CHAPTER – 3
MATERIALS AND METHODS
3. MATERIALS AND METHODS

3.1 SEQUENCE ALIGNMENT

Sequence alignment has become an essential part of biological science. There are now many different techniques and implementations of methods to perform alignment of sequences. Early in the days of protein and gene sequence analysis [73]. It was discovered that the sequences from related proteins or genes were similar, in the sense that one could align the sequences so that many corresponding residues match. This discovery was very important, since strong similarity between two genes is a strong argument for their homology. Bioinformatics is based on it. There are many features of sequence alignments that give interesting information. For example, a closer analysis of the alignment can reveal which parts of the sequences that is likely to be important for the function, if the proteins are involved in similar processes. In parts of the sequence of a protein which are not very critical for its function, the random mutations can accumulate more easily. In parts of the sequence that are critical for the function of the protein, hardly any mutations will be accepted; nearly all changes in such regions will destroy the function.

The basis for comparison of proteins and genes using the similarity of their sequences is that the proteins or genes are related by evolution; they have a common ancestor. Random mutations in the sequences accumulate over time, so that proteins or genes that have a common ancestor far back in time are not as similar as proteins or genes that diverged from each other more recently. Analysis of evolutionary relationships between protein or gene sequences depends critically on sequence alignments [74].
3.1.1 The Evolutionary aspects of Sequence Alignment

One goal of sequence alignment is to enable the researcher to determine whether two sequences display sufficient similarity such that an inference of homology is justified. Although these two terms are often interchanged in popular usage, let us distinguish them to avoid confusion in the current discussion. Similarity is an observable quantity that might be expressed as, say, percent identity or some other suitable measure. Homology, on the other hand, refers to a conclusion drawn from these data that two genes share a common evolutionary history [75]. Genes either are or are not homologous—there are no degrees for homology as there are for similarity.

3.1.2 Goals of Sequence Analysis

Sequences analysis is used to identify the genes which determines the function of each gene. One way to hypothesize the function is to find another gene possibly from another organism whose function is known and to which the new gene has high sequence similarity. This assumes that sequence similarity implies functional similarity, which may or may not be true. But the identification of proteins involved in the regulation of gene expression is mainly. Observed and confirmed by sequence repeats, which also identifies the other functional regions, for example origins of replication (sites at which DNA polymerase binds and begins replication; For example, pseudogenes (sequences that look like genes but are not expressed), sequences responsible for the compact folding of DNA, and sequences responsible for nuclear anchoring of the DNA.
Many of these tasks are computational in nature. Given the incredible rate at which sequence data is being produced, the integration of computer science, mathematics, and biology will be integral to analyzing those sequences.

### 3.1.3 Sequence Similarity

In molecular biology, a common question is to ask whether or not two sequences are related. The most common way to tell whether or not they are related is to compare them to one another to see if they are similar. If we look at two words in the English language, we note that two words that are spelled similarly may mean two completely different things, such as the words pear and tear. Biological sequences that are similar but not exact provide useful information to help discover functional, structural, and evolutionary information. Two sequences in different organisms are homologous if they have been derived from a common ancestor sequence [76]. Two sequences may or may not be homologous regardless of their sequence similarity. However, the greater the sequence similarity, the greater chance there is that they share similar function and or structure.

### 3.1.4 Steps in Alignment of Sequences of DPP-IV

Sequence analysis was performed using two methods, they are: pair wise sequence alignment and multiple sequence alignment. In pair wise sequence alignment our query sequence is aligned with the subject sequence. It is mainly done by the NCBI-BLAST [77], WU-BLAST [78] and EBI-FASTA [79] tools. And the multiple sequence alignment is done by using the EBI-CLUSTAL W [80] tool.
Pair wise alignment is used to compare a novel sequence with those contained in nucleotide and protein databases by aligning the novel sequence with previously characterized genes. The emphasis of this tool is to find regions of sequence similarity, which will yield functional and evolutionary clues about the structure and function of this novel sequence. Regions of similarity detected via this type of alignment tool can be either local, where the region of similarity is based in a particular location, or global, where regions of similarity can be detected across otherwise unrelated genetic code.

3.1.5 Protein Sequence Selection

DPP-IV protein is identified from SWISS PROT database. Six proteins are observed, they are P81425, Q9N2I7, P27487, P28843, P22411 and P14740. BLAST program, protein - protein blastp, from NCBI was selected to scan the query protein sequence against pdb structure database.

Swiss prot ID P81425 was scanned against pdb data base. Similarly the other five sequences Q9N2I7, P27487, P28843, P22411 and P14740are also scanned.
The six sequences in fasta format are as given below.

>P81425|DPP-IV_BOVIN Dipeptidyl peptidase 4 - Bos taurus (Bovine).
MKTPKWLLGLLALAIALVTVITPVVLLTKGDASTDSRRTYTLDYLKN1FRMKFYLNLRW5DHELYLKQ
ENNILLFNAEYGNSSIFLENSTDFEDEFHSINDYSVSPDQYIFLFYNYVQKWRSHTASYDIYDLNKRQLI
TEEIPNNTQWITSSVGHKLAYVWNDIYVKNEPNSPSQRTWTGKDKVINEGTDWYEEEEVSAYSAL
WWSPNSTFLAYAQNDFVEDVPIEYSFDESLEQYPKTVKIPKAGAVNPIKFFVNVINSSLSPNINATSQ
QIVPPGSVLGHDLYCLDVVTWEISLQLWRLRIQNSIMIDCYDRTSRGWISSVGRQHIEHSTTGWVGR
FRPAEPHTSDGSNFYKIISNEEGYKICHIPQFTDKRNCTFITKGAWEVIGIEALTSYLYISNEYKGMPG
ARNLYKIQLNDYKVTCELNPDCQYSSFVSEFQAEKYQLRCSGPGLPLYLHNSNNDKELRVLENNS
DLDQLPQVMPQSKLDFILHLHGTQFWYQMPILPPHFDSKYPPLLLEAYAGPCQKADAIIFRLNWATYLAS
TENIIVASFDGRGSGYQGDHIMAINHNRRLGTFEVEDQIEATRQFSKMFVDDKRIAIWGSYGGYVTSMVL
GAGSGVFCKGIAVAPVSKWEYDSVTERYMGMTDPENLDSYRNSTVMSRAENFKQVEYLLIHGTADDNV
HFQQSAQISKALVDAGVDFQSMWYTEDHGIASSTAHQIHYTHMSHFLKQCFSSL

>Q9N2I7|DPP-IV_FELCA Dipeptidyl peptidase 4 - Felis silvestris catus (Cat).
MKTPKWLLGLLALAIALVTVITPVVLLNKGNDAADSDRRTYTLTDYLKN1FRMKFYLNLRW5DHELYLKQ
DNIIIFNAEYGNSSIFLENSTDFEDEFHSINDYSVSPDQYIFLFYNYVQKWRSHTASYDIYDLNKRQLI
TEEIPNNTQWITSSVGHKLAYVWNDIYVKNEPNSPSQRTWTGKDKVINEGTDWYEEEEVSAYSAL
WWSPNSTFLAYAQNDFVEDVPIEYSFDESLEQYPKTVKIPKAGAVNPIKFFVNVINSSLSPNINATSQ
QIVPPGSVLGHDLYCLDVVTWEISLQLWRLRIQNSIMIDCYDRTSRGWISSVGRQHIEHSTTGWVGR
FRPAEPHTSDGSNFYKIISNEEGYKICHIPQFTDKRNCTFITKGAWEVIGIEALTSYLYISNEYKGMPG
ARNLYKIQLNDYKVTCELNPDCQYSSFVSEFQAEKYQLRCSGPGLPLYLHNSNNDKELRVLENNS
DLDQLPQVMPQSKLDFILHLHGTQFWYQMPILPPHFDSKYPPLLLEAYAGPCQKADAIIFRLNWATYLAS
TENIIVASFDGRGSGYQGDHIMAINHNRRLGTFEVEDQIEATRQFSKMFVDDKRIAIWGSYGGYVTSMVL
GAGSGVFCKGIAVAPVSKWEYDSVTERYMGMTDPENLDSYRNSTVMSRAENFKQVEYLLIHGTADDNV
HFQQSAQISKALVDAGVDFQSMWYTEDHGIASSTAHQIHYTHMSHFLKQCFSSL
>P27487|DPP-IV_HUMAN Dipeptidyl peptidase 4 - Homo sapiens (Human).
MKTPWKVLLGLGAAALVTIITPVELLNLKGTDDATADSRKTITYLKDLYLQLSRLDHELYLKQENNILVFNAEYGNSSVFLENSTDFEGHSINDYSISP9QFILLEYNYVKWHRHSYTAASYDIYDNLNKRQLITEEIPIIITQWVTSPVGHKLVAYWNNDIVKIEPNLPSYRITWTGKEDIYNGITDWWYEEVEVFSAYSA
LWWSPNGTFLAYAQFNDTEPVIEYSFDESQPYPKTRVPYPKAGAVNPTVKFVVNTDLSLSSVTNATS
IQITAPASMLIGDHYLCDVTTAQERISLQWLRRIQYNSVMDCYDESSGRWNCLVARQHIEMSTTTNGVGRFRPSEPHFTLDGAFTKIISNEEGYRHICFQIDKDCFTIKGTWEVIGIEALTSDYLYISNEYKGMPPGGRNLKIQSLDGYTVTCLUDQELNPCQYYSFSFSEAYYQLRCSGPGPLLTYLHSVNDKGLRVLHADN
SALDMLQVQUMPSSKLFIIDIIINETFCTFWQMLILPFDKSSKYPLLDDYAGPSQKADTVFLNLWATYLA
STENIIVASFDGRGSGQYQQDKIMHAINRRLGUFEVQIEAARQFSKFGVFDNKRIAIWWSGYGVTSMV
LGSGSGVFKCGIAVAPVSRWEYEDSYVTERYMGLPTPEDNLDHYNSTVMRSNFKQVEYLLIHGTADDNVHFPQOSAIQSKALVDVGVDFQAMWYTEDHDHGIASSTAQHIYTHMSHFIFCFSLP

>P28843|DPP-IV_MOUSE Dipeptidyl peptidase 4 - Mus musculus (Mouse).
MKTPWKVLLGLGAAALVTIITPVELLNLKGTDDATADSRKTITYLKDLYLQLSRLDHELYLKQENNILVFNAEYGNSSVFLENSTDFEGHSINDYSISP9QFILLEYNYVKWHRHSYTAASYDIYDNLNKRQLITEEIPIIITQWVTSPVGHKLVAYWNNDIVKIEPNLPSYRITWTGKEDIYNGITDWWYEEVEVFSAYSA
LWWSPNGTFLAYAQFNDTEPVIEYSFDESQPYPKTRVPYPKAGAVNPTVKFVVNTDLSLSSVTNATS
IQITAPASMLIGDHYLCDVTTAQERISLQWLRRIQYNSVMDCYDESSGRWNCLVARQHIEMSTTTNGVGRFRPSEPHFTLDGAFTKIISNEEGYRHICFQIDKDCFTIKGTWEVIGIEALTSDYLYISNEYKGMPPGGRNLKIQSLDGYTVTCLUDQELNPCQYYSFSFSEAYYQLRCSGPGPLLTYLHSVNDKGLRVLHADN
SALDMLQVQUMPSSKLFIIDIIINETFCTFWQMLILPFDKSSKYPLLDDYAGPSQKADTVFLNLWATYLA
STENIIVASFDGRGSGQYQQDKIMHAINRRLGUFEVQIEAARQFSKFGVFDNKRIAIWWSGYGVTSMV
LGSGSGVFKCGIAVAPVSRWEYEDSYVTERYMGLPTPEDNLDHYNSTVMRSNFKQVEYLLIHGTADDNVHFPQOSAIQSKALVDVGVDFQAMWYTEDHDHGIASSTAQHIYTHMSHFIFCFSLP
>P22411|DPP-IV_PIG Dipeptidyl peptidase 4 - Sus scrofa (Pig).

