4 global scores:

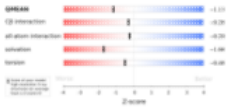
QMEANscore4: 0.695

Estimated absolute model quality:



Z-Score: -1.129

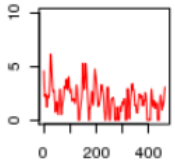
Score components:



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.

Experiment 9: 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
VHFFC
NYEESRGNYLVDVDGNRMLDLYSQISSVPIGYSHPALLKLIQQPQNASMFVNRPALGILPPENFVEKLRQ
SLLSV
APKGMSQLITMACGSCSNENALKTIFMWYRSKERGQRGFSQEELETCMINQAPGCPDYSILSFMGAFHGR
TMGCL
ATTHSKAIHKIDIPSFDWPIAPFPRLKYPLEEFVKENQQEEARCLEEVEDLIVKYRKKKKTVAGIIVEPI
QSEGG
DNHASDDFFRKLRDIARKHGCAFLVDEVQTGGGCTGKFWAHEHWGLDDPADVMTFSKKMMTGGFFHKEEF
RPNAP
YRIFNTWLGDPSKNLLLAEVINIIKREDLLNNAAHAGKALLTGLLDLQARYPQFISRVRGRGTFCSFDTP
DDSIR
NKLILIARNKGVVLGGCGDKSIRFRPTLVFRDHHAHLFLNIFSDILADFK*

>P1;2oatA
structure:2oatA:   28: :  404: :::-1.00:-1.00
-------------------------------------------------------------------
-----
--ERGKGIYLWDVEGRKYFDFLSSYSAVNQGHCHPKIVNALKSQVDKLTLTSRAVLG--EYEEYITKL--
-----
--FNYHKVLPMNTGVEAGETACKLARKW---------GYTVKGIQKYKA---------
KIVFAAGNFWGRTLSAI
SSS-------TDPTSYD-GFGPF----MPGFDIIPYND------LPALERAL-----
QDPNVAAFMVEPIQGEAG
VVVPDPGYLMGVRELCTRHQVLFIADEIQTGLARTGRWLAVDYENV--RPDIVLLG-
KALSGGLYDDDIMLTIKP
GEHGSTYGGNPLGCRVAIAALEVLEEENLAENADKLGIILRNELMKLPS---
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)

```
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:

                                                        .--- 2oatA @2.0    72.0000
                                                       |
                                                    .--+-- 1sffA @1.9     72.5000
                                                   |
                                 .-----------------+------ 1d7uA @2.0     78.0000
                                |
                  .-------------+----------------------- 1s0aA @1.7     80.5000
                 |
       .---------+------------------------------------- 2gsaA @2.4     83.0625
      |
   .--+---------------------------------------------- 1ohvA @2.3

   +----+----+----+----+----+----+----+----+----+----+----+----+----+
83.5050   81.5137   79.5225   77.5312   75.5400   73.5488   71.5575
     82.5094   80.5181   78.5269   76.5356   74.5444   72.5531
```

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.**

Experiment 10: Protein- Ligand docking using Glide protocol

Aim:

To quantify the interaction of the ligand with the protein target using Glide protocol of Schrodinger package.

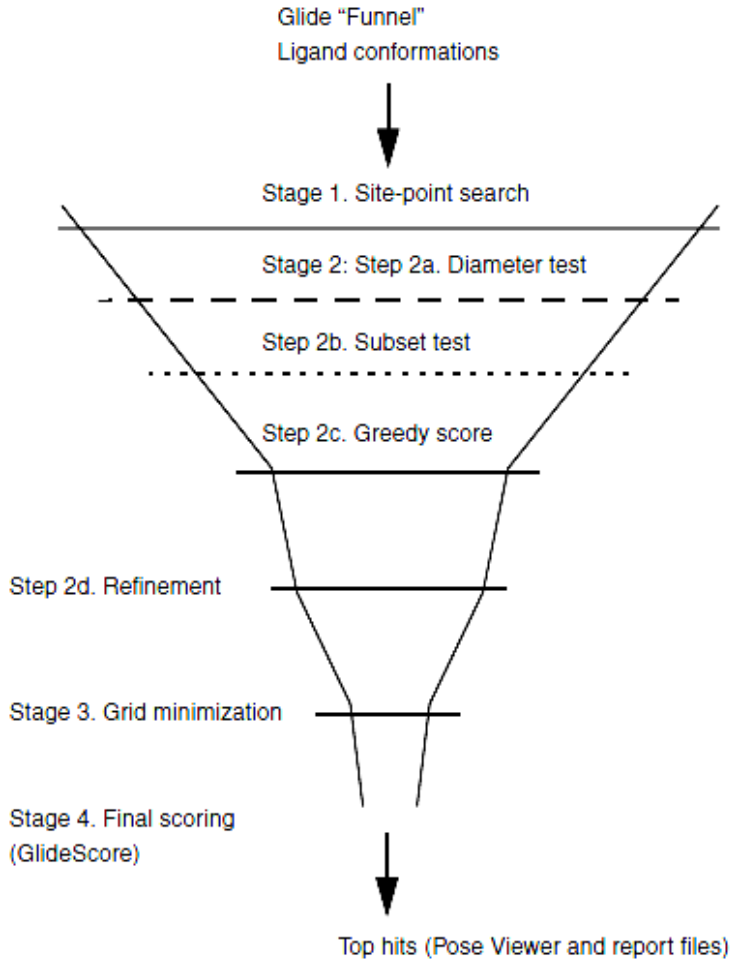
Introduction:

Glide is Grid-based Ligand Docking with Energetics which searches for favorable interactions between one or more ligand molecules and a receptor molecule, usually a protein. Each ligand must be a single molecule, while the receptor may include more than one molecule, e.g., a protein and a cofactor. Glide can be run in rigid or flexible docking modes; the latter automatically generates conformations for each input ligand. The combination of position and orientation of a ligand relative to the receptor, along with its conformation in flexible docking, is referred to as a *ligand pose*.

The ligand poses that Glide generates pass through a series of hierarchical filters that evaluate the ligand's interaction with the receptor. The initial filters test the spatial fit of the ligand to the defined active site, and examine the complementarity of ligand-receptor interactions using a grid-based method patterned after the empirical ChemScore function. Poses that pass these initial screens enter the final stage of the algorithm, which involves evaluation and minimization of a grid approximation to the OPLS-AA nonbonded ligand-receptor interaction energy.

Final scoring is then carried out on the energy-minimized poses. By default, Schrödinger's proprietary GlideScore multi-ligand scoring function is used to score the poses. If GlideScore was selected as the scoring function, a composite *Emodel* score is then used to rank the poses of each ligand and to select the poses to be reported to the user. Emodel combines GlideSore, the nonbonded interaction energy, and, for flexible docking, the excess internal energy of the generated ligand conformation. Glide uses a hierarchical series of filters to search for possible locations of the ligand in the active-site region of the receptor. The shape and properties of the receptor are represented on a grid by several different sets of fields that provide progressively more accurate scoring of the ligand poses. Conformational flexibility is handled in Glide by an

extensive conformational search, augmented by a heuristic screen that rapidly eliminates unsuitable conformations, such as conformations that have long-range internal hydrogen bonds.



The Glide docking hierarchy.

Method:

Receptor Grid Generation

Import your protein structure onto the Workspace

Applications → Glide → Receptor Grid Generation

↓

Receptor

Define receptor:

Select ligand from workspace to exclude it from the grid

↓

Site

Enclosing box: Centroid of workspace ligand

↓

Constraints and Rotatable Groups

Default setting

↓

Start the job

↓

The output grid file will be generated in the format: grid-out.maegz

This grid file will be used for all further operations

Glide: Ligand Docking

Applications → Glide → Ligand Docking

↓

Settings

Receptor Grid: Specify the receptor grid
(Obtained from the grid generation step)
Docking: SP (Standard Precision)

↓

Ligands

Ligands to be docked: Import the ligands (LigPrep output)