MKTPKVLGLGLGIALTVTVAPVLLNKDDAAADSRTYTLTDLKSTFRVFKYTQLWISDHELYK
QENNILLFNAYGNSSIFLENSTFDELGYSTNDSVSPDQRQILFELYNYVKQWRHSYTASYDIYLNDLRQQL
ITEERIPNTQWITWSPVGHKLYAVWNNNDYIKNPNLSSRIGRITWTGKENVYNGVTDWYEEVFSAYSA
LWSPNGTFLAYQFDNTEPIEYFSDESLOYPKTVRIPYPKAGANPTVKFVTFTRTLSPNASTVS
YQIVPPASVLIGDHYLCVGTVTWTEERISLQWIRRAQNSYIDICDYDESTGRWISSVARQHIESTTNGVG
RFRPAEPHFTSDGNSFYKISNEEGKHKFQTDSNCFTITKGAWEVIGIEALTSDLYYISNEHKGMP
GGRNLRYRQLNDYTKVTCLSLCEINPERCQYSSASFSNKKYYQRCFPGPLYLHSSSDKELRVLLEDN
SALDKMLQDVQMPKSKLDDVNLHGTKFQWIMLPPHDFSKKYPLLIEVAGPSCQEVDSVTFRSLWATYLA
STENIIASFDGRSGSYQDSIMHAINRRGTEVEDQIEATRQFSKMGFDKRIAIGWWSYGGYVSMV
LGAGSGVFCKGIAVAPSKWEYYDSVETYREMGLPTTPEDNLDDYRNSTVMRANFKQVEYLLIHGTADDN
VHFQQSAQLSLKALVDAGDFQTMMYDDEHDGIASNMHQAQHIYTHMSHFLQCFSLP

>PI4740|DPP-IV_RAT Dipeptidyl peptidase 4 - Rattus norvegicus (Rat).

MKTPKVLGLGLGIALTVTVAPVLLNKDEAAADSRTYTLAFLKNTFRVFKYSLSRLWSDSELYKQE
NNILLFNAYGNSSIFLENSTFDELGYSTNDSVSPDRLFVLEENVYVKQWRHYESYDIYLNDLRQQLIT
EKKIPNTQWITWSQEGHKLAVWKNDYIKIKEPHLPSHRITSTGKENVIFNGINDWYEEIFGAYSALW
WSPNGTFLAYQFDNTEPIEYFSDESLOYPKTVRIPYPKAGANPTVKFVTFTRTLSPNASTVS
IITAPASVTTGDHYLCVAVWSEDRISLQWLRRINQNYSMAICDYKKTLLVNCPTTTQEHETSATGWCGRF
RPAPFHFTSDGSSFYKIVSDKGYKHCQFKDKRPQVCTITKGAWEVISIEALTSDLYYISNEYKEM
PGGRRNLKQITLDTNKKCLSDINPERCQYSSVSLSKEAKYYQLCRGPGPLYLHSTQDLKELRVLLED
NSALDKMLQDVQMPKSKLDDVNLHGTKFQWIMLPPHDFSKKYPLLIDVAYAGPCSQKAADAIFLNWATYL
ASTENIIASFDGRSGSYQDSIKMHAINEGRLGTLEVEDQIAEARQFLKMGFDKRTIAIWWGWSYGGYVSMV
VLGSGSVFCKGIAVAPSKWEYYDSVETYREMGLPTTPEDNLDDYRNSTVMRANFKQVEYLLIHGTADDN
VHFQQSAQLSLKALVDAGDFQTMMYDDEHDGIASNMHQAQHIYTHMSHFLQCFSLP
3.1.6 BLAST

The **Basic Local Alignment Search Tool (BLAST)** finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families. For performing BLASTp under NCBI website, all protein sequences are subjected to blast analysis using default parameters. An example of blast run image is given below.

![BLAST input sequence showing PDB database chosen for analysis.](image)

**Figure-3.2: BLAST input sequence showing PDB database chosen for analysis.**
The output of blast analysis is given below.

**Distribution of 21 Blast Hits on the Query Sequence**

Mouse-over to show define and scores, click to show alignments

**Figure 3.3:** Blast alignment result showing number of blast hits against query P18266 sequence

<table>
<thead>
<tr>
<th>Sequences producing significant alignments:</th>
<th>(Bits) Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi</td>
<td>60584449</td>
</tr>
<tr>
<td>gi</td>
<td>40399063</td>
</tr>
<tr>
<td>gi</td>
<td>33398001</td>
</tr>
<tr>
<td>gi</td>
<td>52696028</td>
</tr>
<tr>
<td>gi</td>
<td>50313774</td>
</tr>
<tr>
<td>gi</td>
<td>27574040</td>
</tr>
<tr>
<td>gi</td>
<td>62737295</td>
</tr>
<tr>
<td>gi</td>
<td>60393957</td>
</tr>
<tr>
<td>gi</td>
<td>11085918</td>
</tr>
<tr>
<td>gi</td>
<td>30749225</td>
</tr>
<tr>
<td>gi</td>
<td>110990222</td>
</tr>
<tr>
<td>gi</td>
<td>66361328</td>
</tr>
<tr>
<td>gi</td>
<td>56554517</td>
</tr>
<tr>
<td>gi</td>
<td>116670031</td>
</tr>
<tr>
<td>gi</td>
<td>11667019</td>
</tr>
<tr>
<td>gi</td>
<td>86854273</td>
</tr>
<tr>
<td>gi</td>
<td>59319748</td>
</tr>
<tr>
<td>gi</td>
<td>12064602</td>
</tr>
<tr>
<td>gi</td>
<td>12012099</td>
</tr>
<tr>
<td>gi</td>
<td>98327471</td>
</tr>
<tr>
<td>gi</td>
<td>15969366</td>
</tr>
</tbody>
</table>
3.1.7 FASTA

Fasta analysis was performed by pasting sequence in the box given. All parameters are kept as defaults expect matrices. They are discussed in Results section in detail. An example of Fasta input is given below. Database PDB is selected with Blosum 50 matrix.

**Figure-3.4: Summary alignment result of BLAST analysis**

---

The alignment result is shown below: Score = 1359 bits (35%), Expect = 0.0, Method: Composition-based stats, Identities = 848/738 (87%), Positives = 856/738 (95), Gaps = 1/738 (0%)

**Query 39**  80A-0AALSSERFTELSDLTVLYKZANNIILFNAKYNQGI  87
  MKG 40A  ASCA0TTLSDTLTVLYKZANNIILFNAKYNQGI  
  **Subject 1**  80T-DDAASRFEKETLTLYKZANNIILFNAKYNQGI  90

**Query 30**  FLESTDEEF RHEIDY+SDQ+FTPFILFERTURQRHFSTTYTSTTBYERQLLTE 147
  FLESTDEEF RHEIDY+SDQ+FTPFILFERTURQRHFSTTYTSTTBYERQLLTE  
  **Subject 61**  FLESTDEEF RHEIDY+SDQ+FTPFILFERTURQRHFSTTYTSTTBYERQLLTE  120

**Query 140**  8T0C+TSTSFSESKLTVNYKCRDDNYFNNSSIKIITGTOENAIHYGDQDYMST  207
  8T0C+TSTSFSESKLTVNYKCRDDNYFNNSSIKIITGTOENAIHYGDQDYMST  
  **Subject 121**  8T0C+TSTSFSESKLTVNYKCRDDNYFNNSSIKIITGTOENAIHYGDQDYMST  190

**Query 200**  8TSDYVSFDFOTTDFATQPQTOVLIEVSFYVDSLQTFKTEDIPFTEAGANPTYKL  247
  8TSDYVSFDFOTTDFATQPQTOVLIEVSFYVDSLQTFKTEDIPFTEAGANPTYKL  
  **Subject 181**  8TSDYVSFDFOTTDFATQPQTOVLIEVSFYVDSLQTFKTEDIPFTEAGANPTYKL  240

**Query 268**  FV+TD+1++  TNAS++T  FAX+NL  GY+GTVY+TA  +IR+LQGRHRIKVYDM  
  **Subject 241**  FV+TD+1++  TNAS++T  FAX+NL  GY+GTVY+TA  +IR+LQGRHRIKVYDM  300

**Query 320**  DYNSTQKaosaaaqhvnnihestegovqphpdimpdhsdgfnyekehcvxq  357
  DYNSTQKaosaaaqhvnnihestegovqphpdimpdhsdgfnyekehcvxq  
  **Subject 301**  DYNSTQKaosaaaqhvnnihestegovqphpdimpdhsdgfnyekehcvxq  380
**Figure-3.5: Summary table view of Fasta**

<table>
<thead>
<tr>
<th>Alignment</th>
<th>DDB</th>
<th>Source</th>
<th>Length</th>
<th>Identity</th>
<th>Similar%</th>
<th>Overlap</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P08_1UGR_A</td>
<td>mol protein length: 736 Dipeptidyl PPTidas</td>
<td>736</td>
<td>97.076</td>
<td>99.501</td>
<td>734</td>
<td>6.6e-212</td>
</tr>
<tr>
<td>2</td>
<td>P08_1UGR_B</td>
<td>mol protein length: 736 Dipeptidyl PPTidas</td>
<td>736</td>
<td>97.076</td>
<td>99.501</td>
<td>734</td>
<td>6.6e-212</td>
</tr>
<tr>
<td>3</td>
<td>P08_1UGY_A</td>
<td>mol protein length: 740 Dipeptidyl PPTidas</td>
<td>740</td>
<td>97.076</td>
<td>99.501</td>
<td>734</td>
<td>6.6e-212</td>
</tr>
<tr>
<td>4</td>
<td>P08_1UGY_B</td>
<td>mol protein length: 740 Dipeptidyl PPTidas</td>
<td>740</td>
<td>97.076</td>
<td>99.501</td>
<td>734</td>
<td>6.6e-212</td>
</tr>
<tr>
<td>5</td>
<td>P08_1UGY_A</td>
<td>mol protein length: 740 Dipeptidyl PPTidas</td>
<td>740</td>
<td>97.076</td>
<td>99.501</td>
<td>734</td>
<td>6.6e-212</td>
</tr>
<tr>
<td>6</td>
<td>P08_1UGY_B</td>
<td>mol protein length: 740 Dipeptidyl PPTidas</td>
<td>740</td>
<td>97.076</td>
<td>99.501</td>
<td>734</td>
<td>6.6e-212</td>
</tr>
<tr>
<td>7</td>
<td>P08_1UGY_B</td>
<td>mol protein length: 731 Dipeptidyl PPTidas</td>
<td>731</td>
<td>97.062</td>
<td>99.435</td>
<td>731</td>
<td>1.7e-211</td>
</tr>
<tr>
<td>8</td>
<td>P08_1UGY_A</td>
<td>mol protein length: 731 Dipeptidyl PPTidas</td>
<td>731</td>
<td>97.062</td>
<td>99.435</td>
<td>731</td>
<td>1.7e-211</td>
</tr>
<tr>
<td>9</td>
<td>P08_1UGY_A</td>
<td>mol protein length: 731 Dipeptidyl PPTidas</td>
<td>731</td>
<td>97.036</td>
<td>97.001</td>
<td>736</td>
<td>2.3e-211</td>
</tr>
<tr>
<td>10</td>
<td>P08_1UGY_B</td>
<td>mol protein length: 731 Dipeptidyl PPTidas</td>
<td>731</td>
<td>97.036</td>
<td>97.001</td>
<td>736</td>
<td>2.3e-211</td>
</tr>
<tr>
<td>11</td>
<td>P08_1UGY_D</td>
<td>mol protein length: 731 Dipeptidyl PPTidas</td>
<td>731</td>
<td>97.036</td>
<td>97.001</td>
<td>736</td>
<td>2.3e-211</td>
</tr>
<tr>
<td>12</td>
<td>P08_1UGY_C</td>
<td>mol protein length: 731 Dipeptidyl PPTidas</td>
<td>731</td>
<td>97.036</td>
<td>97.001</td>
<td>736</td>
<td>2.3e-211</td>
</tr>
<tr>
<td>13</td>
<td>P08_1UGY_A</td>
<td>mol protein length: 733 Dipeptidyl PPTidas</td>
<td>733</td>
<td>97.039</td>
<td>99.491</td>
<td>728</td>
<td>3.9e-211</td>
</tr>
<tr>
<td>14</td>
<td>P08_1UGY_B</td>
<td>mol protein length: 733 Dipeptidyl PPTidas</td>
<td>733</td>
<td>97.039</td>
<td>99.491</td>
<td>728</td>
<td>3.9e-211</td>
</tr>
<tr>
<td>15</td>
<td>P08_1UGY_D</td>
<td>mol protein length: 733 Dipeptidyl PPTidas</td>
<td>733</td>
<td>97.039</td>
<td>99.491</td>
<td>728</td>
<td>3.9e-211</td>
</tr>
<tr>
<td>16</td>
<td>P08_1UGY_C</td>
<td>mol protein length: 733 Dipeptidyl PPTidas</td>
<td>733</td>
<td>97.039</td>
<td>99.491</td>
<td>728</td>
<td>3.9e-211</td>
</tr>
</tbody>
</table>

**Figure-3.6: First few reported hits from Fasta sequence analysis**
3.1.8 WuBLAST

Similar to Fasta program, Wu Blast program was used to analyze query sequence scan against pdb protein database using default options. The input image is given below.

![Image of WuBLAST interface](image)

**Figure-3.7:** Query sequence subjected to WuBlast run against PDB database
### Figure-3.8: The output of WuBlast alignment

#### 3.1.9 CLUSTALW

ClustalW multiple sequence analysis is performed to determine the number of proteins that share common structural and functional features. As an input to clustalw all sequences in fasta format are pasted in the given box with default options. The output is analyzed for sequences that are aligned for the complete length, scores, alignment, conserved residues, substituted and semi-conserved substituted residue patterns are observed. An example input in ClustalW alignment is given below

<table>
<thead>
<tr>
<th>Alignment</th>
<th>DRID</th>
<th>Source</th>
<th>Length</th>
<th>Score</th>
<th>Identity%</th>
<th>Positive%</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EDE-2BE6_A</td>
<td>mol protein length:738 DIFEPTIDYL PEPTIDASE IV</td>
<td>738</td>
<td>5796</td>
<td>67</td>
<td>94</td>
<td>0.</td>
</tr>
<tr>
<td>2</td>
<td>EDE-2BE6_R</td>
<td>mol protein length:738 DIFEPTIDYL PEPTIDASE IV</td>
<td>738</td>
<td>5796</td>
<td>67</td>
<td>94</td>
<td>0.</td>
</tr>
<tr>
<td>3</td>
<td>EDE-1L2F_A</td>
<td>mol protein length:740 Dipeptidyl peptidase IV</td>
<td>740</td>
<td>5788</td>
<td>67</td>
<td>94</td>
<td>0.</td>
</tr>
<tr>
<td>4</td>
<td>EDE-1L2F_R</td>
<td>mol protein length:740 Dipeptidyl peptidase IV</td>
<td>740</td>
<td>5788</td>
<td>67</td>
<td>94</td>
<td>0.</td>
</tr>
<tr>
<td>5</td>
<td>EDE-1WO7_A</td>
<td>mol protein length:740 Dipeptidyl peptidase IV</td>
<td>740</td>
<td>5786</td>
<td>67</td>
<td>94</td>
<td>0.</td>
</tr>
<tr>
<td>6</td>
<td>EDE-1WO7_B</td>
<td>mol protein length:740 Dipeptidyl peptidase IV</td>
<td>740</td>
<td>5786</td>
<td>67</td>
<td>94</td>
<td>0.</td>
</tr>
<tr>
<td>7</td>
<td>EDE-1PF9_A</td>
<td>mol protein length:731 Dipeptidyl peptidase IV soluble form</td>
<td>731</td>
<td>5748</td>
<td>88</td>
<td>94</td>
<td>0.</td>
</tr>
<tr>
<td>8</td>
<td>EDE-1PF9_B</td>
<td>mol protein length:731 Dipeptidyl peptidase IV soluble form</td>
<td>731</td>
<td>5748</td>
<td>88</td>
<td>94</td>
<td>0.</td>
</tr>
</tbody>
</table>
3.2. 3-D STRUCTURE PREDICTION

Structure prediction of a protein is the one of the method to predict three-dimensional structure of a protein from its amino acid sequence. Prediction is basically different from the inverse problem of protein design [81]. Protein structure prediction is one of the most important goals pursued by bioinformatics and chemistry; it is highly important in medicine such as in drug design and biotechnology.

Prediction of 3D protein structures from amino acid sequences represents one of the most important problems in computational biology. The community-wide Critical Assessment of Structure Prediction (CASP) experiments have been designed to obtain an objective assessment of in the field [82], the practical role of protein structure prediction is now more important than ever. Considerable amounts of protein sequence data are produced by modern large-scale DNA sequencing efforts such as the Human Genome Project. Despite community-wide efforts in structural genomics, the output of experimentally determined protein structures typically by time-consuming and relatively
expensive X-ray crystallography or NMR spectroscopy. The protein structure prediction remains an extremely difficult and unresolved undertaking. Though, structure modeling processes often involve human interventions because the human-expert knowledge combined with biochemical information like function, mutagenesis, catalytic residues, [83][84] etc. could help in both structural assembly and model selection. However, the development of fully-automated algorithms has the advantage in the potential application on proteome-scale structure predictions. Particularly, it allows non-experts to generate structural models for their own sequences through servers. Homology models of proteins are of great interest for planning and analyzing the biological experiments when no experimental three-dimensional structures are available. Building homology models requires specialized programs and up-to-date sequence and structural databases. Homology modeling or comparative modeling methods [85] are able to predict the 3D structure of a protein sequence by using information derived from a homologous protein of known structure. The utility of homology methods is evident when considering the vast numbers of potential coding sequences that are available nowadays. It has been estimated that of the order of 20-30% of these sequences open reading frames, or ORFs can be assigned to a fold classification derived from structure in the PDB protein structural databank. If a 3D model of the protein[86][87] of interest can be derived, it may be usable as the basis for a structure based drug design study, or as input to a molecular docking procedure, which aims to find the binding site for small molecule ligands.

In addition, such models can be a useful aid to the rational design of experiments such as site-directed mutagenesis or in understanding protein stability and function.
Proteins from different sources and sometimes diverse biological functions can have similar sequences, and it is generally accepted that high sequence similarity is reflected by distinct structure similarity. Indeed, the root mean square deviation (RMSD) of the alpha-carbon co-ordinates for protein cores sharing 50% residue identity is expected to be around 1Å. This fact served as the premise for the development of comparative protein modeling also often called modeling by homology or knowledge-based modeling, which is presently the most reliable method. Comparative model building consists of the extrapolation of the structure for a new (target) sequence from the known 3D-structure of related family members (templates). It is expected that two proteins of similar origin and function would have reasonable structural similarity. Therefore it is possible to use the known structure as a template for modeling the structure of the unknown structure. While the high precision structures required for detailed studies of protein-ligand interaction can only be obtained experimentally, theoretical protein modeling provides the molecular biologists with "low-resolution" models which hold enough essential information about the spatial arrangement of important residues to guide the design of experiments. The rational design of many site-directed mutagenesis experiments could therefore be improved if more of these "low-resolution" theoretical model structures were available.

Given a correct alignment on a related template several methods can produce an accurate model, while without a correct alignment no method can produce a good model. There are several computer programs and web servers that automate the comparative modelling process. The easiest method is to use a web server such as Swiss-Modeller [88], 3D-Jigsaw [89] homology or What-IF [90]. Alternative programs such as What-IF modeller can be run locally on the computer.
While the web servers are convenient and useful, the best results in a difficult or unusual modelling case, such as problematic alignments, modelling of loops, existence of multiple conformational states, and modelling of ligand binding, are still obtained by non-automated, expert use of the various modelling tools. All structure prediction techniques depends one way or another on experimental data. Unfortunately, all protein structures contain errors. Sometimes even models with errors may be useful, because some aspects of function can be predicted from just coarse structural features of a model. Hence, verification of the data used in modelling procedures is a prerequisite for good results. Many of the same verification techniques can of course also be used to get an impression of the quality of the model.

### 3.2.1 Query Sequence

>Q9N2I7|DPP-IV_FELCA Dipeptidyl peptidase 4 - Felis silvestris catus (Cat).

```
MKTPWKLGLLGLAALITITVPPVLLNKGNADADSSRRTYTLTDYKNTFRVKFYSLRWVDHYKYQ
DNNILNFNAEYGNSSTIFLENSTFDEFEHSINDYSVSPDGQFILLEYNYVQWRHSYTAASYDIYDNLNKRQLI
TEEKIPNNTQWITWSPEGHKLAYVWKNDVYVKNEPNSSSHRITWTGEENAIYNGIADVYEEEIFSAYSAL
WWSPKGTFLAYAQFNDTQVPLIEYSFYSDESLQYPMTRIPYPKAGAAANPTVKLKVITDNLNPNTATS
EITPPAAMLTGDDYLCVDVANEERILQWLRRIQNYMSVMDIRDNSTGKISSAAQEHIMSTGWG
FRPAEPHTSDGRNFYKIIISNEDGYKHCIRFQIDKKDCTFITKGAEWAGIEALITTDLYYISNEYKGM
GRNLYKIQLNDYTIVACLSELKPERCQOYSVVSFSKEAKYYQLRSCEGPLELHYTLLRHSSDEELLRVLEDNS
ALDKMLQEVQMPSKKLDFIILNETKFQWQMILPSDFTKYPILLDYYAGPCSQKADAIFRLNNAVYL
TENIVASFDGRSGYQQDKHIMHAVNRRTGTFEVEQIEARQFSKMGFVDKKRIAIGWGSYGGYVTSMVL
GAGSGVFKCGIAVAPVRSEYYDSVYTERYMGFLPTQDNLDYYYYNSTMRAENFKQVEYLLIHTADDNV
HFQQSAQISKALVDAGVFQAMWYTEDDHIGASGPASHIQHIYTMESHIKQFCFSLP
```
3.2.2 Subject Sequence