↓

Core, Constraints, Torsional Constraints and Similarity

Default settings

↓

Output

Write pose viewer file and
Perform post-docking minimizations

↓

Start the job

↓

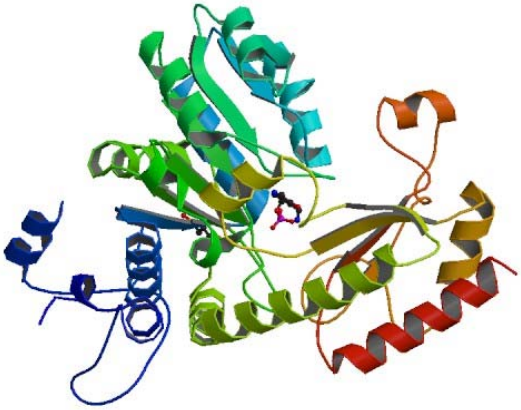
View Results

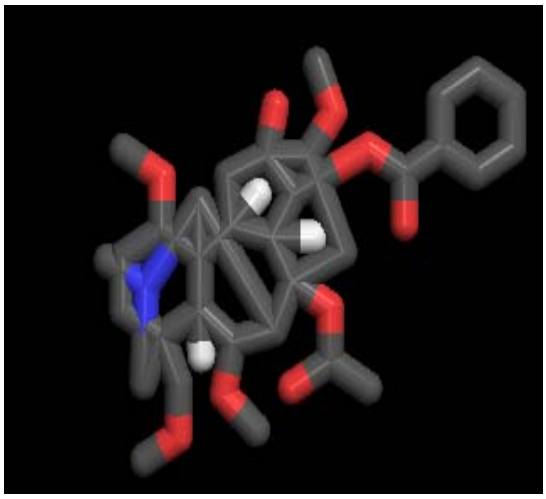
Project → Show table

Analyze the poses ranked based on GlideScore and DockScore
Visualize hydrogen bonding interactions

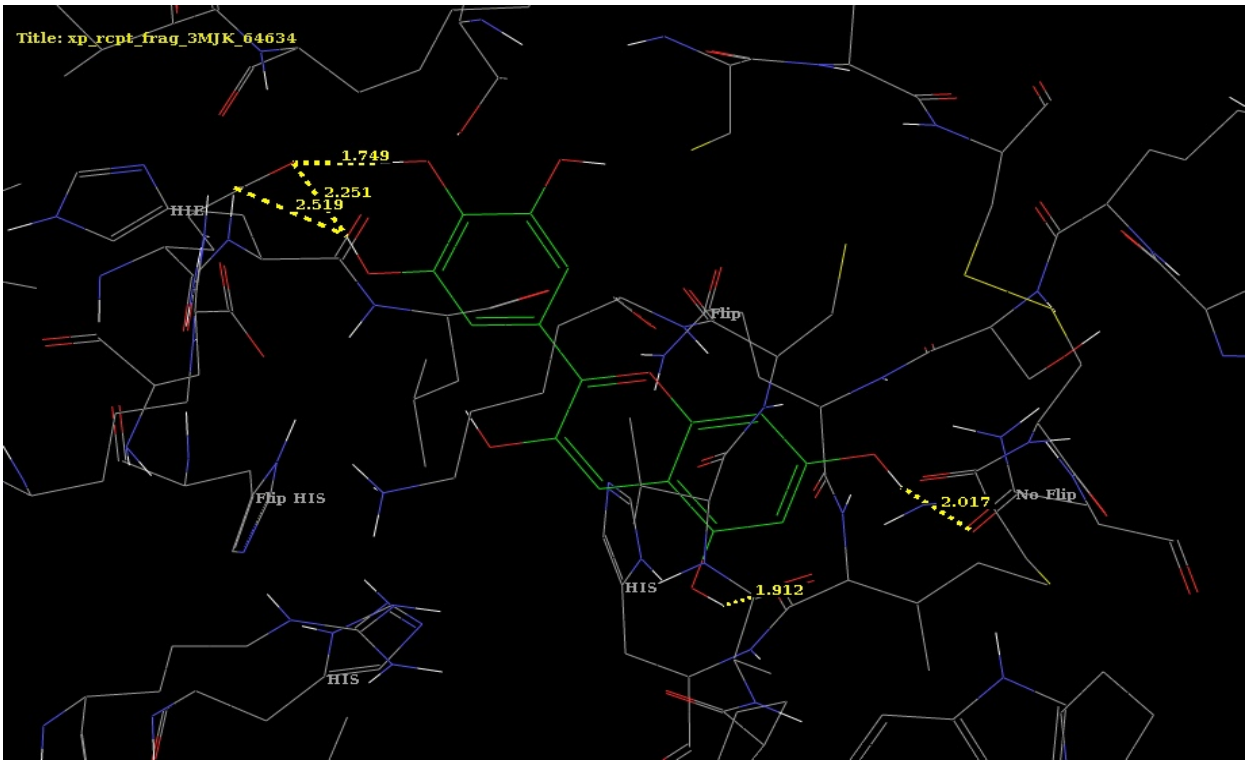
Result:

Protein - 3MKJ **Ligand – Delphinine**





Receptor and Ligand Interaction



3MKJ – Delphinine were successfully docked using glide .