>2BGR:B|pdbid|chain|sequence

NKGTDDATADSRKTYTLTDYLKNTYRLKLKLYSLRISDHEYLYKQENNIFNAEYGNSSVFLENSTFDEFG
HSINDYSISPDGFILLEYNYVKQWRHSYSTASYDIYDLNKRLITEER1PNNTQWVTSPVGKHLYAVVWN
DIYVKIEPNLPSYRITWGTKEDIYYNGITDWSYEEVFSAYSALWSPNFGFLAYQFDNDTEVPLIEYSFY
SDESLQYPKTVRPYPKAGAVNPTKFFVNTDSLSSVTNATSIQITAPASMLGDHYLCEVVTATQERIS
LQWLRRIQNYSVMIDCDYDESSGRWNCLVARQHIEMSTTGWGVGRFRFSEPHTLDGNSFYKIIISNEEGYRH
ICYFQIDKDCFTKGTWVEIGIEALTSDYLYYSISNEYKGMPGRNLKYKIQLSDYTKVTDCLSCELPERC
QYYSVSFKEAKYYQLRCSGPGLPRLYTHSVDKGRVLEDNASLKDMLQVQMPSKKLDFTILNETKFW
YQMIILPPHDFKSKKYPLLVIDYYAPGCSQKADTVRLNWATLASTENIIIVASFDRGRGSGYGQDGKIMHAINR
RLGTFEVDQIEARAQFSKGMFDNKR1IWGWSGYTVSMLGSGGSGVFKCGIAVAPVSRWNEYDYSVT
ERYMGLTPEDNLDHYRNSVMSRAENFKQVEYLLIHTADDNVHFQSQAPIKALVDVGVDFQAMWYDTE
DHGIASTAHQHIYTHMHSFIKQCFSLP
Figure-3.10: PAM 250 FASTA FORMAT
3.2.3 Modeller9v1

MODELLER is a computer program that models three-dimensional structures of proteins and their assemblies by satisfaction of spatial restraints. MODELLER implements an automated approach to comparative protein structure modeling by satisfaction of spatial restraints [91][92]. Briefly the core modeling procedure begins with an alignment of the sequence to be modeled (target) with related known 3D structures (templates). This alignment is usually the input to the program. The output is a 3D model for the target sequence containing all main chain and side chain non-hydrogen atoms. Based on the given an alignment, the model will obtain.

![Figure-3.11: Modeller workspace showing on windows command prompt](image-url)
3.2.4 Methodology

Step 1

Comparative models were constructed for various gene/protein sequences to study the sequences in the structural context and to suggest site directed mutagenesis experiments for elucidating specificity changes in this apparent case of convergent evolution of enzymatic specificity. To perform homology modeling, Blast analysis has been carried out by using PAM and BLOSUM matrices against the protein structure sequence database with the following Swiss P81425, Q9N2I7, P27487, P28843, P22411 and P14740. Out of which Q9N2I7 is taken into consideration as the overlap residues are more and have the much similarity and identity. Thus we select this sequence. So we consider only that sequence.

Step 2

>>PDB:2BGR_B mol:protein length:738 DIPEPTIDYL PEPTIDASE IV (738 aa) inits: 4540 init1: 4540 opt: 4549 Z-score: 5562.6 bits: 1040.0 E(190205): 0  Smith-Waterman score: 4549; 87.8% identity (96.7% similar) in 738 aa overlap (29-765:1-738)

```
  10  20  30  40  50
Q9N2I7 MKTPWKVLLGGLGLAALITIITVPVLLNKG-NDAAADSRRTYTLTDYLKNTFRVFYSL
::: .::: :.:.::: ::::::::: .:.:.::: :::::::::
PDB:2B   NKGTDDATADSRKTYTLTDYLKNTRLKLYSL
       10  20  30

  60  70  80  90  100  110
Q9N2I7 RWVSDHDYLYKDQNNLNFNAEYGNSSIFLENSTFDEFEHVSINDYVSPDGQFILLEYNY
```
PDB:2B  RWISDHEYLYQENNILVFNAEGNSSLFLENSTFDEFGHSINDYSISPDBQFILLEYNY

40  50  60  70  80  90

120  130  140  150  160  170
Q9N2I7  VKQWRHSYTASYDIYDLNKRQITEEKPNNNTQWITWSPEGHKLAYWKNDFVYKNEPNS

180  190  200  210  220  230
Q9N2I7  SSHRITWTGENAIYINGIDWYEEIESAYSALWSPKFTFLAYAQFNQVDVPLEYSF

240  250  260  270  280  290
Q9N2I7  YSDESLQYPMTRIPYPKAGAANPTVKLFVIKTDNLNPNTNATSVEITPPAAMLTGDYYL

300  310  320  330  340  350
Q9N2I7  CDVTWANEERISLQWLRIQNSVMDIRDYNNSTGKNISAAQHEIMSTTGWGRFRPA

360  370  380  390  400  410
Q9N2I7  EPHFTSDGRNFYKISNEDGYKHICRFQIDKDDCFITKGAWVEIGIEALTTYLYISN

54
PDB:2B  EPHTLDGNSFYKIISNEEGYRHICYFQIDKKDCTFITKGTVEVIGEALTSDYLYYISN

340     350     360     370     380     390
420     430     440     450     460     470
Q9N217  EYKGMGGRNLKLYKIQLNDYTQVACLSCELKPERCQYYSVSVFSKEAKYYQLRCGRGPLPLY

PDB:2B  EYKGMGGRNLKLYKIQLSDYTKVTCLSCELNPERCQYYSVSVFSKEAKYYQLRCGRGPLPLY

400     410     420     430     440     450
480     490     500     510     520     530
Q9N217  TLHRSNDEELRVLLEDNSALMDKMLQEVQMSKLDLDFIIILNETKFWYQMLPPFHDTSKKY

PDB:2B  TLHSSVNKGLRLVEDNSALMDKMLQNVQMSKLDLDFIIILNETKFWYQMLPPFHDTSKKY

460     470     480     490     500     510
540     550     560     570     580     590
Q9N217  PLLIDVYAGPCSQKADAIFRLNWATYLASTENIIVASFDOGRGSGYQGDIMHAVNRRGLT

PDB:2B  PLLLDVYAGPCSQKADTVFRLNWATYLASTENIIVASFDOGRGSGYQGDIMHAVNRRGLT

520     530     540     550     560     570
600     610     620     630     640     650
Q9N217  FEVEDQIEARQFSKMGFVDKRIAIGWSGYGMTSMVLGAGSGVFKCIGAVAPVSRWE

PDB:2B  FEVEDQIEARQFSKMGFVDKRIAIGWSGYGMTSMVLGAGSGVFKCIGAVAPVSRWE

580     590     600     610     620     630
660     670     680     690     700     710
Q9N217  YYDSVTERYMGLPTPQDNLDYYKNSVMSRAENFKQVEYLLIGHTADDNVHFQQSAQIS

PDB:2B  YYDSVTERYMGLPTPEDNLDHYRNSTVMSRAENFKQVEYLLIGHTADDNVHFQQSAQIS
Step 3

Series of commands in the modeller9v1 that will generate model with superimposed and optimized structure.

1. Mod9v1 search.py

2. Mod9v1 malign.py

3. Mod9v1 get-model.py

4. Mod9v1 optimize.py

5. Mod9v1 superpose.py
Files in Modeller9v1

1. Search File:

```python
# This will only work if PDB is installed locally and atom_files_directory is set properly.
from modeller import *
from modeller.scripts import sequence_search
log.verbose()
env = environ()
env.io.atom_files_directory = './atom_files'
env.out.prc = 'SHORT'
env.signif_cutoff = (6, 3)
env.search_randomizations = 30
try:
    sequence_search(err, sequence='QG6392.prd', segfile='QG6392.cbn',
                    chains_list='test.cou')
except IOError:
    print "The chains databases are probably not installed on your system."
    print "If desired, you can download these from the Modeller website."
    print "(Data file downloads page) However, consider using profilebuild()" 
    print "instead; see the basic modelling tutorial at the website."
```

Figure-3.12: Modeller workspace showing sequence search file

In this search file the target sequence file name specified as Q9N2I7 with the extensions.

The command “mod9v1 search.py” searches target file by using this file.

2. Alignment file:

Figure-3.13: Modeller workspace showing sequence alignment file
In this file the alignment should be the same as like alignment in the alignment program FASTA that we have taken. The command mod9v1 malign.py will check this alignment and also checks template (2BGR) sequence with template structure. The alignment sequence (2BGR) must match that from the 2BGR (PDB) in the atom files exactly.

3. Get-model file:

```python
# Step 5: model building
# This script should produce two models, 1fdx.B9990001.pdb and 1fdx.B9990002.pdb.

from modeller import *
from modeller.automodel import *  # Load the automodel class

log.verbose()  # To get different models from another script

env = environ(rand_seed=-12341)  # directories for input atom files
env.io.atom_files_directory = '../atom_files'

a = automodel(env,
               sinfile = 'alignment.segment',  # alignment filename
               knowns = ['2BGR'],  # codes of the templates
               sequence = '2BGR',  # code of the target
               assess_methods = assess.DOPE)  # request GA341 assessment

a.starting_model = 1  # index of the first model
a.ending_model = 5  # index of the last model
a.deviation = 4.0  # has to >0 if more than 1 model
a.make()  # do homology modelling
```

Figure-3.14: Modeller workspace showing get-model file
In this get-model file we have to specify the known template structure file name and target protein sequence file name. Here target protein is specified as Q9N2I7 and template structure as 2GBR.

Here certain modifications were made, such as

Starting model=1

Ending model =5

This will generate five models.

4. Optimize file:

```python
# Example for: conjugate_gradients(), molecular_dynamics(), model.switch_triple()

# This will optimize stereochemistry of a given model, including
# non-bonded contacts.
from modeller import *
from modeller.scripts import complete_pdb
from modeller.optimizers import conjugate_gradients, molecular_dynamics, actions

env = environ()
env.io.atom_files_directory = '../atom_files'
env.edat.dynamic_sphere = True

env.libs.topology.read(file='./lib/top_heav.lib')
environment.libs.parameters.read(file='./lib/par.lib')

cod = 'Q9N2I7'
mdl = complete_pdb(env, cod)
mdl.write(file='code+*.ini')

# Select all atoms:
atomel = selection(mdl)

# Generate the restraints:
mdl.restraints.make(atome1, restraint_type='stereo', spline_on_site=True)
mdl.restraints.write(file='code+*.rst')

mpolv = atomel.energy()

# Create optimizer objects and set defaults for all further optimizations
opt = conjugate_gradients(output='REPORT')
mdl = molecular_dynamics(output='REPORT')
```

Figure-3.15: Modeller workspace showing optimized file
In this file modeled protein name has to be mentioned for optimization. Here modeled protein name specified as Q9N2I7 and the command Mod9v1 optimize.py runs the optimization and gets the modeled protein with stable and minimum energy.

5. Superpose:

```python
# Example for: selection.superpose()

# This will use a given alignment to superpose C-alpha atoms of
# one structure (2ctx) on the other (ifas).

from modeller import *

ev = environ()
env.io.atom_files_directory = '../atom_files'

mdl1 = model(env, file='Q9N2I7')
mdl2 = model(env, file='2GBR')
alin = alignment(env, file='alignment.seg.ali', align_codes=('Q9N2I7', '2GBR'))

atmsel = selection(mdl1).only_atom_types('CA')
r = atmsel.superpose(mdl2, alin)

# We can now use the calculated RMS, DRMS, etc. from the returned 'r' object:
RMS = r.rms
DRMS = r.DRMS
print "4d equivalent positions" % r.num_equiv_pos

mdl2.write(file='2ctx.fit')
```

**Figure-3.16: Modeller workspace showing superimposed file**

Here modeled protein file name is specified as (Q9N2I7) and template file name as (2GBR) for superimposition. The command mod9v1 superpose.py runs superimposition of these two proteins and gets RMSD (root mean square deviation) value.
3.3 Ligand Binding Sites Prediction

Identification of protein biochemical functions based on their three-dimensional structures is now required in the post genome-sequencing era. Ligand binding is one of the major biochemical functions of proteins, and thus the identification of ligands and their binding sites is the starting point for the function identification. Most protein function prediction methods identify small ligand molecules for protein structures by applying the principles of molecular complementarity. They assume that the ligand has complementary geometrical and physicochemical properties to the binding site and that similar binding sites bind similar ligands. Structural Genomics initiatives are generating an increasing number of protein structures with very limited biochemical characterisation. The analysis and functional assignment of protein structure is a key challenge and a major bottleneck towards the goal of well-annotated genomes. As shape plays a crucial role in biomolecular recognition and function, the development of shape analysis techniques is important for understanding protein structure-function relationships.

3.3.1 Protein binding site prediction

Proteins have to interact with other molecules like DNA, small molecules called as ligand or other proteins to perform their biological function. Knowledge about where the protein binds to other molecules gives us a better understanding of its biological function. Before discussing protein-protein interaction site prediction, Protein–ligand binding sites are the active sites on protein surface that perform protein functions. Thus, the identification of those binding sites is often the first step to study protein functions.
and structure-based drug design. There are many computational algorithms and tools developed in recent decades, such as LIGSITE, PASS, Q-Site Finder, SURFNET, and so on.

### 3.3.2 Identification of protein-ligand binding site

Proteins not only interact with other proteins but also interact with some small molecules called as ligands. Like NAD - Nicotinamide adenine dinucleotide, AMP - Adenosine 5'-onophosphate etc. Unlike the interaction between proteins, ligands tend to bind to the pockets (cavities) on protein surface. Identification and evaluation of these ligand binding sites are the initial steps for protein structural-based drug design. Characterizing these ligand binding pockets plays an important role in automated ligand docking. In the last decade, a variety of computational methods has been developed for the location of possible ligand-binding sites of proteins. Structure Based Drug Design (SBDD) is a computational approach to lead discovery that uses the three-dimensional structure of a protein to fit drug-like molecules into a ligand binding site to modulate function. Identifying the location of the binding site is therefore a vital first step in this process, restricting the search space for SBDD or virtual screening studies. The detection and characterization of functional sites on proteins has increasingly become an area of interest. Structural genomics projects are increasingly yielding protein structures with unknown functions and binding sites. Binding site prediction was pioneered by pocket detection, since the binding site is often found in the largest pocket. More recent methods involve phylogenetic analysis, identifying structural similarity with proteins of known function and identifying regions on the protein surface with a potential for high binding
affinity. Binding site prediction has been used in several SBDD projects and has been incorporated into several docking tools.

### 3.3.3 Geometry based Protein Pocket Prediction

Most of these pocket detection methods use pure geometric criteria to find clefts on protein surface and do not require any knowledge of the ligands, such as POCKET [93][94], LIGSITE [95] SURFNET [96] CAST [97][98] and PASS [99]. Statistical and empirical studies have shown that the actual ligand binding sites correspond to the largest pocket on a protein surface.

<table>
<thead>
<tr>
<th>Method</th>
<th>Geometric technique</th>
<th>Evaluation dataset</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>POCKET</td>
<td>Identification “protein-solvent-protein” events on 3D grid, 3 directions(X,Y,Z)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LIGSITE</td>
<td>Extension and refinement of POCKET, 7 directions</td>
<td>10 receptor-ligand complexes</td>
<td>7 ligands bind in the 1st, 2 in the 2nd, and 1 in the 3rd largest predicted pocket.</td>
</tr>
<tr>
<td>SURFNET</td>
<td>Place probe sphere between atoms</td>
<td>67 enzyme-ligand structures</td>
<td>83% success rate</td>
</tr>
<tr>
<td>CAST</td>
<td>Alpha shape and triangulation</td>
<td>51 of SURFNET dataset</td>
<td>74% success rate</td>
</tr>
<tr>
<td>PASS</td>
<td>Coat protein by probe sphere layer by layer</td>
<td>30 protein-ligand complexes</td>
<td>ligands bind to the top 3 pockets in 26 cases</td>
</tr>
</tbody>
</table>

Figure-3.17: Summary of Geometry based protein ligand site prediction Methods

**POCKET** introduced the idea of protein-solvent-protein events as key concept for the identification. In **LIGSITE**, the protein is mapped onto a 3D grid. A grid point is part of the protein if it is within 3Å of an atom coordinate; otherwise it is solvent. Next, the x, y,
and z-axes are scanned for pockets, which are characterized as a sequence of grid points, which start and end with the label protein and a period of solvent grid points in between. These sequences are called protein-solvent-protein events. Only grid points that exceed a threshold of protein-solvent-protein events are retained for the final pocket prediction. Since the definition of a pocket in POCKET is dependent on the angle of rotation of the protein relative to the axes, LIGSITE extends POCKET by scanning along the four cubic diagonals in addition to the x, y and z directions. LIGSITE was originally tested on 10 receptor-ligand complexes of which 7 bind in the largest, 2 in the second largest, and 1 in the third largest predicted pocket.

Figure-3.18: POCKET & LIGSITE showing protein-solvent-protein and surface-solvent-surface events
The SURFNET (Laskowski, 1995), program generates molecular surfaces and gaps between surfaces from 3D coordinates supplied in a PDB-format file. The gap regions can correspond to the voids between two or more molecules, or to the internal cavities and surface grooves within a single molecule. The program is particularly useful in clearly delineating the regions of the active site of a protein. It can also generate 3D contour surfaces of the density distributions of any set of 3D data points. All output surfaces can be viewed interactively, along with the molecules or data points in question, using some of the best-known molecular modeling packages. In addition, PostScript output is available, and the generated surfaces can be rendered using various other graphics packages. In SURFNET the key idea is that a sphere, which separates two atoms and which does not contain any atoms, defines a pocket. First, a sphere is placed so that the two given atoms are on opposite sides on the sphere's surface. If the sphere contains any other atoms, it is reduced in size until no more atoms are contained. Only spheres, which are between a radius of 1 to 4\(^\circ\)A are kept. The result of this procedure is a number of separate groups of interpenetrating spheres, called gap regions, both inside the protein and on its surface, which correspond to the protein's cavities and clefts. SURFNET was used to analyze 67 enzyme-ligand structures and the ligand is bound in the largest pockets in 83\% of the cases (Laskowski et al., 1996).
Figure-3.19: SURFNET: Sphere with maximum volume is defined as largest pocket

CAST (Binkowski et al., 2003; Liang et al., 1998) computes a triangulation of the protein's surface atoms using alpha shapes (Edelsbrunner et al., 1995; Edelsbrunner and Mucke, 1994). In the next step, triangles are grouped by letting small triangle flow towards neighbouring larger triangles, which act as sinks. The pocket is then defined as collection of empty triangles. CAST was tested on 51 of 67 enzyme-ligand complexes used for SURFNET (Laskowski et al., 1996) and achieved a success rate of 74%.

PASS (Brady and Stouten, 2000) uses probe spheres to fill cavities layer by layer. First, an initial coating of the protein with probe spheres is calculated. Each probe has a burial count, which counts the number of atoms within 8Å distance. Only probes with count above a threshold are retained. This procedure is iterated until a layer produces no new
buried probe spheres. Then each probe is assigned a probe weight, which is proportional to the number of probe spheres in the vicinity and the extent to which they are buried.

![Image of buried probe spheres and probes with large number of atom contacts](image)

**Figure-3.20: PASS: Pockets or Active site points are the probes with large number of atom contacts**

**BDOCK** uses both the degree of burial and conservation to improve docking. The binding sites of enzyme-inhibitor complexes always involve a very deep buried pocket. It is also believed that the binding sites are more conserved than the rest of surface. Based on this knowledge, First to propose a novel shape complementary scoring function for the initial docking stage, which takes the degree of burial of surface residues into account as different weights. The algorithm of BDOCK is illustrated in the following figure. The numbers represent the buried degrees of surface residues.
3.3.4 Proteins

Proteins play a key role in biological systems. In contrast to DNA, which serves as information storage, proteins are responsible for carrying out functions in biological processes. A better understanding of proteins is crucial to further insights to biological processes. In the following we will explain characteristics of proteins at several levels of detail and emphasize the importance of the three dimensional structure of proteins, related to the functional role.
Protein surfaces typically contain clefts of different shapes and sizes. It has been shown that for enzymes, the active site is commonly found in the largest cleft, Laskowski et al. (1996). This makes the detection of the active site relatively straightforward, and can be performed by a program such as SURFNET, Laskowski (1995). SURFNET does this by placing expandable spheres between protein atoms such that the radii of these spheres do not penetrate the two protein atoms or any nearby atoms. The union of all such spheres is used by SURFNET to describe the 3D cleft shape. Residue conservation can be used to improve the binding pocket prediction, Glaser et al. (2005).

In this contribution, we present an efficient method for comparing protein binding pockets based on a spherical harmonics expansion. Binding pocket shapes are approximated as functions on the unit sphere by describing each surface point by its spherical coordinates and setting $f(\theta, \phi) = r$, Morris et al. (2005). The same methodology can be applied to enhance the description by including other properties such as the electrostatic potential, Ritchie and Kemp (1999, 2000).

3.3.5 The Protein Data Bank

The three-dimensional coordinates of protein structures are usually deposited in the Protein Data Bank [100] the internationally recognized primary depository for all published three-dimensional biological macromolecules. The PDB was founded in 1971 at the Brookhaven National Laboratories containing an initial set of seven protein structures. In 1998 the PDB was put under the responsibility of the Research Collaboratory for Structural Bioinformatics (RCSB) at the Rutgers University, New Jersey (Berman, et al., 2000). Since the beginning of the 1990"s, the number of deposited structures in the PDB has been increasing exponentially. As of today (25/01/2009) the
PDB holds 55,419 models of macromolecule structures of which more than 92% are proteins, 4% are protein-nucleic acid complexes, 2% are DNA, 1% are RNA and few are carbohydrates and antibiotics. 99% of the protein structures were determined by X-ray crystallography (86%) and Nuclear Magnetic Resonance Spectroscopy (14%). 212 structures were solved with electron microscopy. Initially users were allowed to deposit their theoretical models from ab initio or homology modelling calculations, but this practice was stopped in 2006 (Berman, et al., 2000). For the latest statistics on the PDB’s content and growth, see the PDB statistics on the homepage of the RCSB

3.3.6 The worldwide Protein Data Bank

In 2003, the worldwide Protein Data Bank (wwPDB) (Berman, et al., 2003) was announced by its three founding members, namely RSCB, European Macromolecular Structure Database (MSD-EBI) and the Japanese PDB depository (PDBj) with the aim to sustain the PDB as the single non-profit and worldwide accessible depository for structural models of biological macromolecules. In the years since the PDB was founded, new experimental techniques have been developed to determine the structure of macromolecules. Automated scripts and applications have replaced most of the manual curation of deposited data and the Internet has evolved changing the way users submit access and receive data from the PDB. All these innovations have required at certain times in the past adjustments to the data content and data format that, along with disagreements between curators and depositors, introduced in consistencies in the PDB archive. It was recognised that a global effort on international level would be necessary to unify the PDB"s data content and data format, which led to the founding of the PDB. The
remediation project completed at the end of 2007 within the PDB addressed the inconsistencies mentioned above. The project standardized chemical nomenclatures and labelling of amino acids, nucleic acids and small molecules, removed differences between sequences in the different depositories of the founding organizations, updated citations to primary references, databases and taxonomies and improved the representation of large assemblies and viruses.

3.3.7 Bioinformatics on Proteins

With the advent of genomics, hundreds of genomes were deciphered and the sequence information has provided a sound basis for looking at functional modules and biological networks as the underlying principles of biological processes. The technologies employed in biomedical research are becoming more and more data driven. Not only data collection and numerical analysis, but more and more data and information integration are required to interpret experimental data. Bioinformatics as a contemporary approach to uncover the underlying principles of interactions of proteins on all scales from the molecule to the population has evolved rapidly for good reasons. The concept of describing biological processes as conceptual knowledge based models aims to bring light into the unlimited number of experimental observations.
3.3.8 Brookhaven Protein Data Bank (PDB) [www.rcsb.org/pdb]

- The PDB was founded in the early 1970s
- Repository for 3-D biological macromolecular data (especially proteins)
- Crystallographers and NMR-Scientists deposit their experimentally derived structures of proteins
- each protein has 4-letter ID (PDB-ID); e.g.: 1ACB

The Protein Data Bank (pdb) file format is a textual file format describing the three dimensional structures of molecules held in the Protein Data Bank [101]. Most of the information in that database pertains to proteins, and the pdb format accordingly provides
for rich description and annotation of protein properties. However, proteins are often crystallized in association with other molecules or ions such as water, ions, nucleic acids, drug molecules and so on, which therefore can be described in the pdb format as well.

![Figure-3.23: Image of PDB showing 32,520 Protein Structures by year 2005](image)

3.3.9 PDB File Format:

- Developed in the early 70s
- most important: *human* readable (not machine readable)
- and compatible to FORTRAN-programs
- devided in HEADER section and ATOM records
- HEADER contains miscellaneous information
- ATOM records contain xyz-coordinates of the atoms
PDB - online

- Search Field for PDB-IDs
- News, current Information etc.
- Links to important sub-pages (e.g. file format guide)
- http://www.pdb.org

Figure 3.24: Image showing the PDB online Database
HEADER

Describes molecule and gives deposition date

HEADER    HYDROLASE(SERINE PROTEASE) 08-NOV-91 1ACB 1ACB 2

CMPND

Compound – name of the molecule

COMPND    ALPHA-CHYMOTRYPSIN (E.C.3.4.21.1) COMPLEX WITH EGLIN C 1ACB 3

SOURCE

organism

SOURCE    OXEN (BOS TAURUS) AND LEECH (HIRUDO MEDICINALIS) 1ACB 4

AUTHOR

M.BOLOGNESI,F.FRIGERIO,A.CODA,L.PUGLIESE,C.LIONETTI, 1ACB 5

AUTHOR   2 E.MENEGATTI,G.AMICONI,H.P.SCHNEBLI,P.ASCENZI 1ACB 6

REVDAT

History of the modifications made to an entry since its release.

REVDAT    1 31-OCT-93 1ACB 0 1ACB 7

JRNL

The JRNL record contains the primary literature citation that describes the experiment which resulted in the deposited coordinate set.
REMARK

Records can contain free-form annotation, but they also accommodate standardized information; for example, the REMARK 350 BIOMT records describe how to compute the coordinates of the experimentally observed multimer from those of the explicitly specified ones of a single repeating unit.

REMARK records present experimental details, annotations, comments, and information not included in other records.
SEQRES

Sequence of each chain; be aware not all AA mentioned here must be available in
COORDINATES-Section as well!

SEQRES 1 E 245 CYS GLY VAL PRO ALA ILE GLN PRO VAL LEU SER GLY LEU 1ACB 78
SEQRES 2 E 245 SER ARG ILE VAL ASN GLY GLU GLU ALA VAL PRO GLY SER 1ACB 79

HELIX/SHEET/TURN

Secondary structure elements as provided by the crystallographer (subjective)

HELIX 1 H1 ASN E 165 LYS E 169 1 1ACB 108
SHEET 1 CH1 7 TRP E 29 ASP E 35 0 1ACB 111
TURN 1 C01 ASP E 35 GLY E 38 1ACB 129

SSBOND

Disulfide bridges

SSBOND 1 CYS E 1 CYS E 122 1ACB 145

CRYST1, ORIGX1-3, SCALE1-3

Crystallographic parameters

SSBOND 1 CYS E 1 CYS E 122 1ACB 145
ATOM records describe the coordinates of the atoms that are part of the protein. For example, the first ATOM line above describes the alpha-N atom of the first residue of peptide chain A, which is a proline residue; the first three floating point numbers are its x, y and z coordinates and are in units of Angstroms. The next three columns are the occupancy, temperature factor, and the element name, respectively.

The ATOM records present the atomic coordinates for standard residues.

```plaintext
ATOM      1  N   CYS E   1       2.323 -16.405  18.812  1.00 43.48                                               1ACB 158
ATOM      2  CA  CYS E   1       3.017 -15.136  18.786  1.00 35.11                                              1ACB 159
ATOM      3  C   CYS E   1       4.134 -15.068  19.799  1.00 32.90                                               1ACB 160
ATOM      4  O   CYS E   1       4.173 -15.810  20.772  1.00 41.38                                               1ACB 161
ATOM      5  CB  CYS E   1       2.052 -13.969  19.139  1.00 31.14                                              1ACB 162
ATOM      6  SG  CYS E   1       1.246 -14.085  20.788  1.00 34.72                                              1ACB 163
ATOM      7  N   GLY E   2       4.993 -14.081  19.607  1.00 21.94                                               1ACB 164
ATOM      8  CA  GLY E   2       6.057 -13.734  20.499  1.00 20.45                                               1ACB 165
TER
    Terminates a chain
ATOM   2293  O   GLY I  70      12.671  16.511  11.135  1.00 24.53           1ACB2450
ATOM   2294  OXT GLY I  70      12.385  14.435  10.809  1.00 32.36        1ACB2451
```
HETATM

Describe coordinates of hetero-atoms, that is those atoms which are not part of the protein molecule.

Coordinates for Heteroatoms (Water and other non protein atoms)

COULMN | DATA TYPE | FIELD | DEFINITION
--------|-----------|-------|-------------------
1 - 6   | Record tag | 'ARN   ' | Atom serial number.
7 - 11  | Integer    | serial | Atom name.
13 - 16 | Atom name   | name   | Alternate location indicator.
18 - 20 | Residue name| resName| Residue name.
22      | Character   | chainID| Chain identifier.
23 - 26 | Integer     | residEq| Residue sequence number.
27      | AChar       | code   | Code for insertion of residues.
31 - 38 | Real(6.3)   | x      | Orthogonal coordinates for X in Angstroms.
39 - 46 | Real(6.3)   | y      | Orthogonal coordinates for Y in Angstroms.
47 - 54 | Real(6.3)   | z      | Orthogonal coordinates for Z in Angstroms.
55 - 60 | Real(6.2)   | occupancy | Occupancy.
61 - 66 | Real(6.2)   | tempFactor | Temperature factor.
73 - 76 | LString(4)  | segID  | Segment identifier, left-justified.
77 - 78 | LString(2)  | element| Element symbol, right-justified.
79 - 80 | LString(2)  | charge | Charge on the atom.

Figure -3.25: PDB File Format - ATOM entries
Through the years the file format has undergone many changes and revisions. Its original format was dictated by the width of computer punch cards (80 columns). The most recent revision is 3.2.

3.3.10 PDB – Quality of structure:

The PDB is a historical archive. Its contents are not uniform, but reflect the knowledge of the time as well as the data management practices[102]. This may produce incomplete query results.

Spelling errors abound,
e.g. 23 versions of Escherichia coli:
$COLI
COLI
E. COLI
E.COLI
ESCHERCHIA COLI
ESCHERICH $COLI
ESCHERICHIA $ COLI
ESCHERICHIA COLI
ESCHERICHIA COLI.
EXCHERICHIA COLI

While creating datasets for further analysis of proteins, the quality of each structure might be of interest:

- Resolution
3.3.11 PROCHECK → WHAT IF report

- linked at PDB & http://www.cmbi.kun.nl/gv/pdbreport/ and checks the protein for correctness for example Missing atoms, Symmetry, Bond lengths, Torsion angles, Ramachandran plot, Hydrogen bonds, Bump checks and much more

- We get warnings and errors if the structure explored is not within the range of 'normal values' for these aspects.

![Phenylalanine 2D and 3D structures](image-url)

**Figure-3.26:** Image of Proteins structure prediction methods with its advantages

**Figure-3.27:** Image of Phenylalanine 2D and 3D structures
| ATOM | 2567 | N   | PHE B 175 | 7.821 | -25.530 | -22.848 | 1.00 | 8.71 |
| ATOM | 2568 | CA  | PHE B 175 | 8.845 | -25.172 | -21.877 | 1.00 | 9.41 |
| ATOM | 2569 | C   | PHE B 175 | 9.449 | -23.798 | -22.169 | 1.00 | 10.02 |
| ATOM | 2570 | O   | PHE B 175 | 10.664 | -23.613 | -22.103 | 1.00 | 10.37 |
| ATOM | 2571 | CB  | PHE B 175 | 9.928 | -26.251 | -21.848 | 1.00 | 9.53 |
| ATOM | 2572 | CG  | PHE B 175 | 10.969 | -26.137 | -22.982 | 1.00 | 10.03 |
| ATOM | 2573 | CD1 | PHE B 175 | 12.356 | -25.819 | -22.988 | 1.00 | 10.51 |
| ATOM | 2574 | CD2 | PHE B 175 | 11.725 | -27.211 | -23.402 | 1.00 | 10.25 |
| ATOM | 2575 | CE1 | PHE B 175 | 11.821 | -27.095 | -22.869 | 1.00 | 11.17 |
| ATOM | 2576 | CE2 | PHE B 175 | 12.282 | -26.086 | -24.008 | 1.00 | 10.95 |
| ATOM | 2577 | CZ  | PHE B 175 | 10.953 | -26.335 | -23.622 | 1.00 | 11.38 |

Figure-3.28: Image of ATOM entries with its amino acids

Molecular visualization software capable of displaying pdb files (28):

**Visualization tools:**

- Jmol
- PyMOL
- RasMol
- VMD
- Gabedit
- Molden
- Molekel
- Cn3D

Figure-3.29: Image showing the visualization models with its tools
3.4 Introduction to Python:

Traditionally, computer science programs have emphasized system programming languages over scripting languages. However, scripting languages would seem to offer a number of benefits, particularly for the introductory programming sequence. Scripting languages generally have simpler syntax and semantics than system languages. Because of dynamic typing and interpretation they are very flexible and encourage experimentation. The very high-level nature allows students to build more sophisticated and interesting projects with less implementation effort. Probably, the lack of interest in scripting languages has stemmed from the perception that they are “toy” languages and not suited to general purpose programming. While that may have been true of early scripting languages (e.g. UNIX shell scripts), it is certainly not true of modern variants. Python is a scripting language which supports many programming paradigms, including object oriented, imperative, functional, structured and even Meta programming. It is easy to read, self documenting and comes with a huge collection of libraries which make it possible to do almost anything in very little code. This allows for simplicity of use, while maintaining its extreme usefulness. With Python, there are different methods of programming and can be made easier. For instance, modules can be created in order to divide up the coding needed and plus it can be used as a benefit to catch specific mistakes. To use this would involve uniform arguments and parameters. Many Python programmers report substantial productivity gains and feel the language encourages the development of higher quality, more maintainable code."
Python is simple to use, but it is a real programming language, offering much more structure and support for large programs than shell scripts or batch files can offer. On the other hand, Python also offers much more error checking than C, and, being a very-high-level language, it has high-level data types built in, such as flexible arrays and dictionaries. Because of its more general data types Python is applicable to a much larger problem domain than Awk or even Perl, yet many things are at least as easy in Python as in those languages. Python allows you to split your program into modules that can be reused in other Python programs. It comes with a large collection of standard modules that you can use as the basis of your programs or as examples to start learning to program in Python. Some of these modules provide things like file I/O, system calls, sockets, and even interfaces to graphical user interface toolkits like Tk. Python is an interpreted language, which can save you considerable time during program development because no compilation and linking is necessary. The interpreter can be used interactively, which makes it easy to experiment with features of the language, to write throw-away programs, or to test functions during bottom-up program development. It is also a handy desk calculator.

3.5 Prediction of Binding Pockets in Proteins

The objective of the project is to generate a pocket detection program based on Site Finder algorithm.

- The algorithm uses the same interface as Q-Site Finder.
- The program works by scanning along all 3D-cubes of a grid surrounding the protein. The program also scans all cubic diagonals.
Since the protein is scanned in fourteen directions, each grid point can be defined to be part of a site up to fourteen times. Grid points are only retained if they are defined to be part of a site at least ten times.

The program generates a pocket PDB file for each PDB entry with its coordinates calculated from atom coordinates of particular PDB entry.

This data will be obtained using 25 PDB entries that correspond to the GOLD data set.

Only large pockets are considered with degree of buriedness value 12 as precision.

Restricting the size of the pocket is important for reducing the search space required for docking and de novo drug design or site comparison.
3.6 Molecular Docking

Molecular Docking is a study of how two or more molecular structures fit together, for example drug and enzyme or receptor of protein, fit together. Docking as an efficient in silico [103] screening tool is playing an ever-increasing role in rational drug design. Molecular docking is used to predict the structure of the intermolecular complex formed between two or more molecules. The most interesting case is the protein-ligand interaction, because of its applications in medicine. Ligand is a small molecule, which interacts with protein's binding sites. Binding sites are areas of protein known to participate in forming non-covalent, non-bonded interactions with pharmacophoric groups of compounds. There are several possible mutual conformations in which binding may occur. These are commonly called binding modes. Docking is carried out using a computer program in order to dock computer-generated representations of small molecules to a receptor [104]. Good complementarity of a molecule indicates that the molecule is potentially a good binder. The outcome of a docking exercise normally includes some sort of affinity prediction for the molecules investigated, yielding a relative rank-ordering of the docked compounds with respect to affinity.

Technically speaking, the placement of the molecules in the region of interest i.e the receptor-binding site is referred to as ‘docking’, whereas the prediction of affinity is referred to as ‘scoring’. Although termed ‘docking programs’, the programs used nowadays are designed to carry out both the tasks. The distinction between docking and scoring defines also the two major technical challenges faced by docking programs: to predict the binding mode of a molecule correctly also referred to as ‘pose prediction’,
where ‘pose’ refers to the orientation and conformation of a molecule at the receptor binding site and to predict the binding affinity of compounds or to produce a relative rank-ordering for a number of compounds in a reliable manner.

Docking accuracy and scoring reliability are directly associated with two interrelated issues in the docking process: the searching algorithm and the scoring function. Molecular docking can be divided into two separate problems. The search algorithm should create an optimum number of configurations that include the experimentally determined binding modes. These configurations are evaluated using scoring functions to distinguish the experimental binding modes from all other modes explored through the searching algorithm [105]. A rigorous searching algorithm would go through all possible binding modes between the two molecules.

Some common searching algorithms include

- Molecular dynamics
- Monte Carlo methods
- Genetic algorithms
- Fragment-based methods
- Point complementary methods
- Distance geometry methods
- Tabu searches
- Systematic searches
3.6.1 DOCKING:

Docking is often used to foretell the binding orientation of small molecule drug candidates to their protein targets in order to in turn predict the affinity and activity of the small molecule. Hence docking plays an important role in the rational design of drugs.

3.7 Pdb Sum

PDB sum is the database which was used to provide and give summary of every macromolecular structure deposited in the Protein Data Bank (PDB).

Figure-3.30: Image showing the PDB SUM of the Glutathione S-transferase sequence
3.8 3D Jigsaw

It builds three-dimensional models for proteins based on homologues of known structure. It’s a protein comparative modeling server for predicting the structure and function of our protein sequence. 3D-JIGSAW takes part in the Critical assessment of fully automated structure prediction servers. Version 2.0 is out, with a new interactive mode to build models, a domain-oriented template search procedure (Domain Fishing), alignment accuracy measures and a new energy minimization algorithm to refine models.

3.9 CPH Models 3.0

It is a protein homology modeling server and in this, the template recognition is based on profile-profile alignment guided by secondary structure and exposure predictions. Automated neural-network based protein modelling server for tertiary structure prediction.

3.10 Drug Bank

The Drug Bank database is a unique bioinformatics and cheminformatics resource that combines detailed drug i.e. chemical, pharmacological and pharmaceutical data with comprehensive drug target with sequence, structure, and pathway information. The database contains nearly 4800 drug entries including >1,350 FDA-approved small molecule drugs, 123 FDA-approved biotech (protein/peptide) drugs, 71 nutraceuticals and >3,243 experimental drugs. Additionally, more than 2,500 non-redundant protein (i.e. drug target) sequences are linked to these FDA approved drug entries. Each Drug Card
entry contains more than 100 data fields with half of the information being devoted to drug/chemical data and the other half devoted to drug target or protein data.

### 3.10.1 Zinc Database Search

ZINC, a free database of commercially-available compounds for virtual screening. ZINC contains over 13 million purchasable compounds in ready-to-dock, 3D formats. ZINC is provided by the Shoichet Laboratory in the Department of Pharmaceutical Chemistry at the University of California, San Francisco (UCSF).

![ZINC data base home page](image)

**Figure-3.31: Image showing ZINC data base home page.**

### 3.10.2 Isis Draw

ISIS/Draw is a program from MDL that is free for non-commercial use. We can use it to draw chemical structures, and export them for viewing as 3D models. ISIS
means Integrated Scientific Information System. ISIS/Draw is mainly a 2D drawing program, it has some 3D rotation features and can interface with Rasmol for 3D visualization and rendering. ISIS/Draw also includes structure and reaction validation features and can calculate elementary properties such as formula and molecular weight.

In this study the bivalent inhibitors of Glutathione S transferase were drawn by using ISIS/Draw 2.3 and that ligands were used for docking.

![Figure-3.32: Image showing ISIS Draw home page.]

### 3.11 TSAR

TSAR software of version 3.3 was used to study the QSAR derivatives. It has TSAR project window, to which molecular data is entered through import/export file
system. Multiple regression analysis is done by taking physiochemical properties and biological activity. Then a graph was plotted in between actual values and predicted values. A description of the basic operation of TSAR and fundamental aspects of the application with which we need to be familiar, includes the TSAR interface in knowing how to work with projects, data and views. When we work with TSAR graphical interface, the first screen that is displayed is the main TSAR window. This is called a project window. Any data that we handle in TSAR is organized into projects and you view that data using the project Window. Menu bar displays menu items that give access to drop-down menus. Toolbar contains action buttons that provide shortcuts to the most frequently used menu options. View tabs allow us to move between different views of the currently displayed project. Status bar displays general messages about the status of the current project and displays single progress messages. Scroll bar allow us to move around the window area and display information that is beyond the window border.

Figure-3.33: Image showing TSAR Homepage for 2D to 3D Conversion
3.12 Active Site Identification

Active site is identified using CASTp Server (Computed Atlas of Surface Topography of Proteins) binding sites and active sites of proteins and DNAs are often associated with structural pockets and cavities. CASTp server uses the weighted Delaunay triangulation and the alpha complex for shape measurements. It provides identification and measurements of surface accessible pockets as well as interior inaccessible cavities, for proteins and other molecules. It measures analytically the area and volume of each pocket and cavity, both in solvent accessible surface (SA, Richards' surface) and molecular surface (MS, Connolly's surface). It also measures the number of mouth openings, area of the openings, and circumference of mouth lips, in both SA and MS surfaces for each pocket.

You can request calculation for a particular molecule. The results will be shown on the screen or emailed to you. The emailed results include measured parameters for pockets, cavities and mouth openings, as well as listing of wall atoms and mouth atoms for each pocket. In addition, a downloadable PyMOL plug-in will help you to visualize the pocket of your interest.
3.13 WebLab Viewer lite

WebLabViewer provided a very easy-to-use, user-friendly approach for molecule-display. WebLab ViewerLite analyze organic and inorganic structures, proteins, DNA/RNA, and crystals. WebLabViewer has been a cut-down version of a commercial viewer/editor-program now sold by Accelrys. Apparently Accelrys has discontinued the free version of what has now become DS Viewer Pro. WebLab ViewerLite is developed by Molecular Simulations Inc. and is used by 36 users of Software Informer. The most popular versions of this product among our users are: 3.1, 3.2 and 4.0. The list of features is more or less similar to that of the two other viewing-packages (Rasmol and Chime), in addition, WebLabViewer could generate and display surfaces and features a very easy to use interface. Display soft surfaces and solvent accessible surfaces. Visualize organic and
inorganic crystal structures in a variety of display styles. Display proteins using C-alpha Wire, C-alpha stick, Line Ribbon, Flat Ribbon, Solid Ribbon, Tube and Schematic display styles. Color by Amino Acid, Amino Acid chain, pKa, hydrophobicity, and secondary type.

3.14 AUTODOCK

AutoDock is molecular modeling simulation software. It is especially effective for Protein-ligand docking. It is one of the most cited docking software in the research community [106. It is currently maintained by The Scripps Research Institute and Olson Laboratory.
AutoDock is a suite of automated docking tools. It is designed to predict how small molecules, such as substrates or drug candidates, bind to a receptor of known 3D structure. It has applications in: X-ray crystallography, structure-based drug design, lead optimization, virtual screening (HTS), combinatorial library design, protein-protein docking, and chemical mechanism studies [107]. The procedure developed for AutoDock uses a Monte Carlo simulated annealing technique for configurational exploration with a rapid energy evaluation using grid based molecular affinity potentials, thus combining the advantages of a large search space and a robust energy evaluation. This has proven to be a powerful approach to the problem of docking a flexible substrate into the binding site of a static protein.

3.15 Conversion of 2d to 3d structure by using Tsar

- The conversion of 2D structure into 3D structure can be done by using Tsar Software. This conversion is very useful for 3D visualization of 2D structure.
- The purpose of this conversion in this study was to do docking with Autodock software.

**Steps involved in this conversion**

- Open the Tsar and Molecules need to be converted were imported into Tsar.
- Click on the structure option in Tsar, and follow these three steps
  1. Corina-Make3D,
  2. Charge2-derive charges
  3. Cosmic-optimize 3D
- These three steps resulted in conversion of 2D structure into 3D.
• The structure came with this Tsar was fully optimized and stable structure
• After this we can export these 3D converted molecules into our files.

**The three steps in tsar for 3d conversion**

1. corina - Make 3D
2. charge2 - Derive Charges
3. cosmic - Optimize 3D

![Figure-3.36: Image Showing TSAR workspace with 2D to 3D conversion](image-url)
3.16 Protein-Ligand Binding Affinity Analysis:

3.16.1 Protein data bank:

The RCSB PDB provides a variety of tools and resources for studying the structure of biological macromolecules and their relationship to sequence, function, and disease. The RCSB is a member of the www.PDB whose mission is to ensure that the PDB archive remains an international resource with uniform data. This site office used for browsing, searching and reporting that utilize the data resulting from ongoing efforts to create a more consistent and comprehensive archive.

PDB database was searched for the presence of DPP-IV structures.
Table 3.1.: List of DPP-IV Protein structures from PDB database

<table>
<thead>
<tr>
<th>PDB ID CODE</th>
<th>EXPT METHOD</th>
<th>RESOLUTION</th>
<th>LIGANDS</th>
<th>TITLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A) 2BUB</td>
<td>X-RAY DIFFRACTION</td>
<td>2.66</td>
<td>FPB</td>
<td>Crystal structure of human dipeptidyl peptidase iv (cd26) in complex with a reversed amide inhibitor</td>
</tr>
<tr>
<td>B) 2G5T</td>
<td>X-RAY DIFFRACTION</td>
<td>2.30</td>
<td>ACF</td>
<td>Crystal structure of human dipeptidyl peptidase IV (DPPIV) complexed with cyanopyrrolidine (C5-pro-pro) inhibitor 21ag</td>
</tr>
<tr>
<td>C) 2G5P</td>
<td>X-RAY DIFFRACTION</td>
<td>2.40</td>
<td>ADF</td>
<td>Crystal structure of human dipeptidyl peptidase IV (DPPIV) complexed with cyanopyrrolidine (C5-pro-pro) inhibitor 21ac</td>
</tr>
<tr>
<td>D) 2G63</td>
<td>X-RAY DIFFRACTION</td>
<td>2.00</td>
<td>AAF</td>
<td>Crystal structure of human dipeptidyl peptidase IV (DPPIV) complexed with cyanopyrrolidine (C5-pro-pro) inhibitor 24b</td>
</tr>
<tr>
<td>E) 2I03</td>
<td>X-RAY DIFFRACTION</td>
<td>2.40</td>
<td>AXD</td>
<td>Crystal structure of human dipeptidyl peptidase 4 (DPP IV) with potent alkynyl cyanopyrrolidine (ABT-279)</td>
</tr>
<tr>
<td>F) 2OGZ</td>
<td>X-RAY DIFFRACTION</td>
<td>2.10</td>
<td>U1N</td>
<td>Crystal structure of DPP-IV complexed with Lilly aryl ketone inhibitor</td>
</tr>
<tr>
<td>G) 2IIT</td>
<td>X-RAY DIFFRACTION</td>
<td>2.35</td>
<td>872</td>
<td>Human dipeptidyl peptidase 4 in complex with a diazepan-2-one inhibitor</td>
</tr>
</tbody>
</table>
The materials required for docking studies include the following:

1. The first step in this direction involves the determination of the x-ray crystallography structures of proteins selected from the pdb database

2. The x-ray structures with the ligands are to be selected.

3. They are tabulated on the basis of the experimental data, ligands present and the resolution of the x-ray studies

   On the basis of above criteria 7 entries were found from the pdb database.

The no of disallowed regions residues were found from the Ramachandran plot statistics. The protein having the least no of residues in the disallowed regions is selected and taken for docking studies

3.16.2 Ramachandran Plot statistics

<table>
<thead>
<tr>
<th>No. of Residues</th>
<th>%-tage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most favoured regions [A, B, L]</td>
<td>1135</td>
</tr>
<tr>
<td>Additional allowed regions [a, b, l, p]</td>
<td>178</td>
</tr>
<tr>
<td>Generously allowed regions [~a, ~b, ~l, ~p]</td>
<td>7</td>
</tr>
<tr>
<td>Disallowed regions [XX]</td>
<td>0</td>
</tr>
<tr>
<td>Non-glycine and non-proline residues</td>
<td>1320</td>
</tr>
<tr>
<td>End-residues (excl. Gly and Pro)</td>
<td>2</td>
</tr>
<tr>
<td>Glycine residues</td>
<td>80</td>
</tr>
<tr>
<td>Proline residues</td>
<td>54</td>
</tr>
<tr>
<td>Total number of residues</td>
<td>1456</td>
</tr>
</tbody>
</table>
Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and $R$-factor no greater than 20.0 a good quality model would be expected to have over 90% in the most favoured regions [A, B, L].

Table 3.2: Showing the no of disallowed regions of proteins from the database

<table>
<thead>
<tr>
<th>PDB ID CODE</th>
<th>NO OF RESIDUES IN DISALLOWED REGIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>2BUB</td>
<td>5</td>
</tr>
<tr>
<td>2G5T</td>
<td>5</td>
</tr>
<tr>
<td>2G5P</td>
<td>8</td>
</tr>
<tr>
<td>2G63</td>
<td>8</td>
</tr>
<tr>
<td>2I03</td>
<td>9</td>
</tr>
<tr>
<td>2OGZ</td>
<td>2</td>
</tr>
<tr>
<td>2IIT</td>
<td>0</td>
</tr>
</tbody>
</table>

From the above table 2IIT showed nil residues in the disallowed regions.
3.16.3 PDB Summary data of 2IIT

Figure-3.37: PDB summary data of 2IIT
Figure-3.38: Image displaying various protein data bank features

Figure-3.39: Reference article for 2IIT
Schematic diagram of interactions between protein chains. Interacting chains are joined by coloured lines, each representing a different type of interaction, as per the key above. The area of each circle is proportional to the surface area of the corresponding protein chain. The extent of the interface region on each chain is represented by the black wedge whose size signifies the interface surface area. Statistics for this interface are given below.

**Interface statistics**

<table>
<thead>
<tr>
<th>Chain</th>
<th>No. of interface residues</th>
<th>Interface area (Å²)</th>
<th>No. of salt bridges</th>
<th>No. of disulphide bonds</th>
<th>No. of hydrogen bonds</th>
<th>No. of non-bonded contacts</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>49</td>
<td>2170</td>
<td>-</td>
<td>-</td>
<td>35</td>
<td>325</td>
</tr>
<tr>
<td>B</td>
<td>48</td>
<td>2183</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Figure-3.40: Image of Protein-protein interactions
Figure-3.41: Key residue interactions at the protein-protein surface

Key: Salt bridges, Disulphide bonds, Hydrogen bonds, Non-bonded contacts

The number of H-bond lines between any two residues indicates the number of potential hydrogen bonds between them. For non-bonded contacts, which can be plentiful, the width of the striped line is proportional to the number of atomic contacts.

Residue colours: Positive (H,K,R); negative (D,E); S,T,N,O = neutral; A,V,L,I = aliphatic; F,Y,W = aromatic; P,O = Pro&Oly; C = cysteine.
3.16.4 Protein 2IIT.pdb

The structures of the protein are downloaded from http://www.rcsb.org/pdb/ as PDB files. PDB files have an extension of 2IIT.pdb the 3D structure of the protein with its ligand was taken from pdb and its appropriate inhibitory molecules were also taken from the research article “(3R)-4-[(3R)-3-Amino-4-(2,4,5-tri.uorophenyl)butanoyl]-3-(2,2,2- tri.uoroethyl)-1,4-diazepan-2-one, a selective dipeptidyl peptidase IV inhibitor for the treatment of type 2 diabetes [by Tesfaye Biftu and et al] for docking studies

3.16.5 ACTIVE SITE RESIDUES:

The DPP-IV protein is a dimer i.e., it is having two chains. The active site residues found in the 5 A° region surrounding the ligand 8729001(A) are as follows:

Tyr 547 (A), Ser 630 (A), Tyr 666 (A), Tyr 631(A), Tyr 662(A), Val 711 (A), His 740 (A), Asn 710 (A), Glu 206 (A), Glu 205 (A), Phe 357 (A)

Figure-3.42: showing the active site residues of 2IIT protein
3.17 Inhibitors

A set of six inhibitory molecules were taken for docking studies with DPP-IV structure. The six molecules were drawn and tested for their activity with the protein selected from database. These inhibitory molecules were selected from the article on the basis of their activity (IC$_{50}$ values). 2 highly active, 2 moderately active and 2 low active molecules were selected.

The Table 5.3 below shows the molecules with their respective IC$_{50}$ values. In the given table Mol 15 and Mol 22 have low activity, Mol 1 and Mol 18 have moderate activity

<table>
<thead>
<tr>
<th>Molecule id</th>
<th>Activity(IC$_{50}$) in nM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mol 1</td>
<td>2.6</td>
</tr>
<tr>
<td>Mol 15</td>
<td>160</td>
</tr>
<tr>
<td>Mol 18</td>
<td>6.6</td>
</tr>
<tr>
<td>Mol 22</td>
<td>140</td>
</tr>
<tr>
<td>Mol 24</td>
<td>0.91</td>
</tr>
<tr>
<td>Mol 26</td>
<td>0.29</td>
</tr>
</tbody>
</table>

The molecules were drawn using ISIS draw 2.3 and then were converted to mol files. These molecules were then exported to TSAR where they were converted to 3D molecules and their energy was minimized. The molecule were then saved as mol2 files and used for docking.
Figure 3.43: Showing the molecules designed from the article
3.18 Software Tools

To perform docking studies Auto dock 3.0.5, ISIS draw 2.3, TSAR 3.0 and web lab viewer 4.0 were used as computational tools

3.18.1 AUTO DOCK

Auto dock is a suite of automated docking tools. It is designed to predict how small molecules, such as substrates, bind to a receptor of known 3D structure. Auto dock actually consists of two main programs: auto dock performs the docking of the ligand to a set of grids describing the target protein; auto grid pre-calculates these grids. In addition to using them for docking, the atomic affinity grids can be visualized. A graphical user interface called auto dock tools or ADT was utilized to generate grids and calculate dock scores and evaluate the conformers.

3.18.2 Docking Steps:

3.18.2.1 Preparing macromolecule 2IIT.pdb

Macromolecule was loaded using the option “load molecule” from the menu. Hydrogens were added from edit menu. Repair module was loaded from file menu to check for missing atoms and they were rectified. Charges were checked for integral values. Pdbqs file was written as macromolecule name. pdbqs

3.18.2.2 Preparing ligand

The ligand was loaded using input molecule option. The rigid root was selected by picking the atom from rigid root option. The rotatable bonds were defined using the option rotatable bonds. Write pdbq option was used to save the resultant pdbq format, which adds the charge parameter to the ligand.

3.18.2.3 Preparing the grid

The macromolecule was chosen using the option choose macromolecule from macromolecule under grid menu. The map types were set by reading formatted file of the ligand. Grid was set by the following parameters
Number of points in X-direction: 40
Number of points in Y-direction: 40
Number of points in Z-direction: 40
Centre (co-ordinates taken from the crystal ligand central atom):
  X- direction:  39.324
  Y- direction:  49.739
  Z- direction:  37.653

The gpf (Grid Parameter File) file was written as ligand name. gpf.

The macromolecule was chosen using the option choose macromolecule from macromolecule under docking menu. The ligand parameters (ligand name.pdbq file) were set. The Lamarckian Genetic algorithm parameters were employed to run protein–ligand docking. Docking run parameters were set to default. A dpf (Dock Parameter File) file was written

3.18.2.4 Run

Auto grid was started. Once, the run was complete, AutoDock was started

3.18.2.5 Calculation of results

From the docking logs option the dlg file was read. The different conformations were seen through show conformations. The best conformer was saved through choose docked conformation from molecules option